PROBABILITY AND STATISTICS: THEORY AND APPLICATIONS

D. A. S. FRASER
University of Toronto

DUXBURY PRESS
North Scituate, Massachusetts
PREFACE

WHY

In recent years important developments and trends have taken place in probability and statistics. Greatly increased attention is being given to the analysis of data, particularly to the development of inference methods that can be applied easily to data without restrictive models. The availability of desk and large-scale computers has opened up large areas of applications—the simple and direct completion of rather complicated analyses, on the one hand, and the numerical derivation of needed distributions and percentage points, on the other hand. The likelihood function is emerging as a basic means of assessing data and making inferences, greatly extending the range of standard methods. And in a related way the likelihood function as it depends on the sample space has strengthened theoretical methods and largely eclipsed the one-time prominence of sufficiency. The increased attention to the analysis of data has also highlighted the essential unity of statistical inference, with estimation, tests, and confidence regions appearing as closely related components. The availability of computers has also led to the development of exact inference methods for distribution forms other than the familiar normal, thus allowing proper attention to longer-tailed distributions. On the instructional side there is an increased need to organize introductory probability in a way that is adequate and appropriate to both the study of statistics and the continued study of probability. Conditional probability is receiving new attention and wider application; and on a larger scale there is an increasing need for the clear organization of basic probability concepts. Paralleling these important positive developments and trends there has been an unfortunate deemphasis—almost neglect—of basic principles of experimentation, no doubt caused in part by the broader attention that has been given to design and analysis details.

These developments and trends are important to all of probability and statistics, and particularly important to an introductory course in probability and statistics; for clearly they bear on almost all areas of theory and application. Existing texts do not reflect the important positive trends just described. For example, likelihood function analysis is almost entirely absent from existing texts; indeed, the likelihood function itself is incorrectly defined in most texts. And quite generally the texts omit or neglect reasonable contact with the very fundamental principles of experimentation.

This text was designed for a full-year course on probability and statistics, blending theory and application. The text was developed to bring together in an organized way the important positive trends just described and to focus on the basic principles of experimentation. The material was assembled to provide a broad coverage course in itself or a basic course for continued study in probability or in statistics.

The text presupposes a one-year course in calculus and assumes a concurrent
second course in calculus introducing multiple integration and the elements of linear algebra.

WHAT

The likelihood function is used prominently. Its correct definition leads to current likelihood analysis methods of inference and it also provides by far the shortest route to the classical concepts of sufficiency and minimal sufficiency.

The normal distribution is treated as just one of many distribution forms that can be proposed for applications. Recent theoretical results and computer methods are used to develop inference results for other distribution forms.

Conditional probability is given special and detailed attention. Recent examinations of conditional probability affirm it as the most basic concept of probability and statistics; it is probability itself, in applications.

The three methods of inference—estimation, significance tests, and confidence intervals—are introduced together as unified components of statistical inference. The introduction is made in an applied inference context using symmetry and large samples, two basic tools from probability theory; the usual formal theories are developed later.

The organization is in terms of concepts—concepts of probability and concepts of statistics. The notation is flexible, suitable for the informality of data analysis and for the formality of probability theory. An underlying theme is the basic goal of science—experimentation and the determination of cause–effect relations.

A large number of illustrations and geometrical pictures are used to clarify, illustrate, and give insight. In particular, plots are included for all the common distributions, both regular distributions and derived central and noncentral distributions. Thus some feeling for the form of a distribution accompanies the functional expression of the distribution.

Special attention is given to a broad selection of examples and problems covering theory, applications, data analysis, and computer usage. In addition, most chapters have a supplemental section that elaborates on theoretical points of interest, or provides the background material for references within chapters, or extends some interesting point or development within a chapter. These supplements are assembled at the back of the book; they can be used as supplemental reading for the particularly concerned student or as part of the basic material for a much more detailed and intensive course.

HOW

The material in this text is organized around the basic concepts of probability and statistics. These concepts are introduced section by section, chapter by chapter, building a solid, integrated framework for probability and for statistics. For example, the binomial distribution for independent Bernoulli trials is derived in Chapter 3, after the discussion of marginal distributions and after the treatment of independence. Probability and statistics have a logical framework and the concepts are introduced in a logical progression.

For probability theory, the most fundamental concepts are those of event and probability. Thus in Chapter 1 probabilities are introduced on a class of events; this
introduction is, of course, immediately paralleled in terms of sample spaces. The initial introduction involving just probabilities and events has the great advantage of allowing events to be treated later as primary entities. Various sample spaces are then easily examined, sample spaces suggested by convenience perhaps or by utility; and many basic results are seen trivially to be independent of the supporting sample space.

Chapter 2 is concerned with probability on real spaces. Again the essential concepts are those of event and probability. At this stage there is no need for functions—or random variables—for they deflect attention from the basic ingredients, the probabilities and the events.

Chapter 3 is concerned with marginal probability, with the use of functions to map into a simpler space or to restrict attention to a smaller class of events. Here the capital-letter notation $X$ is first used—where functions are really involved. Earlier sections use the quite flexible lowercase letter notation $x$, where of course the essential use of $x$ is to form events $\{x : x \leq 3\}$ on a sample space. The dual notation is consistent and it emphasizes that the function notation serves to refer back to an earlier sample space; for example, $(X \leq 3)$ is an abbreviation for $\{s : X(s) \leq 3\}$, where the lowercase letter $s$ is used for events on the earlier sample space. Indeed, the capital-letter notation should be used with extreme caution. For with conditional probability involving, say, normal$(\mu, 1)$ "random variables" $X_1, \ldots, X_n$, it does matter what the functions $X_1, \ldots, X_n$ are—different results with different functions. Accordingly, we avoid such "random variable" functions unless the functions are given explicitly.

Conditional probability in theory and in applications is a primary but often neglected concept. Chapter 4 focuses on the definition of conditional probability and its importance in applications. A selection model is used in Section 4.1; it gives a precise concrete interpretation for conditional probabilities.

Various aspects of means and variances are presented in Chapter 5. The development is organized so that most formulas are immediately available in multivariate form—no change in essential form, just the substitution of vectors for reals. Section 5.4 includes a wide range of problems recording means and variances for all the common distributions. An important section, 5.5, is included covering conditional means and variances.

Some results on limiting distributions and limiting functions are surveyed in Chapter 6.

For statistics the central theme is statistical inference leading to experimental design and the determination of cause-effect relations. In Chapter 7 statistical inference is introduced in the context of applications. This introduction uses the symmetry and large-sample results from the chapters on probability theory. And the applied context allows an easy and informal first contact with estimation, tests of significance, and confidence intervals—three closely related aspects of statistical inference; examples are taken from applications.

The more formal development of statistics in Chapter 8 is organized around the concepts of likelihood. On the applied side we obtain the likelihood function as a means of directly assessing data; the direct use of likelihood for statistical inference is a recent trend in statistics and it has great flexibility for applications. The likelihood function is introduced in Section 8.1B for data on the lifetime of crucibles used in an electrolytic reduction process. On the theoretical side we obtain the mapping that produces likelihood functions, called the likelihood statistic; this concept replaces the classical concept of a minimal sufficient statistic and has greater justifications for its use. Normal distribution theory and exponential models are introduced at this stage;
and as part of this we obtain the exact distributions for inference based on the normal model.

In Chapter 9 estimation theory is introduced, initially in an applied context of combining available estimates; a surveyor wishing to adjust data is used as an illustration in Section 9.1. This combining of estimates is generalized progressively and leads directly to the Gauss–Markov theorem itself. Then a full statistical model is introduced and the standard estimation results are obtained involving maximum likelihood, the information inequality, and the use of sufficiency and completeness.

In Chapter 10 hypothesis testing theory is developed; this includes the fundamental lemma, uniformly most powerful tests, and chi-square tests for qualitative data. Supplemental sections examine UMP unbiased and invariant tests and some formalities of confidence region theory.

Simple and general linear models are examined in Chapter 11, not just for normal error but for arbitrary error forms. For the location and scale model a computer program is available for the distribution of the \( t \)-statistic and the distribution of the standard deviation. The common distribution theory for quadratic forms is special to the normal error model; the development in this chapter gives this distribution theory as a special case but also provides inference methods more generally.

The concluding Chapter 12 surveys the basic principles of experimentation and introduces factorial designs; randomized blocks and covariance analysis are included in the supplemental sections. An adequate introduction to statistics must examine the elements of experimental design, as these elements comprise the fundamentals of science itself.

Answers to selected exercises and problems are recorded at the back of the book. A solutions manual for instructors, prepared by Gordon Fick with my participation, is available to instructors through the publisher.

A computer program for the inference methods in Chapter 11 has been developed by Gordon Fick. For copies write to: Statistics Section, Att.: Gordon Fick, Mathematics Department, University of Toronto, Toronto, Canada, M5S 1A1 (a card deck is available for a small service charge).

ACKNOWLEDGMENTS

I wish to thank the many students who have participated in the development of the present material through several revisions over a six-year period. In particular, Lane Bishop, David Brenner, Michael Evans, Gordon Fick, Georges Monette, Kai W. Ng, and Allan Wilks have worked very closely on the final revision and assisted immensely with the intricacies inherent in the proof stage.

My deep appreciation goes to Gordon Fick for his help and advice during the development of the book, for many fruitful suggestions during proofing, and for his central involvement with the solutions manual and the computer program for the nonnormal inference methods. My deep appreciation also goes to Frances Mitchell for her very careful, patient, and concerned typing of the manuscript and the solutions manual.

D. A. S. FRASER
CONTENTS

PREFACE

CHAPTER 1 PROBABILITY
  1.1 Introduction / 1
  1.2 Events / 9
  1.3 Probabilities / 19
  1.4 Calculating Probabilities / 28

CHAPTER 2 PROBABILITY ON THE LINE AND PLANE
  2.1 Probability on the Real Line: Discrete / 40
  2.2 Probability on the Real Line: Absolutely Continuous / 51
  2.3 Probability on the Plane / 69
  2.4 Change of Variables / 80

CHAPTER 3 MARGINAL PROBABILITY
  3.1 Data Reduction / 98
  3.2 Marginal Probability on the Line and Plane / 105
  3.3 Statistical Independence / 119
  3.4 Counting Sample Points / 128
  3.5 Marginal Distributions for Frequencies / 136

CHAPTER 4 CONDITIONAL PROBABILITY
  4.1 Definition and Interpretation / 152
  4.2 Conditional Probability on the Plane / 163

CHAPTER 5 MEAN VALUE FOR REAL AND VECTOR DISTRIBUTIONS
  5.1 Definition of the Mean / 180
  5.2 The Location of a Distribution / 186
  5.3 The Scaling of a Distribution / 193
  5.4 The Location and Scaling of a Sample / 207
  5.5 Conditional Location and Scaling / 217
  5.6 Moments of a Distribution / 224
  5.7 Moment-Generating and Characteristic Functions / 230
CHAPTER 6 LIMITING DISTRIBUTIONS AND LIMITING FUNCTIONS

6.1 Limiting Distributions / 244
6.2 The Weak Law of Large Numbers and The Central Limit Theorem / 254

CHAPTER 7 STATISTICAL INFERENCE

7.1 The Statistical Model and the Data / 268
7.2 Inference from Large Samples / 275
7.3 Inference for Finite Populations / 286
7.4 Inference from Symmetry / 292

CHAPTER 8 THE LIKELIHOOD FUNCTION IN STATISTICAL INFERENCE

8.1 The Observed Likelihood Function / 308
8.2 Distribution Theory for the Normal Model / 319
8.3 Statistical Inference for the Normal Model / 325
8.4 An Important Statistic / 332
8.5 Models That Have Exponential Form / 341
8.6 Inference from Large Samples Using Likelihood / 348

CHAPTER 9 ESTIMATION

9.1 Some Traditional Methods / 362
9.2 Linear Least Squares / 370
9.3 Combining Unbiased Estimators / 379
9.4 The Information Inequality for Unbiased Estimators / 389
9.5 Likelihood and Completeness for Unbiased Estimators / 398

CHAPTER 10 TESTING STATISTICAL HYPOTHESES

10.1 The Fundamental Lemma / 410
10.2 Composite Hypotheses / 421
10.3 Some Other Methods in Hypothesis Testing / 430
10.4 Testing a Statistical Model / 442

CHAPTER 11 LINEAR MODELS

11.1 The Location-Scale Models / 456
11.2 Calculating Projections and Residuals / 468
11.3 The Regression Model / 476
11.4 Inference with the Regression Model / 482

CHAPTER 12 THE DESIGN OF EXPERIMENTS

12.1 One-Factor Design / 490
12.2 Two-Factor Design / 501
APPENDIX I SUPPLEMENTARY MATERIAL 511

1.5 Probability / 511
2.5 Probability on the Line and Plane / 520
3.6 Marginal Probability / 526
4.3 Conditional Probability / 537
5.8 Mean Value / 544
6.3 Limiting Functions / 554
7.5 Testing Randomness / 563
9.6 Bayes Estimation / 572
10.5 Confidence Intervals and Confidence Regions / 579
12.3 Three-Factor Design / 584
12.4 Randomized Blocks / 590
12.5 Analysis of Covariance / 594

APPENDIX II TABLES 599

ANSWERS 605

INDEX 615
1

PROBABILITY

1.1

INTRODUCTION

In the past, most scientific investigations, particularly quantitative investigations, have been concerned with deterministic relationships: for example, with the extension of a spring as determined by the force applied, or the pressure in a flask containing a gas as determined by the temperature, or the rate of a chemical process as determined by the amount of a catalyst, or the height reached by a projectile as determined by the force applied, or a response variable as a function of an input variable.

A very close investigation of such systems, however, shows that results do not, in fact, reproduce exactly. That is, for any prescribed initial conditions, the results are found to vary from repetition to repetition by small amounts. This variation can derive from many sources of variation—from error in the measurement of the response variable, from variation in the initial materials and conditions (although kept constant as nearly as possible), and from variation in the internal operation of the system being investigated.

More recently, quantitative investigations have become prominent in the biological and social sciences. In these areas the relationships are no longer deterministic in quite the same way. The almost negligible variation found in the traditional physical system is replaced by a very prominent variation that can nearly conceal response effects that come from changes in the input variables. And the moderate need in the "deterministic" context for an adequate description of variation becomes an overwhelming need in these new areas.

*Probability theory provides the description for variation in systems operated under essentially constant conditions.* In addition, it can be combined with deterministic theory to provide the description for variation in systems operated under conditions that are changed by design. Probability theory is a *descriptive* part of science, describing variation; it is thus applicable in essentially all areas of scientific endeavor. Probability theory is developed and discussed in Chapters 1 through 6.

Variation is found in the full range of physical, biological, and social systems. In one direction we have the system of tossing a coin and recording a specified characteristic of the result. This is an example of *games of chance*, which involve coins, dice, cards, and roulette wheels. Such games of chance represent the historical origins of probability theory when prominent mathematicians assisted French gamblers in the seventeenth century. In the other direction, that of scientific areas, we have, for example, the measurement of a reaction time of a particular kind of animal in relation to treatment by a drug. In probability theory the games of chance provide a wide spectrum of examples in which input conditions are relatively easily controlled and kept
constant; these examples are used freely in the chapters on probability theory. The more scientific examples are involved with statistical theory and are largely withheld until later chapters on that theory. Sometimes illustrations are chosen from everyday activities (e.g., meeting buses at a terminal, selecting from a menu at a restaurant, or wrapping Christmas packages). We try to avoid such illustrations, as they seem to promote casual application without attention to conditions needed for validity.

Statistical theory is developed and discussed in Chapters 7 through 12. This theory builds on the use of probability to describe variation and is concerned with operating the system, obtaining observations on the system, and drawing conclusions concerning unknowns of the system; most important, it is concerned with designing input conditions so as to obtain conclusions concerning input–response and cause–effect relationships. Statistical theory is a prescriptive part of science, prescribing the scientist's method in many areas of scientific endeavor. There will be more on this beginning in Chapter 7.

For the study of variation we must control and hold constant the various input variables or factors affecting the system under investigation. We must then decide what is to be recorded concerning the result of a performance of the system. For the coin-tossing example we could record whether the event "heads" \( H \) occurs; or we could record whether the event "tails" \( T \) occurs. And for the reaction-time example we could record whether the reaction time lies in the interval \([2.5, 3.5]\). Then for any chosen event or characteristic, we can examine a large number of independent repetitions of the system and record the proportion of occurrences of the event in those repetitions; for the event \( H \) the proportion might be \( 0.53 \). Of course, with another large number of repetitions, the proportion may be slightly different. But it is a physical phenomenon that the proportion of occurrences for the event tends to a limit as the number of repetitions increases indefinitely; the limit is called the probability of the event. A system with this property is called a stable system or a random system. In cases where the proportion behaves systematically or does not tend to a limit, experience indicates that the input variables have not been controlled and kept constant; in other words, the system is not random.

The probability of an event is viewed as a property of a performance in the future or of a concealed performance in the past. The interpretation for such a probability is in terms of the long-run proportion of occurrences.

Consider this in more detail.

A EVENTS

Events are the primary elements of probability, statistics—indeed of science itself; as an example, consider the event "recovery" in the pharmaceutical investigation of a new antibiotic. As a simple example involving a random system, we will examine the rolling of two dice. An investigator may be interested in the event "double 1"; let \( D \) designate this event. Or he may be interested in the event "the dice are less than 3 inches from each other"; let \( L \) designate this event. Or he may be interested in the event "a particular die is within 10º of being lined up with the rectangular table"; let \( W \) designate this event.

We can think of an event as a set of descriptions for the response or outcome on a performance of the system. Or we can think of an event as a proposition concerning the outcome. Or we can think of an event in the everyday sense of a kind of event.
Thus an event can occur or not occur on a performance of the system; and, correspondingly, the set of descriptions can hold or not hold, and the proposition can be true or false. We use capital letters to designate events, for example $A$, $B$, $C$, . . . .

Consider some operations on events:

1. The event "'$A$ intersect $B'$" designated $A \wedge B$ or $AB$, is the event "'both $A$ and $B'$." For example, $D \wedge L$ is the event "'two 1s less than 3 inches from each other.'" The operation intersection can also be called meet or conjunction.

2. The event "'$A$ union $B'$" designated $A \vee B$, is the event "'either $A$ or $B$ (or both).'" For example, $D \vee L$ is the event "'double 1 or the dice are less than 3 inches from each other.'" The operation union can also be called join or disjunction.

3. The event "'$A$ complement'" designated $A^c$, is the event "'not $A$.'" For example, $L^c$ is the event "'the dice are at least 3 inches from each other.'" The operation complementation can also be called negation.

As a special event, let $\emptyset$ designate the null event; this handles, for example, the event "'$A \wedge A^c'$," which cannot occur. And let $\mathcal{U}$ designate the universal event; this handles, for example, the event "'$A \vee A^c'$," which necessarily occurs.

Consider some definitions for events:

1. The events $A$ and $B$ are disjoint if $A \wedge B = \emptyset$, that is, if they cannot occur simultaneously. Disjoint events can also be called mutually exclusive events. For example, $A$ and $A^c$ are disjoint.

2. The event $A$ is contained in the event $B$, designated $A \subseteq B$, if $A \wedge B = A$. This can be interpreted as: the occurrence of $A$ implies the occurrence of $B$.

3. The class of events $\{B_1, \ldots, B_k\}$ is a partition of the event $A$ if $A = B_1 \vee B_2 \vee \cdots \vee B_k$ and if events $B_i$ and $B_j$ with $i \neq j$ are disjoint. For example, $\{DL, DL^c\}$ is a partition of the event $D$, "'double 1.'"

**B EXAMPLE**

Consider the tossing of a certain first coin; we shall consider a second coin later. We might be interested in a variety of events: a certain face is upward, the coin is within a certain distance of the edge of the table, the orientation of the coin as determined by the upward face is within a certain angle of the directions established by the edge of the table. Suppose that we are interested in the event $H_1$, "'heads for this coin.'"

Consider the operations applied to the event $H_1$. If we ignore the possibility that the coin stands on edge, then $H_1^c$ is the event, say $T_1$, "'tails for this coin.'" The event $H_1 \wedge T_1 = \emptyset$ is the null event, and $H_1 \vee T_1 = \mathcal{U}$ is the universal event. Thus, accepting the operations, we become interested effectively in the class $\mathcal{B} = \{ \emptyset, H_1, T_1, \mathcal{U} \}$ of events. Note from an elementary point of view that we have produced a partition of the universal event:

```
\mathcal{U} \rightarrow H_1 \\
     \rightarrow T_1
```

The coin was tossed 100 times and for each event of interest a record was kept of whether or not it occurred on each repetition. For this it is convenient to let 1 designate
TABLE 1.1
Recording the event $A$

<table>
<thead>
<tr>
<th>Event $A$</th>
<th>Indicator $I_A(R_1)$</th>
<th>$I_A(R_2)$</th>
<th>$I_A(R_{100})$</th>
<th>Frequency $\sum_{i=1}^{100} I_A(R_i)$</th>
<th>Proportion $\bar{P}(A) = \sum_{i=1}^{100} I_A(R_i)/100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\emptyset$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$H_1$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>53</td>
<td>0.53</td>
</tr>
<tr>
<td>$T_1$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>47</td>
<td>0.47</td>
</tr>
<tr>
<td>$\Omega$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>100</td>
<td>1.00</td>
</tr>
</tbody>
</table>

occurrence and 0 designate nonoccurrence. The record for an event $A$ is then the succession of values of an indicator function $I_A$; thus $I_A(R_i)$ is equal to 1 or 0 according as the event $A$ occurs or does not occur on the $i$th repetition $R_i$. For the succession of 100 performances we obtained heads, tails, tails, . . . , heads; the tabulation is as given in Table 1.1.

For any event $A$ the total $\sum I_A(R_i)$ is the number of occurrences or the frequency of $A$ in the 100 repetitions; it is a measure of the event $A$, call it the weight of occurrence of $A$. The average $\bar{P}(A) = \sum I_A(R_i)/100$ is the proportion of occurrences of $A$ in the 100 repetitions; it is a measure of the event $A$, call it the estimated probability of $A$.

If we were to examine another 100 repetitions of the system, we would find typically that the estimated probabilities $\bar{P}(H_1)$ and $\bar{P}(T_1)$ were slightly different. Necessarily, of course, $\bar{P}(\emptyset) = 0$ and $\bar{P}(\Omega) = 1$.

For a larger batch of repetitions, say 1000, we could calculate estimated probabilities in a similar way. And for different batches of 1000 we would find typically that the estimated probabilities tend to vary less than for batches of 100.

As the number of repetitions is increased indefinitely, we would observe the physical phenomenon that the estimated probability for an event $A$ approaches a limiting value, say $P(A)$; this is a measure of the event $A$, the long-run measure of $A$, call it the probability or true probability of $A$.

Of course, beyond a certain point the coin may show appreciable wear, the initial conditions for the tossing of the coin would be different, and we would expect the probability of the event to be slightly changed.

Now suppose that the coin is symmetric, one face to the other, except for very delicate labeling that indicates heads for one face and tails for the other face. And suppose that the tossing of the coin is done carefully to ensure dynamic symmetry between the two faces. Then we do not need the large number of repetitions. For the symmetry between the two faces implies symmetry between the true measures for the two faces: $P(H_1) = P(T_1)$. Of course, the sum of the proportions for the two faces is always 1. Thus $P(H_1) + P(T_1) = 1$, and it follows that

\[
\bar{P}(\emptyset) = 0, \quad \bar{P}(H_1) = 1/2 = \bar{P}(T_1), \quad \bar{P}(\Omega) = 1;
\]

note that we are using a bar over $P$ to indicate a probability evaluated from symmetry. This gives us an approximation to the true probabilities, to the degree that symmetry represents the coin tossing.

Now consider the model that we have constructed for the variation in the coin-tossing system. For the events of interest we have a class
\[ \mathcal{B} = \{ \emptyset, H_1, T_1, \cup \} , \]

which is closed under the standard operations \( \land, \lor, \) and \( ^c \) on events. And for the long-run evaluation of the events in \( \mathcal{B} \) we have a measure or function \( P \) defined on \( \mathcal{B} \).

From the 100 performances, we have the probability measure \( P \) given by

<table>
<thead>
<tr>
<th>( \emptyset )</th>
<th>( H_1 )</th>
<th>( T_1 )</th>
<th>( \cup )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P )</td>
<td>0</td>
<td>0.53</td>
<td>0.47</td>
</tr>
</tbody>
</table>

as an approximation to the true measure; and from symmetry we have the probability measure \( P \) given by

<table>
<thead>
<tr>
<th>( \emptyset )</th>
<th>( H_1 )</th>
<th>( T_1 )</th>
<th>( \cup )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P )</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

as an approximation to the true measure.

In summary: we have constructed a model \( (\mathcal{B}, P) \), where \( \mathcal{B} \) is an appropriate class of events and \( P \) is the measure that gives the probability for each of the events in \( \mathcal{B} \). In an application one or other of the two approximations might be used for \( P \)—with appropriate cautions.

\section{Example (continuation)}

Consider the tossing of the preceding coin together with a second coin. And suppose that we are interested in the event \( H_1 \), "heads for the first coin," and the event \( H_2 \), "heads for the second coin." We are now observing in greater depth.

The standard operations applied to the events \( H_1 \) and \( H_2 \) generate 16 events altogether. For example, \( H_1^2 = T_1 \), \( H_2^2 = T_2 \), say; and, for example, \( (H_1H_2)^c \) is the event "at least one tail." The various events can be labeled in a simple manner and assembled as the class \( \mathcal{B}_2 \):

\[ \mathcal{B}_2 = \{ \emptyset, H_1H_2, H_1T_2, T_1H_2, T_1T_2, H_1, T_1, H_2, T_2, H_1H_2 \lor T_1T_2, H_1T_2 \lor T_1H_2, (H_1H_2)^c, (H_1T_2)^c, (T_1H_2)^c, (T_1T_2)^c, \cup \} \]

The events in the original class \( \mathcal{B} \) have been underlined. Thus note that the new class, \( \mathcal{B}_2 \), is an enlargement of \( \mathcal{B} \); that is, \( \mathcal{B} \subset \mathcal{B}_2 \). Also note from the point of view of partitions that we have a finer partition of the universal event:
For some later considerations, three of the events in $\mathcal{B}_2$ are of somewhat special interest and we give them alternative designations:

$$E_0 = T_1 T_2, \quad E_1 = H_1 T_2 \lor T_1 H_2, \quad E_2 = H_1 H_2.$$ 

Note that $E_j$ is the event "exactly $j$ heads for the two coins."

The pair of coins was tossed 100 times, in fact, the 100 tosses discussed earlier when we mentioned only the first coin. The outcomes, with the increased description, were heads–heads, tails–tails, . . . , heads–tails, giving the tabulation shown in Table 1.2. Note that it would suffice to tabulate the four elementary events $H_1 H_2$, $H_1 T_2$, $T_1 H_2$, $T_1 T_2$. For there is additivity such that for any other event we can obtain the indicator, the frequency, or the proportion by adding the corresponding values for component elementary events. Thus $H_1 H_2 \lor T_1 T_2$ has disjoint components $H_1 H_2$ and $T_1 T_2$ and $\hat{P}(H_1 H_2 \lor T_1 T_2) = \hat{P}(H_1 H_2) + \hat{P}(T_1 T_2)$. Also note that $\hat{P}$ here is an extension of $\hat{P}$ in the earlier example; the earlier values of $\hat{P}$ are underlined.

We have now constructed the model $(\mathcal{B}_2, \hat{P})$ for the coin-tossing system. The class $\mathcal{B}_2$ is an appropriate class of events, an enlargement of the original class $\mathcal{B}$. And $\hat{P}$ is the measure that gives the probability for each of the events in $\mathcal{B}_2$; it is an extension of the original $P$ to the enlarged class of events. From the 100 performances we have the measure $\hat{P}$ given in Table 1.3.

Now suppose that each coin is symmetric, one face to the other, except for delicate labeling that indicates heads for one face and tails for the other face. And suppose that the tossing is done carefully to ensure dynamic symmetry. Then we do not need the large number of repetitions. For the symmetry among the four events $H_1 H_2$, $H_1 T_2$, $T_1 H_2$, $T_1 T_2$ gives $P(H_1 H_2) = P(H_1 T_2) = P(T_1 H_2) = P(T_1 T_2)$; and the sum of the proportions for the four events is 1. Accordingly.

### Table 1.2

<table>
<thead>
<tr>
<th>Event $A$</th>
<th>Indicator $I_1(R_1)$</th>
<th>Indicator $I_4(R_2)$</th>
<th>Indicator $I_{100}(R_{100})$</th>
<th>Frequency $\sum_{i=1}^{100} I_i(R_i)$</th>
<th>$\hat{P}(A)$ = $\sum_{i=1}^{100} I_i(R_i)/100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\emptyset$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$H_1 H_2 = E_2$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>0.25</td>
</tr>
<tr>
<td>$H_1 T_2$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>28</td>
<td>0.28</td>
</tr>
<tr>
<td>$T_1 H_2$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>23</td>
<td>0.23</td>
</tr>
<tr>
<td>$T_1 T_2 = E_0$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>24</td>
<td>0.24</td>
</tr>
<tr>
<td>$H_1$</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>53</td>
<td>0.53</td>
</tr>
<tr>
<td>$T_1$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>47</td>
<td>0.47</td>
</tr>
<tr>
<td>$H_2$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>48</td>
<td>0.48</td>
</tr>
<tr>
<td>$T_2$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>52</td>
<td>0.52</td>
</tr>
<tr>
<td>$H_1 H_2 \lor T_1 T_2$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>49</td>
<td>0.49</td>
</tr>
<tr>
<td>$H_1 T_2 \lor T_1 H_2 = E_1$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>51</td>
<td>0.51</td>
</tr>
<tr>
<td>$(H_1 H_2)^c$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>75</td>
<td>0.75</td>
</tr>
<tr>
<td>$(H_1 T_2)^c$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>72</td>
<td>0.72</td>
</tr>
<tr>
<td>$(T_1 H_2)^c$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>77</td>
<td>0.77</td>
</tr>
<tr>
<td>$(T_1 T_2)^c$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>76</td>
<td>0.76</td>
</tr>
<tr>
<td>$\Omega$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>100</td>
<td>1.00</td>
</tr>
</tbody>
</table>
TABLE 1.3
Measure $P$ obtained from 100 performances

<table>
<thead>
<tr>
<th>$\mathcal{B}_2$</th>
<th>$P$</th>
<th>$\mathcal{B}_2$</th>
<th>$P$</th>
<th>$\mathcal{B}_2$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\emptyset$</td>
<td>0</td>
<td>$H_1$</td>
<td>0.53</td>
<td>$(H_1H_2)^c$</td>
<td>0.75</td>
</tr>
<tr>
<td>$H_1H_2$</td>
<td>0.25</td>
<td>$T_1$</td>
<td>0.47</td>
<td>$(H_1T_2)^c$</td>
<td>0.72</td>
</tr>
<tr>
<td>$H_1T_2$</td>
<td>0.28</td>
<td>$H_2$</td>
<td>0.48</td>
<td>$(T_1H_2)^c$</td>
<td>0.77</td>
</tr>
<tr>
<td>$T_1H_2$</td>
<td>0.23</td>
<td>$T_2$</td>
<td>0.52</td>
<td>$(T_1T_2)^c$</td>
<td>0.76</td>
</tr>
<tr>
<td>$T_1T_2$</td>
<td>0.24</td>
<td>$H_1H_2 \lor T_1T_2$</td>
<td>0.49</td>
<td>$\emptyset$</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$H_1T_2 \lor T_1H_2$</td>
<td>0.51</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\bar{P}(H_1H_2) = \bar{P}(H_1T_2) = \bar{P}(T_1H_2) = \bar{P}(T_1T_2) = 0.25$;

and the additivity mentioned above gives the probabilities for the other events. Thus from symmetry we obtain the measure $P$ given in Table 1.4.

TABLE 1.4
Measure $P$ obtained from symmetry

<table>
<thead>
<tr>
<th>$\mathcal{B}_2$</th>
<th>$P$</th>
<th>$\mathcal{B}_2$</th>
<th>$P$</th>
<th>$\mathcal{B}_2$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\emptyset$</td>
<td>0</td>
<td>$H_1$</td>
<td>1/2</td>
<td>$(H_1H_2)^c$</td>
<td>3/4</td>
</tr>
<tr>
<td>$H_1H_2$</td>
<td>1/4</td>
<td>$T_1$</td>
<td>1/2</td>
<td>$(H_1T_2)^c$</td>
<td>3/4</td>
</tr>
<tr>
<td>$H_1T_2$</td>
<td>1/4</td>
<td>$H_2$</td>
<td>1/2</td>
<td>$(T_1H_2)^c$</td>
<td>3/4</td>
</tr>
<tr>
<td>$T_1H_2$</td>
<td>1/4</td>
<td>$T_2$</td>
<td>1/2</td>
<td>$(T_1T_2)^c$</td>
<td>3/4</td>
</tr>
<tr>
<td>$T_1T_2$</td>
<td>1/4</td>
<td>$H_1H_2 \lor T_1T_2$</td>
<td>1/2</td>
<td>$\emptyset$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$H_1T_2 \lor T_1H_2$</td>
<td>1/2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

D COMMENTS

For our probability model we have chosen events as the basic entities; events are examined formally in Section 1.2. Alternatively, we could have chosen a labeling for the various possible outcomes at some degree or depth of observation. However, we would then have had to change the outcome description in going from the example in Section B to that in Section C.

The choice of events as the basic entities is slightly more complicated, but it does pay some rich dividends. For example, many results in later chapters are available immediately in an invariant form, invariant of the degree or depth of labeling of the outcomes of performances; we thus avoid the need for direct proofs of the invariance. Labeling, however, is a convenience and we shall introduce it in Section 1.2 as an easy device for presenting events.

E EXERCISES

1 Consider the tossing of a coin as described in Section B but now allow for the rare event $E$, "the coin stands on edge."
Chap. 1: Probability

(a) Record the class $\mathcal{B}'$ of events generated from $H$, $T$, and $E$ by the standard operations: it has eight elements and the operations can be used to form convenient designations.
(b) Tabulate the estimated probability measure $\hat{P}$ for the events in $\mathcal{B}'$; use the data in Section B.

2 A water sample can be described as pure, $P$ (no coliform bacteria); tolerable, $T$ (from 1 to 10 coliform bacteria); or unsafe, $U$ (more than 10 coliform bacteria). Present the 8 possible events obtained by the standard operations; give a brief description of each.

3 A lot containing 1000 trailer wheel bearings can be described as premium, $P$ (no defective bearings); tolerable, $T$ (from 1 to 5 defective bearings); or unsatisfactory, $U$ (more than 5 defective bearings). Present the 8 possible events obtained by the standard operations; give a brief description of each.

4 The offspring of a certain mating of mice can be described as male, $M$; female, $F$; black fur, $B$; brown fur, $R$, where $\{B, R\}$ is a partition of the universal event. Record the class $\mathcal{B}$ of 16 events generated from the given events.

F PROBLEMS

5 A water sample can be described as pure, $P$ (no coliform bacteria); safe, $S$ (from 0 to 10 coliform bacteria); or unsafe, $U$ (more than 10 coliform bacteria); compare with Exercise 2. For each of the 8 events as determined in Exercise 2, record a simple designation in terms of the events $P$, $S$, and $U$.

6 The binary numbers 0 and 1 have been used for the indicator functions in the tabulations of results in Sections B and C. The binary numbers can also be used to form alternative labels for events in relation to the "smallest," or elementary, events. Consider Exercise 1. The label $(1, 0, 0)$ can be used for $H$, $(0, 1, 0)$ for $T$, and $(0, 0, 1)$ for $E$; and then, for example, $(1, 1, 0)$ for $H \lor T$, and so on, where the three positions correspond to the three elementary events and 1(0) corresponds to the presence (absence) of the elementary event as a component.

(a) In terms of this labeling, show that
$$\mathcal{B}' = \{(i_1, i_2, i_3): i_j = 0, 1, j = 1, 2, 3\}.$$ 
(b) Verify that
$$\begin{align*}
(i_1, i_2, i_3) \land (i'_1, i'_2, i'_3) &= (\min (i_1, i'_1), \min (i_2, i'_2), \min (i_3, i'_3)), \\
(i_1, i_2, i_3) \lor (i'_1, i'_2, i'_3) &= (\max (i_1, i'_1), \max (i_2, i'_2), \max (i_3, i'_3)), \\
(i_1, i_2, i_3)^\complement &= (1 - i_1, 1 - i_2, 1 - i_3).
\end{align*}$$

7 Let $\mathcal{G} = \{A_1, \ldots, A_k\}$ be a partition of the universal event $\mathcal{U}$ into $k$ different nonnull events. Let $\mathcal{B}$ be the set of events generated from $\mathcal{G}$ by the standard operations; then in the manner used in Problem 6, $\mathcal{B}$ can be presented in the following form:
$$\mathcal{B} = \{(i_1, \ldots, i_k) : i_j = 0, 1, j = 1, \ldots, k\}.$$ 
Determine the operations $\land$, $\lor$, and $\complement$ in terms of this representation.

8 Let $\mathcal{G} = \{A_1, A_2\}$ be a class of events for the outcome of a certain system. And let $\mathcal{B}$ be the set of events generated from $\mathcal{G}$ by the standard operations. Then it follows easily that the number of elements in $\mathcal{B}$ (the cardinality of $\mathcal{B}$) is either 2, 4, 8, or 16. For each of these values give necessary and sufficient conditions in terms of the events $A_1, A_2, A_1^c, A_2^c, A_1 A_2$, and $A_1^c A_2^c$. 
1.2

EVENTS

We now consider events in more detail. There is, of course, an easy way of doing this—by means of sets on a sample space—and we examine this easy way in Sections B and C. The easy way, however, is at a price—complications arise whenever we change our depth of observation of the real-world system under investigation. To rise above these complications we first examine, in Section A, events as basic entities of the mathematical model. And, indeed, this is worthwhile. For, later, we shall have the substantial advantage of knowing that various results for events and probabilities are independent of the sample space used.

We examine probabilities in Section 1.3, and some rules for calculating probabilities in Section 1.4.

A ALGEBRA OF EVENTS

The standard operations $\land$, $\lor$, and $^c$ are almost trivial for the events in the examples of Sections 1.1B and 1.1C. We now need these operations for more general situations and we must examine the basic rules for them: for example, $A \lor B = B \lor A$.

Consider a random system, and let $\mathcal{B}$ be the class of events of interest. We now examine properties that go naturally with such a class $\mathcal{B}$.

We have interpreted an event as a set of descriptions or as a proposition concerning the outcome of the system being investigated. And we have presented the operations $\land$, $\lor$, and $^c$ as logical operations on these propositions. Some of the basic rules for these logical operations are now assembled. They do form a somewhat unwieldy collection, but fortunately our subsequent analysis will not need to work directly with them. Rather, they are assembled here to crystallize the formal framework that we are working in.

The rules of logic are best presented in an abstract setting. For this, let $\mathcal{B} = \{A, B, \ldots\}$ be a class of undefined elements $A, B, \ldots$; in an application the elements will designate events for the outcome of the system. And let $\land$ and $\lor$ be operations that combine two elements in $\mathcal{B}$ to form a single element of $\mathcal{B}$, and $^c$ be an operation that takes a single element of $\mathcal{B}$ into a single element of $\mathcal{B}$; in an application the operations will correspond to the logical operations discussed in Section 1.1. A class $\mathcal{B}$ with operations that satisfy the natural properties for events is called a Boolean algebra.

**DEFINITION 1**

A Boolean algebra is a class $\mathcal{B} = \{A, B, \ldots\}$ with operations $\land$, $\lor$, and $^c$ such that

(a) Reflexive: $AA = A$ \quad $A \lor A = A$

(b) Commutative: $AB = BA$ \quad $A \lor B = B \lor A$

(c) Associative: $A(BC) = (AB)C$ \quad $A \lor (B \lor C) = (A \lor B) \lor C$
(d) Distributive: \( A(B \lor C) = AB \lor AC \quad A \lor (BC) = (A \lor B)(A \lor C) \)

(e) Distributive: \( (AB)^c = A^c \lor B^c \quad (A \lor B)^c = A^cB^c \)

(f) Reflexive: \( (A^c)^c = A \)

(g) There are distinguished elements \( \emptyset \) and \( \Omega = \emptyset^c \) such that
\[
AA^c = \emptyset, \quad A \lor A^c = \Omega.
\]
\[
A \emptyset = \emptyset, \quad A\Omega = A, \quad A \lor \emptyset = A, \quad A \lor \Omega = \Omega.
\]

As examples consider the classes \( \mathcal{B} \) and \( \mathcal{B}_2 \) for the coin tossing discussed in Sections 1.1B and 1.1C. In most cases we will want our class of events to be more than a Boolean algebra. Consider a physicist who counts \( \alpha \)-particles in a 7.5-second interval from a radioactive source. Let \( E_i \) be the event "\( i \) particles in the 7.5-second interval." The physicist might be interested in the event "more than five particles"; in symbols this would be \( E_5 \lor E_6 \). . . . Thus we must be able to apply our operations a countable number of times. For this we define a Boolean \( \sigma \)-algebra; the \( \sigma \) denotes countable. In the following definition we use \( A \subseteq B \); recall that this is formally defined by the condition \( A \land B = A \).

**DEFINITION 2**

A Boolean \( \sigma \)-algebra \( \mathcal{B} \) is a Boolean algebra such that

(a) For any sequence of elements \( A_1, A_2, \ldots \) in \( \mathcal{B} \), there is an intersection element \( C \) (denoted \( \bigcap_{i=1}^\infty A_i \)) in \( \mathcal{B} \): \( C \) is contained in (\( \subseteq \)) each \( A_i \); and any other element in \( \mathcal{B} \) that is contained in each \( A_i \) is also contained in \( C \).

(b) For any sequence of elements \( A_1, A_2, \ldots \) in \( \mathcal{B} \), there is a union element \( D \) (denoted \( \bigvee_{i=1}^\infty A_i \)) in \( \mathcal{B} \): \( D \) contains (\( \supseteq \)) each \( A_i \); and any other element in \( \mathcal{B} \) that contains each \( A_i \) also contains \( D \).

In terms of our logical framework, Boolean \( \sigma \)-algebras have applications throughout science and technology. For consider the pharmaceutical assessment of the antibiotic mentioned in Section A; the investigator would be interested in "recovery" or its complement for each patient in a clinical trial, as well as for combinations of these.

We could easily construct an example that has the physicist count \( \alpha \)-particles. But we wait, as we will soon have a wealth of examples based on a sample space, the concept to be developed next.

In conclusion, we assume that the class \( \mathcal{B} \) of events for our random system is a Boolean \( \sigma \)-algebra.

**B THE SAMPLE SPACE**

Consider a random system and let \( \mathcal{B} \) be the class of events of interest. As part of this we have, of course, implicitly decided on the depth of observation, the degree to which we are going to recognize characteristics of the result of a performance. For
example, with the coin-tossing system we could recognize only the face showing for the
first coin, or only the faces showing for the first and the second coins, or perhaps only
certain characteristics of location and orientation of the coins.

Let $s$ designate a possible outcome of the system, given an agreed depth of
observation; we call such an outcome a **sample point**. We now define: The sample
space is the set $\mathcal{S} = \{s\}$ of possible outcomes, the set of sample points. As alternative
notation we might use $t$ and $\mathcal{S}$ or, say, $x$ and $\mathcal{X}$: As an application, consider the
following pharmaceutical illustration.

**EXAMPLE 1**

**Life or death:** Consider the administration of a drug to a rabbit obtained from a
certain genetic strain. And suppose that we have decided to observe only whether
the animal lives $x_1 = 0$ or dies $x_1 = 1$ in a test period. Then the sample space is
$\mathcal{S} = \{x_1\} = \{0, 1\}$.

Now suppose that we have decided to observe whether the animal lives $x_1 = 0$ or
dies $x_1 = 1$ and whether a second animal tested lives $x_2 = 0$ or dies $x_2 = 1$. Then
the sample space becomes

$$
\mathcal{S}_2 = \{(x_1, x_2) : x_1 = 0, 1\} = \{(0, 0), (0, 1), (1, 0), (1, 1)\}

= \{0, 1\} \times \{0, 1\} = \{0, 1\}^2;
$$

this is the Cartesian product of the set $\{0, 1\}$ with itself. The sample spaces $\mathcal{S}$ and
$\mathcal{S}_2$ are presented in Figure 1.1.

**EXAMPLE 2**

**Performance rating:** Consider a psychology experiment in which a subject's
performance is rated from 1 to 6. For two subjects let $x_i$ be the rating for the $i$th
subject. Then the sample space is

$$
\mathcal{S} = \{(x_1, x_2) : x_i = 1, 2, \ldots, 6\}

= \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\} = \{1, 2, 3, 4, 5, 6\}^2;
$$

this is the Cartesian product of the set $\{1, 2, 3, 4, 5, 6\}$ with itself.

---

**FIGURE 1.1**

Sample spaces $\mathcal{S}$ for a single animal and $\mathcal{S}_2$ for two animals. Also, sample spaces
$\mathcal{S}$ and $\mathcal{S}_2$ in Example 3.
Sometimes we shall use a lowercase letter, say $s$, to refer to a particular outcome in $\mathcal{S}$. And sometimes we shall use it to refer to an arbitrary outcome in $\mathcal{S}$, as, for example, in defining a subset of $\mathcal{S}$ such as $\{s : s > 5\} = \{s > 5\}$; recall the example on $\alpha$-particles. There are some real advantages to keeping the dual usage and not making a formal distinction in the notation. Of course, the context will always make clear the distinction as needed.

Sometimes it is convenient to let the sample space be larger than needed, to be a convenient and simple mathematical space. For example, in the rating of a single subject’s performance, the sample space is $\{1, 2, 3, 4, 5, 6\}$, but it may be convenient to use the space of natural numbers $\mathbb{N} = \{1, 2, 3, \ldots\}$, or the space of integers $\mathbb{Z} = \{\ldots, -1, 0, 1, \ldots\}$, or even the space $\mathbb{R}$ of real numbers. Of course, we shall have ample opportunity later to indicate that certain sample points in effect are not possible outcomes.

### C EVENTS REPRESENTED BY SUBSETS

Consider a random system. Let $\mathcal{B}$ be the class of events of interest, and let $\mathcal{S}$ be the sample space corresponding to a chosen depth of observation. Now consider an event $B$ in $\mathcal{B}$. The event can be viewed as a proposition concerning the outcome. Accordingly, we can examine each outcome in $\mathcal{S}$, collect those outcomes for which the proposition is true, and obtain a subset, say $A$, of the space $\mathcal{S}$. The event $B$ is thus represented as a subset $A$ of $\mathcal{S}$, the truth set for the event.

Let $\mathcal{A} = \{A\}$ consist of the subsets of $\mathcal{S}$ that correspond to events in $\mathcal{B} = \{B\}$. Often it will be convenient to refer to $\mathcal{A}$ itself as the class of events for the system. Clearly the operation intersection $\wedge$ for events becomes the operation intersection $\cap$ for sets, union $\vee$ becomes union $\cup$, and complement $^c$ becomes ordinary complement $^c$; and, of course, the null event $\emptyset$ becomes the empty set $\emptyset$, and disjoint events $B_1$ and $B_2$ ($B_1 \cap B_2 = \emptyset$) become disjoint sets $A_1$ and $A_2$ ($A_1 \cap A_2 = \emptyset$).

Sometimes $\mathcal{A}$ may consist of all the subsets of the sample space, as in the following examples.

### EXAMPLE 3

**Tossing two coins:** Consider the tossing of two coins as discussed in Sections 1.1B and 1.1C. And suppose that we recognize only the face of the first coin: let $x_1 (= 0, 1)$ be the number of heads showing for that coin. Then $\mathcal{S} = \{x_1\} = \{0, 1\}$. The class of all subsets of $\mathcal{S}$ is

$$\mathcal{A} = \{\emptyset, \{1\}, \{0\}, \{0, 1\}\},$$

where $\emptyset$ now designates the empty set. The class $\mathcal{A}$ has four elements, and they can be put in one-to-one correspondence with the elements of $\mathcal{B}$ in Section 1.1B. Now suppose that we observe the face on the first coin and the face on the second coin: let $x_i$ be the number of heads for $i$th coin. Then $\mathcal{S}_2 = \{(x_1, x_2)\} = \{0, 1\}^2$. The class of all subsets of $\mathcal{S}_2$ is

$$\mathcal{A}_2 = \{\emptyset, \{(0, 0)\}, \ldots, \{(0, 0), (0, 1)\}, \ldots, \{(0, 0), (0, 1), (1, 0)\}, \ldots, \mathcal{S}_2\}.$$
The class $\mathcal{A}_2$ has 16 elements and they can be put in one-to-one correspondence with the elements of $\mathcal{B}_2$ in Section 1.1C. Note that the event $H_1$ is represented by the set \{1\} relative to $S$ and by the set \{(1, 0), (1, 1)\} relative to $S_2$. See Figure 1.1 and compare with Example 1.

**EXAMPLE 4**

Tossing two dice: Suppose that we recognize only the points on the first die and the points on the second die: let $x_i$ be the number of points on the $i$th die. The sample space is $S = \{(x_1, x_2): x_1 = 1, 2, \ldots, 6\}$; see Figure 1.2. The event $L$, "at least one 6," can be represented by the subset

$L = \{(1, 6), (2, 6), \ldots, (6, 6), (6, 5), \ldots, (6, 2), (6, 1)\}$.

The event $T$, "total points is 7," can be represented by the subset

$T = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$.

Note, for example, that $L \cap T = \{(1, 6), (6, 1)\}$ is the event "a 1 and a 6." Compare with Example 2.

**EXAMPLE 5**

One card from a deck: A card is dealt from a well-shuffled deck of 52 playing cards and its designation is recorded. The sample space is

$S = \{\text{AS, KS, \ldots, 2S, AH, \ldots, 2H, AD, \ldots, 2D, AC, \ldots, 2C}\}$,

where, for example, AS designates "ace of spades" and 2C designates "2 of clubs"; $S$ contains 52 points. Let $\mathcal{G}$ be the class of all subsets of $S$; then

$\mathcal{G} = \{\emptyset, \{\text{AS}\}, \ldots, \{2C\}, \{\text{AS, KS}\}, \ldots, \{3C, 2C\}, \ldots, S\}$.

A little calculation shows that $\mathcal{G}$ has $2^{52}$ elements (corresponding to the presence or absence of each sample point).

---

**FIGURE 1.2**
Sample space $S$ for two dice. Also, sample space $S_2$ in Example 2.
With a finite or countable sample space $\mathcal{S}$ it is quite natural to take $\mathcal{G}$ to be the class of all subsets of $\mathcal{S}$. The class of all subsets of a space $\mathcal{S}$ is called the power set of $\mathcal{S}$ and designated $\mathcal{P}(\mathcal{S})$.

In Example 5 the sample space has 52 points and we noted that the class $\mathcal{G}$ has $2^{52}$ points, a very large number. Perhaps this should be a warning. Indeed, if we consider a sample space as large as the real line $\mathbb{R}$, as, for example, with a single measurement of a physical quantity, then it can be shown that the power set of $\mathbb{R}$ is too big, too complicated, and in fact impossible for some quite reasonable problems involving variation. Often we must settle for a class $\mathcal{G}$ smaller than the power set; we now examine properties for such a class.

**D  ALGEBRA OF SUBSETS**

Consider a sample space $\mathcal{S}$ and let $\mathcal{G}$ be a class of subsets of $\mathcal{S}$. If $\mathcal{G}$ is to work as a class of events, we must be able to apply the standard operations $\cap$, $\cup$, $\epsilon$, $\cup^c$, and $\cap^c$ and remain with the elements of $\mathcal{G}$. Such a class of subsets is called a $\sigma$-algebra:

**DEFINITION 3**

A $\sigma$-algebra is a nonempty class $\mathcal{G} = \{A\}$ of subsets of a space $\mathcal{S}$ such that

(a) Closed under $\sigma$-union: If $A_1, A_2, \ldots \in \mathcal{G}$, then $\cup^c_i A_i \in \mathcal{G}$.

(b) Closed under complement: If $A \in \mathcal{G}$, then $A^c = \mathcal{S} - A \in \mathcal{G}$.

Closure under $\sigma$-intersection can be proved from (a) and (b); see Problem 20.

The combination $(\mathcal{S}, \mathcal{G})$ of a sample space $\mathcal{S}$ and a $\sigma$-algebra $\mathcal{G}$ over $\mathcal{S}$ is called a measurable space.

In Section A we assumed that our class of events satisfied the logic axioms of a Boolean $\sigma$-algebra. It is natural, then, to inquire whether our present class $\mathcal{G}$, given by Definition 3, satisfies the axioms of a Boolean $\sigma$-algebra. Fortunately, the answer is "yes": a $\sigma$-algebra $\mathcal{G}$ over a sample space $\mathcal{S}$ is automatically a Boolean $\sigma$-algebra. Thus quite freely, from now on, we shall express our algebra of events as a $\sigma$-algebra over a suitable sample space. Of course, we should remember, as indicated by Example 3, that many different sample spaces can be used to present a given Boolean $\sigma$-algebra of events; the advantage of the Boolean approach is that it gives us an invariant framework for building the probability model.

The verification that the axioms of a Boolean $\sigma$-algebra hold for subsets of a space is straightforward. For example, consider $(A_1A_2)^c = A_1^c \cup A_2^c$. A point $s$ belongs to the left side, if and only if $s$ does not belong to $A_1A_2$: $s$ does not belong to $A_1A_2$ if and only if $s$ does not belong to $A_i$ for some $i$; $s$ does not belong to $A_i$ for some $i$ if and only if $s$ belongs to $A_i^c$ for some $i$; $s$ belongs to $A_i^c$ for some $i$ if and only if $s$ belongs to the right side. This argument can be abbreviated as

\[
 s \in (A_1A_2)^c \iff s \notin A_1A_2 \iff s \notin A_i \text{ for some } i \\
 \iff s \in A_i^c \text{ for some } i \iff s \in A_i^c \cup A_i^c.
\]

Consider now some examples in which the class $\mathcal{G}$ is smaller than the power set.
**EXAMPLE 6**

*One measurement by a physicist:* Consider a physicist measuring a current flow, and suppose that he makes a single measurement $y$. The sample space is $\mathcal{S} = \{ y : y \in \mathbb{R} \} = \mathbb{R}$.

The real line; note that we are allowing any positive or negative value for $y$. The physicist might be interested in the event that $y$ falls in an interval, say $(1, 2)$; or in a finite union of intervals, say $(-2, -1) \cup (1, 2)$. Certainly we would want the class $\mathcal{G}$ to include all the intervals $(a, b)$. Accordingly, we take $\mathcal{G}$ to be the smallest $\sigma$-algebra that includes all the intervals; of course, any finite or countable union of intervals is an element of $\mathcal{G}$. Problem 25 shows that it is meaningful to talk about a smallest $\sigma$-algebra. An element of $\mathcal{G}$ is called a *Borel set* on the real line; and the class $\mathcal{G}$ is called the *Borel class* on the real line and designated $\mathcal{B}^1$. It is of some interest that the class of countable unions of intervals is only a very small step toward the Borel class; and also that the Borel class is only a very small step toward the power set $\mathcal{P}(\mathbb{R})$.

**EXAMPLE 7**

*k measurements by a physicist:* Consider the physicist in Example 6 and suppose that he makes $k$ measurements $y_1, \ldots, y_k$ on the current flow. The sample space then is

$$\mathcal{S} = \{(y_1, \ldots, y_k) : y_i \in \mathbb{R} \} = \mathbb{R}^k.$$

The $k$-dimensional space over the reals. The physicist might be interested in the event that $y' = (y_1, \ldots, y_k)$ falls in a rectangle, say, $(a_1, b_1) \times \cdots \times (a_k, b_k)$; or in some finite union of rectangles; see Figure 1.3. Certainly we would want the class $\mathcal{G}$ to include all the rectangles $(a_1, b_1) \times \cdots \times (a_k, b_k)$; and we would want to include rectangles partially closed in various ways. Accordingly, we take $\mathcal{G}$ to be the smallest $\sigma$-algebra that includes all such rectangles; this is called the *Borel class* on $\mathbb{R}^k$ and is designated $\mathcal{B}^k$. It is of interest that the class of countable

---

**FIGURE 1.3**

Set $A$, a finite union of rectangles.
unions of rectangles is only a very small step toward the Borel class; and also that
the Borel class is only a very small step toward the power set \( \mathcal{P}(\mathbb{R}^k) \).

In general, if the sample space \( S \) is \( \mathbb{R}^1 \) or say \( \mathbb{R}^k \), we shall take the algebra of events
to be the Borel class over that space. In a sense the Borel class represents the logical
consequences of considering intervals and assuming closure under \( \sigma \)-union and
\( \sigma \)-intersection.

**E INDICATOR FUNCTION FOR AN EVENT**

We now introduce the *indicator function* \( I_A \) as an alternative way of representing a
subset \( A \). We shall see later that indicator functions provide a simple mechanical route
to otherwise complicated results. Recall the brief use of indicators in Sections 1.1B
and 1.1C.

**DEFINITION 4**

The *indicator function* \( I_A \) for a subset \( A \subset S \) is the real-valued function on \( S \) given
by

\[
I_A(s) = \begin{cases} 
1 & \text{if } s \in A \\
0 & \text{otherwise.}
\end{cases}
\]

For an example, see Figure 1.4.

Now, instead of thinking of a subset \( A \) of a space \( S \), we can think of a special kind
of real-valued function \( I_A \) over \( S \). Of course, an indicator function takes only the values
0 and 1, but perhaps we can think ahead to more general functions that take, say,
"part" of a sample point or take a sample point with a positive or negative "weight."

Consider some of the basic properties that make indicator functions very useful.

1. For any set \( A \), \( I_A = 1 - I_{A^c} \).

Proof  The proof can be given in the abbreviated form used in Section D:

\[
I_A(s) = 1 \iff s \in A \iff s \notin A^c \iff I_{A^c}(s) = 0 \iff 1 - I_{A^c}(s) = 1.
\]

2. For any sets \( A_1, A_2, \ldots \),
Sec. 1.2: Events

\[ I \cap_{A_i} = \prod_{1}^{\infty} I_{A_i} \]

3. For any disjoint sets \( A_1, A_2, \ldots \),

\[ I \cup_{A_i} = \sum_{1}^{\infty} I_{A_i} \]

The proofs for these follow the pattern indicated for property 1.

As a small indication of the usefulness of indicator functions, consider the following calculation: for any sets \( A \) and \( B \),

\[ I_{A \cup B} = 1 - I_{(A \cup B)^c} = 1 - I_{A^c B^c} = 1 - (1 - I_A)(1 - I_B) = I_A + I_B - I_{AB}. \]

F Exercises: Games of Chance

1. Consider tossing a die and recording \( y_1 \), the number of points showing.
   (a) Record the sample space \( S \) for \( y_1 \).
   (b) Present the following events as subsets of \( S \): \( A \), "the number is even"; \( B \), "the number is divisible by 3"; and \( C \), "the number is less than 4"; \( A \cap B, A \cap C \), and \( A \cup B \).

2. Consider tossing a die and a coin and recording \( (y_1, y_2) \), where \( y_1 \) is the number of points on the die and \( y_2 \) is the number of heads.
   (a) Record the sample space \( S \) for \( (y_1, y_2) \).
   (b) Present the following events as subsets of \( S \): \( y_1 \) is even"; \( y_1 \) is divisible by 3"; \( y_1 + y_2 = 1 \); \( y_1 + y_2 = 2 \); and \( y_1 + y_2 \leq 2 \).

3. Two dice are tossed and \( (y_1, y_2) \) is recorded, where \( y_i \) is the number of points on the \( i \)th die.
   (a) Record the sample space \( S \) for \( (y_1, y_2) \).
   (b) Present the following events as subsets of \( S \): \( A \), "the total points is 6"; \( B \), "the total points is 7"; \( |y_2 - y_1| = 3 \); and \( |y_2 - y_1| = 4 \). Record these events on a diagram.

4. Two dice are tossed and \( (y_{(1)}, y_{(2)}) \) is recorded, where \( y_{(1)} \) is the smaller number showing and \( y_{(2)} \) is the larger.
   (a) Record the sample space \( S \) for \( (y_{(1)}, y_{(2)}) \).
   (b) Present the following events as subsets of \( S \): \( A \), "the total points is 6"; \( y_{(2)} - y_{(1)} = 3 \); and \( A \cap \{ y_{(2)} - y_{(1)} = 3 \} \). Record these events on a diagram.

5. Three coins are tossed; let \( y_i = 0(1) \) according as the \( i \)th coin is tails (heads).
   (a) Record in full the sample space \( S \) for \( (y_1, y_2, y_3) \).
   (b) Let \( H_i \) be the event "heads on the \( i \)th toss." Present \( H_1, H_1 H_2, \) and \( H_1 H_2 H_3 \) as subsets of \( S \).
   (c) Let \( E_j \) be the event "exactly \( j \) of the events \( H_1, H_2 \), and \( H_3 \) occur." Present \( E_0, E_1, E_2, \) and \( E_3 \) as subsets of \( S \).

6. Consider an urn containing 95 green balls and 5 black balls. The balls are mixed and two balls are drawn in succession, giving \( (y_1, y_2) \), where \( y_i = g(b) \) if the \( i \)th ball is green (black). Record the sample space \( S \) for \( (y_1, y_2) \). Present the following events as subsets of \( S \): \( E_0, E_1, \) and \( E_2 \), where \( E_j \) is the event "exactly \( j \) black balls."

7. An urn contains 3 green balls, marked \( g_1, g_2, \) and \( g_3, \) and one black ball, marked \( b \). The balls are thoroughly mixed and two balls are drawn in succession, yielding \( (y_1, y_2) \), where \( y_i \) is the mark on the \( i \)th ball. Record the sample space \( S \). Present the following events as subsets of \( S \): \( E_0, E_1, \) and \( E_2 \), where \( E_j \) is the event "exactly \( j \) black balls."

8. A deck of 52 playing cards is shuffled and two cards are dealt in succession face up. Describe the sample space \( S \) for \( (y_1, y_2) \), where \( y_i \) is the designation on the \( i \)th card dealt.
How many points does $S$ contain? Describe the subset $A$ for the event "both are spades."

How many points does $A$ contain?

9 A deck of 52 playing cards is shuffled and a card is dealt; the card is returned to the deck, the
deck is shuffled, and a second card is dealt. Describe the sample space for $(y_1, y_2)$, where $y_i$
is the designation on the $i$th card dealt. How many points does $S$ contain? Describe the
subset $A$ for the event "both are spades." How many points does $A$ contain?

10 Four cards are numbered 1, 2, 3, and 4. The cards are shuffled face down and dealt in
succession, giving $(y_1, y_2, y_3, y_4)$; as the cards are dealt, the dealer calls the numbers 1, 2, 3,
and 4 in succession.

(a) Present the sample space $S$ for $(y_1, y_2, y_3, y_4)$.

(b) Let $A_i$ be the event "the $i$th card is called correctly." Present $A_1$ as a subset of $S$.

(c) Let $E_j$ be the event "exactly $j$ of the events $A_1, A_2, A_3,$ and $A_4$." Present $E_2, E_3,$ and $E_4$
as subsets of $S$.

G EXERCISES: INDUSTRY AND SCIENCE

11 Transformers are mass-produced and packaged in lots of 100. A particular lot is subjected to
a sampling plan: two transformers are sampled in succession, tested, and $(y_1, y_2)$ is recorded,
where $y_i = g(b)$ if the $i$th transformer is good (bad). Record the sample space $S$ for $(y_1, y_2)$ for
a typical lot; certain exclusions would occur with exceptional lots. Present the following events
as subsets of $S$: $E_0, E_1,$ and $E_2$, where $E_j$ is the event "exactly $j$ bad items."

12 A physicist records $(y_1, y_2)$, where $y_1$ ($y_2$) is the number of $\alpha$-particles in a first (second)
7.5-second interval.

(a) Present the sample space $S$ for $(y_1, y_2)$ and make a diagram.

(b) Present the following events as subsets of $S$ and mark them on the diagram: the average
$\bar{y} = (y_1 + y_2)/2 \leq 3.5$; max $(y_1, y_2) \leq 4$; and min $(y_1, y_2) \leq 2$.

13 A psychologist records $(y_1, \ldots, y_n)$, where $y_i$ is the logarithm of the reaction intensity for the
$i$th subject. For simplicity take $n = 2$. Present the following events as subsets and mark
them on a diagram: $\bar{y} = (y_1 + y_2)/2 \leq 5$; max $(y_1, y_2) \leq 5$; and min $(y_1, y_2) \leq 3$.

14 Details as in Exercise 13 with $n = 2$. Let $y_{(1)}$ be the smallest, $y_{(2)}$ the second smallest, \ldots \ $y_{(n)}$
the largest of $y_1, \ldots, y_n$; $y_{(1)}, \ldots, y_{(n)}$ are called the order statistics. Present the following events
as subsets and mark them on a diagram: $y_{(2)} \leq 4$; $y_{(1)} \leq 2$; and $y_{(2)} - y_{(1)} \leq 2$.

15 A chemist makes $n$ determinations $(y_1, \ldots, y_n)$ on the rate of a chemical reaction; positive and
negative values are possible. The sample space $S = R^n$. For simplicity take $n = 2$: thus
$S = \mathbb{R}^2$. Present the following events as subsets and mark them on a diagram: $y_1 \leq 5$;
y_2 \leq 5; y_1 + y_2 \leq 5; \text{and } |y_2 - y_1| \leq 5$.

16 Details as in Exercise 15 with $n = 2$. Present the following events as subsets and mark them
on a diagram: $y_{(1)} \leq 5$; $y_{(2)} \leq 5$; and $y_{(2)} - y_{(1)} \leq 5$. For notation, see Exercise 14.

17 Details as in Exercise 15 with $n = 2$. The sample average is $\bar{y} = (y_1 + \cdots + y_n)/n$; the
sample variance is $s_y^2 = \Sigma (y_i - \bar{y})^2/(n - 1)$. Present the following events as subsets and
mark them on a diagram: $\bar{y} - 4.1 \leq 3/\sqrt{2}$; and $s_y^2 \leq 4$.

18 Details as in Exercise 15 with $n = 2$. The $t$-statistic relative to the location 4.1 is

$$t(y_1, \ldots, y_n) = \frac{\bar{y} - 4.1}{s_y / \sqrt{n}}.$$ Present the following event as a subset and mark it on a diagram: $t(y_1, y_2) \leq 3$.

H PROBLEMS

19 Consider subsets $A_i$ of a sample space $S$. The relation $(A_1 A_2) \cap = A_1 \cup A_2$ discussed in
Section C can be generalized: prove De Morgan's laws:

$$(\cap_i A_i)^c = \cup_i A_i^c; \quad (\cup_i A_i)^c = \cap_i A_i^c.$$
Thus complement of intersection is union of complements; complement of union is intersection of complements.

20 Let $\mathcal{A}$ be a $\sigma$-algebra of subsets of a sample space $\mathcal{S}$. Prove that $\mathcal{A}$ is closed under $\sigma$-intersection.

21 Consider a sample space $\mathcal{S}$. An indicator function is a map from $\mathcal{S}$ into $\{0, 1\}$. The class of all indicator functions can be designated $2^\mathcal{S}$, referring to binary-valued functions on $\mathcal{S}$.

(a) Show that the power set $\mathcal{P}(\mathcal{S})$ is equivalent to $2^\mathcal{S}$. Compare with Problems 1.1.6 and 1.1.7.

(b) Prove that

$$I_{\bigcap A_i} = \max_i I_{A_i}, \quad I_{\bigcap A_i} = \min_i I_{A_i}.$$ 

22 Let $A_1, \ldots, A_m$ be subsets of the sample space $\mathcal{S}$; let $I_i$ be the indicator function for the set $A_i$; and let $E_j$ be the event "exactly $j$ of the events $A_1, \ldots, A_m"$'. Prove that

$$I_{E_0} = (1 - I_1) \cdots (1 - I_m)$$

$$= 1 - \sum_{i_1} I_{i_1} + \sum_{i_1 < i_2} I_{i_1} I_{i_2} - \cdots + (-1)^m I_{1 \cdots m}$$

$$= 1 - T_1 + T_2 - \cdots + (-1)^m T_m,$$

where $T_r = \sum_{i_1 < \cdots < i_r} I_{i_1} \cdots I_{i_r}$.

23 Let $\mathcal{A}_0 = \{A_i\}$ be a class of subsets $A$ on a sample space $\mathcal{S}$. Show that there exists a $\sigma$-algebra on $\mathcal{S}$ that contains $\mathcal{A}_0$.

24 (continuation) Let $\{\mathcal{A}_a\}$ be the set of all $\sigma$-algebras $\mathcal{A}_a$ such that $\mathcal{A}_0 \subseteq \mathcal{A}_a$. $\alpha$ ranges over some index set. Show that $\mathcal{A} = \cap \mathcal{A}_a$ is a $\sigma$-algebra over $\mathcal{S}$.

25 (continuation) If $\mathcal{A}_a \subseteq \mathcal{B}_a$, then we say that $\mathcal{B}_a$ is smaller than $\mathcal{B}_p$. Show that it is meaningful to speak of the smallest $\sigma$-algebra containing a class $\mathcal{A}_0$ of subsets of $\mathcal{S}$.

1.3

PROBABILITIES

Probabilities are what the probability model is all about. In this section we examine the simple mathematical framework for probabilities. Then in the next section we examine various computational rules for handling probabilities.

A MEASURING LENGTH, AREA, AND PROBABILITY

Consider the coin tossing in Sections 1.1C and 1.2C with sample space $\mathcal{S}_2 = \{(x_1, x_2)\} = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$,

where $x_i$ is the number of heads for the $i$th coin. The symmetry discussed in Section 1.1C gives a probability $1/4$ at each of the four points in $\mathcal{S}_2$. And the additivity mentioned there shows that the probability for any set of points is the sum of the values at component points.

From a slightly different point of view we can think of the four points in $\mathcal{S}_2$ as four points on the plane $R^2$; and we can think of attaching weights of value $1/4$ at each of the four points. We could then picture breaking off some rectangle or even some more complicated set and measuring its weight; we would obtain, of course, the value given by the symmetric $P$ discussed in Section 1.1C.
And in a more general way, we could think of the plane $R^2$ as being a plate of varying density. We could then break off some rectangle or even some more complicated set and measure its weight; for any set $A$, we would obtain its weight measure, say $\mu(A)$. If the plate were of uniform density and with unit weight for unit area, the measure $\mu$ gives the area for any set on the plane.

And in a simpler way we could think of the line $R$ as being a rod of varying density. We could then break off some interval or even some more complicated set and measure its weight; for any set $A$, we would obtain its weight measure, say $\mu(A)$. If the rod were of uniform density with unit weight for unit length, the measure $\mu$ gives the length of any set on the line. We shall see that measuring probability is a simple special case of such examples of measuring weight.

**B FORMALITIES OF MEASURING**

Consider a sample space $S$ with events given by a class $\mathcal{C}$ of subsets of $S$.

We will find the notion of a partition very convenient. For the present notation, we have:

**DEFINITION 1**

A class $\{A_1, A_2, \ldots\}$ of subsets of $S$ is a countable partition of a subset $A$ if $A = A_1 \cup A_2 \cup \cdots = \bigcup_{i=1}^{\infty} A_i$ and if $A_i$ and $A_j$ are disjoint for $i \neq j$.

In a sense we have taken the set $A$ and broken it into a countable number of subsets $A_1, A_2, \ldots$; see Figure 1.5. And if we think of the plane as a plate of varying density, we have taken a piece $A$ of the plane and broken it into a countable number of pieces $A_1, A_2, \ldots$. Note that the definition covers the case of a finite partition by having most of the $A_i = \emptyset$.

Now consider the formal definition of a measure $\mu$. For this we allow the value $+\infty$, as, for example, the weight of the whole plane viewed, say, as a plate of uniform thickness.

**FIGURE 1.5**

Set $A$ partitioned into $A_1$, $A_2$, $\ldots$. 
DEFINITION 2

A measure \( \mu \) is a function from \( \mathcal{G} \) into \([0, \infty]\) such that

(a) Nontrivial: \( \mu(A) < \infty \) for some \( A \) in \( \mathcal{G} \).
(b) \( \sigma \)-additive: if \( \{A_1, A_2, \ldots\} \) is a countable partition of \( A \) with \( A_i \) in \( \mathcal{G} \), then

\[
\mu(A) = \sum_{i=1}^{\infty} \mu(A_i).
\]

It is straightforward to prove that \( \mu(\emptyset) = 0 \) and to prove finite additivity:

\[
\mu(\bigcup_{i} A_i) = \sum_{i} \mu(A_i),
\]

where the \( A_i \) are disjoint; see Problems 11 and 12.

EXAMPLE 1

Coin tossing from Sections 1.1B and 1.1C: For observing only the first coin we have \( \mathcal{S} = \{0, 1\} \). Suppose that we consider the points of \( \mathcal{S} \) as points on the real line and then consider subsets of the real line. The measure \( P \) based on the 100 repetitions can be presented in terms of sets rather than the events in Section 1.1B:

\[
P(A) = 0 \quad \text{if neither } 0, 1 \in A
\]

\[
= 0.53 \quad \text{only } 1 \in A
\]

\[
= 0.47 \quad \text{only } 0 \in A
\]

\[
= 1.00 \quad \text{both } 0, 1 \in A.
\]

For observing both coins we have \( \mathcal{S}_2 = \{0, 1\}^2 \). Suppose that we consider the points of \( \mathcal{S} \) as points on the plane, and then consider subsets of the plane. The measure \( P \) based on the 100 repetitions can be presented in terms of sets on the plane; we define \( P(A) \) to be the earlier \( P(B) \) of Section 1.1C, where \( B \) is the event corresponding to the subset \( A \); see Figure 1.6.

**FIGURE 1.6**
Probability of at least one head based on 100 repetitions.
EXAMPLE 2

Cardinality: The number of elements in a set \( A \) is a measure of the set. For notation, let

\[ c(A) = \text{number of elements in } A = \text{cardinality of } A. \]

Then for a \( \sigma \)-algebra \( \mathcal{G} \) over a sample space \( \mathcal{S} \), the function \( c \) is a measure, called the \textit{cardinality measure}.

For a sample space such as \( R^1 \) or \( R^2 \) \ldots \, we have that every nonfinite set has cardinality measure \( +\infty \). We obtain a more useful measure if we count only those points that fall in some designated subset \( \mathcal{S}_0 \) of interest; accordingly, we define a measure \( \mu_1 \) by

\[ \mu_1(A) = c(A \cap \mathcal{S}_0). \]

For example, consider the plane \( R^2 \) and let \( \mathcal{S}_0 \) be the set \( S_2 \) described in the preceding example. Then for the set \( A \) in Figure 1.6, we have \( \mu_1(A) = 3 \).

EXAMPLE 3

If we think of unit weights attached to the points of \( \mathcal{S}_0 \) in the preceding example, the measure \( \mu_1 \) gives the weight for a set. It is easy to think more generally of attaching different weights at various points in a designated subset \( \mathcal{S}_0 \). Consider a finite or countable set \( \mathcal{S}_0 = \{a_1, a_2, \ldots \} \) and let \( m(a_i) \geq 0 \) be the weight attached to the point \( a_i \). Then we define a measure \( \mu_2 \) by

\[ \mu_2(A) = \sum_{a_i \in A} m(a_i). \]

where the summation is over all points \( a_i \) in \( A \). If we define \( m(s) = 0 \) for points \( s \) not in \( \mathcal{S}_0 \), we can write

\[ (1) \quad \mu_2(A) = \sum_{s \in A} m(s). \]

where only a finite or countable number of terms have \( m(s) \) different from zero and the summation is effectively over such terms. It is trivial to verify that \( \mu_2 \) satisfies the \( \sigma \)-additivity and thus satisfies the formal Definition 2 for a measure. The measure \( \mu_2 \) is called \textit{symmetric} if the weights \( m(a_i) = m \) are all equal. Note then that

\[ (2) \quad \mu_2(A) = m\mu_1(A). \]

C THREE KINDS OF MEASURE

We have been considering measures that range from one extreme with value \( +\infty \) for every set except \( \varnothing \) to near simplicity in the other direction as in Example 1. Consider some simple kinds of measures:
DEFINITION 3

A measure \( \mu \) is a \( \sigma \)-finite measure if \( S \) can be expressed as a countable union \( \bigcup_{i=1}^{\infty} A_i \) of sets \( A_i \) each having finite measure \( \mu(A_i) < \infty \).

Consider the plane \( R^2 \) as a plate of uniform density and unit weight per unit area. Then the weight measure gives the area of sets on the plane. It is \( \sigma \)-finite; for we can break the plane into a countable number of unit squares each of area 1. The cardinality function on the line or plane is not \( \sigma \)-finite. If we do not allow the value \(+\infty\) for a measure, we obtain a finite measure:

DEFINITION 4

A measure \( \mu \) is a finite measure if \( \mu(S) < \infty \).

The measure \( \mu_2 \) defined in Example 3 is a finite measure in the case that \( \sum_{i=1}^{\infty} m(a_i) < \infty \). For the coin-tossing example (Example 1) the measure for the whole space is 1:

DEFINITION 5

A measure \( \mu \) is a probability measure if \( \mu(S) = 1 \).

For a probability measure we shall usually use the letter \( P \) in place of \( \mu \).

D MEASURING PROBABILITY

From the simple coin-tossing example we can picture probability as a total weight of 1 distributed around the points of the sample space. The additivity mentioned in Section 1.1 becomes the \( \sigma \)-additivity of a measure, and the total weight of 1 is the special property for probabilities. We summarize this as:

DEFINITION 6

A probability measure \( P \) is a nonnegative function on a \( \sigma \)-algebra \( \mathcal{G} \) such that

(a) \( \sigma \)-additive: if \( \{A_1, A_2, \ldots\} \) is a countable partition of \( A \), then

\[
P(A) = \sum_{i=1}^{\infty} P(A_i),
\]

(b) Normed: \( P(S) = 1 \).

Of course for a probability measure \( P \), we have \( P(\emptyset) = 0 \); and we have finite additivity:

\[
P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i),
\]

provided that the \( A_i \) are disjoint (Problems 11 and 12). Note that we can think of the measure \( P \) as being defined on the corresponding Boolean \( \sigma \)-algebra.

We can obtain a fairly general illustration of a probability measure by specializing the measure \( \mu_2 \) in Example 3. We use the notation \( p(a_i) \) for the weight at \( a_i \) in \( S_0 \), and
we require that the sum $\sum_{i}^{n} \mu(A_i) = 1$ is equal to unity. This gives the measure $P_2$:

$$P_2(A) = \sum_{a_i \in A} \mu(a_i).$$

This is a probability measure; for as a special case of $\mu_2$ it is $\sigma$-additive; and of course $P_2(\emptyset) = \sum \mu(a_i) = 1$. The measure $P_2$ is called a discrete measure on the space $\mathcal{S}$.

**EXAMPLE 4**

*Coin tossing from Sections 1.1B and 1.1C:* For observing only the first coin, we take $\mathcal{S}_0 = \{0, 1\}$ and $\mathcal{S}$ to be the real line $R$. The measure based on the 100 repetitions is obtained by defining $\mu$ as follows:

$$\mu(0) = 0.47, \quad \mu(1) = 0.53.$$  

The measure based on symmetry is obtained by defining $\mu$ as follows:

$$\mu(0) = 1/2, \quad \mu(1) = 1/2.$$  

For observing both coins we take $\mathcal{S}_0 = \{0, 1\}^2$ and $\mathcal{S}$ to be the plane $R^2$. The measure based on the 100 repetitions is obtained by defining $\mu$ as follows:

$$\mu(0, 0) = 0.24, \quad \mu(1, 0) = 0.28,$$

$$\mu(0, 1) = 0.23, \quad \mu(1, 1) = 0.25;$$

see Figure 1.6. The measure based on symmetry is obtained by defining $\mu$ as follows:

$$\mu(0, 0) = 1/4, \quad \mu(1, 0) = 1/4,$$

$$\mu(0, 1) = 1/4, \quad \mu(1, 1) = 1/4.$$

**EXAMPLE 5**

*Binomial frequencies:* Consider some distributions of probability on the real line $\mathcal{S} = R$. For a simple example take $\mathcal{S}_0 = \{0, 1, 2\}$ and define $\mu$ by

$$\mu(0) = 1/4, \quad \mu(1) = 1/2, \quad \mu(2) = 1/4;$$

see Figure 1.7. This arises with a pair of symmetric coins if we recognize only the events $E_0$, $E_1$, and $E_2$, that is, recognize only the number of heads showing; this is easily checked from the concluding results in Example 4. For a more general example, take $\mathcal{S}_0 = \{0, 1, \ldots, n\}$ and define $\mu$ by

$$\begin{align*}
\begin{array}{c}
0 \\
\frac{1}{2} \\
\frac{1}{4} \\
\frac{1}{8} \\
\end{array} \\
0 \quad 1 \quad 2 \quad x
\end{align*}$$

**FIGURE 1.7**

Binomial$(2, 1/2)$ probabilities.
(4) \[ p(s) = \frac{n^s}{s!} \frac{n!}{(n-s)!} (1/2)^n; \]

note that \( n \) factorial is defined by \( n! = n(n-1) \cdots 2 \cdot 1 \) for positive integers \( n \) and \( n! = 1 \) for \( n = 0 \). These probabilities on \( S_0 \) add up to 1 by the binomial theorem. We will see later that this distribution occurs with \( n \) symmetric coins if we recognize only the number of heads showing. This distribution is called the binomial distribution. The same distribution arises in the context of Example 1.2.1 (drug administered to rabbits: observe life or death in a test period) if we recognize only the number of deaths occurring and if the probability of a death is 1/2. More on this in Chapter 3, when we have developed the necessary concepts.

**EXAMPLE 6**

**Poisson frequencies:** For a slightly different example on the real line, take \( S_0 = \{0, 1, 2, \ldots\} \) and define \( p \) by

\[ p(s) = \frac{\lambda^s e^{-\lambda}}{s!} \quad \text{for} \quad s \in S_0, \]

where \( \lambda \) is a nonnegative number. We check that the total is 1:

\[ \sum p(s) = e^{-\lambda} \sum_{s=0}^{\infty} \frac{\lambda^s}{s!} = e^{-\lambda} e^{\lambda} = 1. \]

This distribution arises for the physicist counting the number of \( \alpha \)-particles emitted by a radioactive source in a 7.5-second interval (mentioned in Section 1.2A); the value for \( \lambda \) would be the long-run average number of particles in a 7.5-second interval; in Section 2.1D we shall return to this illustration in more detail. This distribution is called the \( \text{Poisson}(\lambda) \) distribution and arises in many applications where we count the number of occurrences of some event in an interval of space or time: red blood cells in a square on a blood slide; accidents of a particular type in a week. See Figure 1.8 for the \( \text{Poisson}(\lambda) \) distribution with \( \lambda = 2 \).

![Figure 1.8](image)

**FIGURE 1.8**  
Poisson(\( \lambda \)) probabilities with \( \lambda = 2 \).
E SYMMETRIC PROBABILITY DISTRIBUTIONS

For the coin-tossing example in Section 1.1C, we saw how symmetry among the faces for each coin gave symmetry among the four elementary events \( H_1H_2, H_1T_2, T_1H_2, \) and \( T_1T_2. \) We will see many examples with such symmetry; in fact, almost all the traditional games of chance are based on symmetry. For an appropriate general model we now define a symmetric probability measure: a probability measure \( P \) as given by (3) is symmetric with respect to \( S_0 = \{ a_1, \ldots, a_N \} \) if the probability weights \( p(a_i) = p \) are all equal. Since \( \sum_1^N p(a_i) = 1, \) it follows that \( p(a_i) = 1/N; \) thus

\[
\bar{P}(A) = \frac{1}{N} \sum_{a \in A} \frac{1}{N} c(A \cap S_0) = \frac{c(A \cap S_0)}{c(S_0)}.
\]

Note that the probability is the number of \( S_0 \)-points in \( A \) divided by the total number of \( S_0 \)-points. This is sometimes called the classical definition of probability, owing to its origins in the traditional games of chance.

EXAMPLE 7

Tossing two symmetric dice (Example 1.2.4): Consider two symmetric dice and suppose that they are tossed symmetrically. Let \( y_i \) be the number of points showing on the \( i \)th die. The space of possible outcomes is \( S_0 = \{ y_1, y_2 : y_i = 1, \ldots, 6 \} = \{ 1, 2, 3, 4, 5, 6 \}^2. \)

There are 36 points in \( S_0 \) and the symmetry of the dice and tossing gives symmetry among the 36 points. This implies that the probability measure is the symmetric measure (6) with \( N = 36. \)

\[
\bar{P}(A) = \frac{\text{number of possible points in } A}{\text{total number of possible points}} = \frac{c(A \cap S_0)}{36}.
\]

Thus we have the probability \( 1/36 \) at each of the marked points in Figure 1.2. And for the events discussed in Example 1.2.4, we have

\[
\bar{P}(L ' ' \text{at least one } 6') = 11/36,
\]
\[
\bar{P}(T ' ' \text{total points is } 7') = 6/36,
\]
\[
\bar{P}(T \cap L) = 2/36.
\]

F THE PROBABILITY MODEL

Consider a random system. We now have the mathematical structure that is needed to describe the variation observed from performance to performance of the system.

In Section 1.2 we considered the class \( \mathcal{B} \) of events that we are interested in. The natural properties for such a class have been summarized in the requirement that \( \mathcal{B} \) be a Boolean \( \sigma \)-algebra. In practice we shall often use a sample space \( S \) as a convenience; the events can then be expressed easily as subsets \( A \) of \( S, \) subsets that belong to a \( \sigma \)-algebra \( \mathcal{G} \) over \( S. \) Of course, this convenience comes at a price: if we decide to ob-
serve more about the system, then we shall require a new sample space and a new
σ-algebra; and we will then have a different representation for each of the original events
in (B).

In the present section we have considered the function P that gives the long-run
proportion of occurrences for any event. The natural properties for such a function
have been summarized in the requirement that P be a probability measure. Our model,
then, is (B, P), where B is the Boolean σ-algebra and P is the probability measure.
(B, P) is called a Boolean probability space. If we include the sample space S, then the
model becomes (S, B, P) and is called a probability space.

G EXERCISES: GAMES OF CHANCE

1 Three symmetric coins are tossed; let x_i = 0(1) according as the i-th coin is tails (heads).
   (a) Record the effective sample space S_0 for the outcomes (x_1, x_2, x_3).
   (b) Determine the symmetric probability measure \( \overline{P} \).
   (c) Calculate \( \overline{P}(E_0) \), \( \overline{P}(E_1) \), \( \overline{P}(E_2) \), and \( \overline{P}(E_3) \), where \( E_j \) is the event "exactly j heads."

2 Tossing two symmetric dice (Example 7). Calculate the probabilities \( P(T_0) \), \( P(T_1) \), \ldots , \( P(T_{12}) \),
   where \( T_j \) is the event "total of the points is j."

3 For tossing two coins and counting the heads on each, the effective sample space is
   \( S_0 = \{ (x_1, x_2) : x_i = 0, 1 \} \). Suppose that the coins are similar but possibly biased. Then in
   Section 3.5A we shall see that the measure \( P_2 \) in (3) is appropriate with
   \[
p(x_1, x_2) = p^x q^{1-x} n^x q^{1-x}, \quad x_i = 0, 1,
\]
   where \( p \), \( q \geq 0 \) and \( p + q = 1 \).
   (a) Tabulate the values of the function p for the four sample points.
   (b) Calculate the probabilities \( P(H_1) \) and \( P(T_0) \), where \( H_1(T_0) \) are the events "heads (tails) for
       the first coin."

4 Balls in an urn (Exercise 1.2.7). An urn contains 3 green balls marked g_1, g_2, and g_3
   and 1 black ball marked b. The balls are thoroughly mixed and two balls are drawn in succession,
   yielding \( (y_1, y_2) \), where \( y_i \) is the mark on the i-th ball drawn.
   (a) Determine the symmetric probability measure.
   (b) Calculate \( P(E_0) \), \( P(E_1) \), and \( P(E_2) \), where \( E_j \) is the occurrence of exactly j black balls.

5 Cards from a deck (Exercises 1.2.8 and 1.2.9). Calculate the symmetric probability measure
   for (a) two cards without replacement (Exercise 1.2.8); (b) two cards with replacement
   (Exercise 1.2.9).

6 Calling cards (Exercise 1.2.10). Four cards are numbered 1, 2, 3, and 4. The cards are
   thoroughly shuffled and dealt face up in succession as the dealer calls successively 1, 2, 3,
   and 4.
   (a) Record the symmetric probability measure on the sample space \( S \) described earlier.
   (b) Calculate \( P(E_0) \), \( P(E_1) \), \( P(E_3) \), and \( P(E_4) \), where \( E_j \) is the event "exactly j cards called
       correctly." Another version has a secretary shuffling four letters and placing them in the
       four addressed envelopes.

H EXERCISES: INDUSTRY AND SCIENCE

7 Transformers in a lot: A lot contains 4 transformers serial marked 1, 2, 3, and 4; the first 3
   are good (g) and the last defective (b). The transformers are thoroughly mixed and 2 are
   sampled in succession (a random sample of 2).
   (a) Define an appropriate sample space having a symmetric probability measure; record the
       symmetric probability measure.
   (b) Calculate \( P(E_0) \), \( P(E_1) \), and \( P(E_2) \), where \( E_j \) is the occurrence of j defective transformers.

8 Lifetime of transistors: The lifetime \( t \) of a certain transistor is such that \( u = e^{-t/5} \) with values
in \((0, 1)\) has a very simple probability measure describing its variation: \(P((a, b)) = b - a\) for any interval \((a, b)\) contained in \((0, 1)\).
(a) Calculate \(P(0.2 < u < 0.8), P(0.4 < u < 0.6)\).
(b) Calculate the probability that the first decimal place of \(u\) is even.

9 \textit{Round-off error}: The values of a certain function are rounded off to the nearest integer. The error \(e\) is found to behave very much as the outcome of a random system that produces values in \((-0.5, 0.5)\) and has a simple probability measure: \(P((a, b)) = b - a\) for any interval \((a, b)\) contained in \((-0.5, 0.5)\).
(a) Calculate \(P(|e| < 0.3), P(|e| < 0.1)\).
(b) Calculate the probability that the first decimal place is even.
(c) Calculate the probability that the error is positive and the first decimal place is even.

10 \textit{Round-off error (continuation)}: The values for a pair of functions are rounded off to the nearest integers. The error \((e_1, e_2)\) is found to behave very much as the outcome of a random system that produces values in \((-0.5, 0.5)^2\) and has a simple probability measure: \(P(A) = \text{area of } A\) for any simple set contained in \((-0.5, 0.5)^2\).
(a) Calculate \(P(|e_1| < 0.1)\) and \(P(|e_2| < 0.1)\).
(b) Calculate \(P(\max |e_i| < 0.1)\).
(c) Calculate \(P((e_1^2 + e_2^2)^{1/2} < 0.1)\).

\section*{PROBLEMS}

11 The definition of a measure \(\mu\) required that there be a set \(A\) that has \(\mu(A) < \infty\). Use the definition of a measure to prove the "obvious" that \(\mu(\emptyset) = 0\).

12 \textit{(continuation)} The definition of a measure is in terms of \(\sigma\)-additivity. Prove finite additivity,

\[\mu(\bigcup_{i=1}^{k} A_i) = \sum_{i=1}^{k} \mu(A_i),\]

where the \(A_i\) are disjoint.

13 \textit{(continuation)} Prove monotonicity: If \(A \subseteq B\), then \(\mu(A) \leq \mu(B)\).

14 The Poisson\((\lambda)\) probability measure in Example 6 has

\[P((-\infty, b)) = \sum_{s=0}^{b} \frac{\lambda^s}{s!} e^{-\lambda} \quad \text{for } b = 0, 1, 2, \ldots .\]

Show that \(P((-\infty, b))\) is a monotone-decreasing function of \(\lambda (\lambda > 0)\). \textit{(Hint: Differentiate with respect to \(\lambda\) and simplify.)}

15 The binomial\((n, p)\) probability measure has

\[P((-\infty, b)) = \sum_{s=0}^{b} \frac{n!}{s!(n-s)!} p^s(1 - p)^{n-s} \quad \text{for } b = 0, 1, \ldots, n,\]

where \(p \in [0, 1]\). Show that \(P((-\infty, b))\) is a monotone-decreasing function of \(p\).

\section*{1.4 CALCULATING PROBABILITIES}

We have been examining the mathematical framework for probabilities. We now examine some convenient rules that are useful for calculating probabilities.
starting point is Definition 1.3.6 and we derive the formulas deductively. The properties in the definition can be summarized as (a) $\sigma$-additivity, and (b) total probability is 1.

### A ADDITIVITY

We have noted in Section 1.3 that a probability measure $P$ has the *null property* $P(\emptyset) = 0$ and the *additivity property*

$P(\bigcup_{i=1}^{k} A_i) = \sum_{i=1}^{k} P(A_i)$

provided that the $A_i$ are disjoint. The additivity property is often used in its simplest form with two component sets. For convenience we record this as a proposition:

---

### PROPOSITION 1

If $A$ and $B$ are disjoint, then

$P(A \cup B) = P(A) + P(B)$.

A special form of this additivity can be very useful:

---

### PROPOSITION 2

The probability of an event $A$ is 1 minus the probability of the complementary event:

$P(A) = 1 - P(A')$.

*Proof* The sample space $\mathcal{S}$ can be partitioned into the two sets $A$ and $A'$; thus

1 = $P(\mathcal{S}) = P(A) + P(A')$,

and the proposition follows by rearrangement.

---

### EXAMPLE 1

**Symmetric coin**: A symmetric coin is tossed 4 times and $(x_1, x_2, x_3, x_4)$ is observed, where $x_i = 0(1)$ according as the $i$th coin is tails (heads). We have symmetry among the $2^4 = 16$ sample points; thus we use the symmetric measure in Section 1.3E. The event "at least one head" has many sample points, but the complementary event is very simple:

$P(\text{at least one head}) = 1 - P(\text{all tails})$

= $1 - 1/16 = 15/16$.

---

### B PARTIAL ADDITIVITY

If two events $A$ and $B$ are not disjoint, then the formula for $P(A \cup B)$ must reflect the overlap:
PROPOSITION 3

For arbitrary events $A$ and $B$,

\[(3) \quad P(A \cup B) = P(A) + P(B) - P(AB).\]

Proof The event $A \cup B$ can be partitioned into three events, $A \cup B = AB^c \cup AB \cup A^cB$; see Figure 1.9. Additivity then gives

\[
P(A \cup B) = P(AB^c) + P(AB) + P(A^cB) \\
= [P(AB^c) + P(AB)] + [P(A^cB) + P(AB)] - P(AB) \\
= P(A) + P(B) - P(AB).
\]

Sometimes the probabilities $P(A)$, $P(B)$, and $P(AB)$ are available in an immediate and direct way:

EXAMPLE 2

Symmetric dice: Two symmetric dice are tossed and the number of points on each is observed:

\[
P(\text{at least one six}) = P(\text{first die is a 6}) + P(\text{second die is a 6}) - P(\text{both dice show 6}) \\
= 1/6 + 1/6 - 1/36 = 11/36.
\]

The preceding proposition can be generalized to cover three events, $A$, $B$, and $C$:

PROPOSITION 4

For arbitrary events $A$, $B$, and $C$:

\[(4) \quad P(A \cup B \cup C) = P(A) + P(B) + P(C) \\
- P(AB) - P(BC) - P(AC) + P(ABC).
\]

\[\text{\textsc{Figure 1.9}}\]

Event $A \cup B$ can be partitioned as $AB^c \cup AB \cup A^cB.$
This can be proved from Proposition 3, or by the method used with Proposition 3, or by a simple procedure with indicator functions to be presented in Supplement Section 1.5A.

C MONOTONICITY

Probabilities have a very simple, almost trivial, monotone property:

PROPOSITION 5

Probability is a monotone function of events: if \( A \subseteq B \), then \( P(A) \leq P(B) \).

Proof If \( A \subseteq B \), then \( B \) can be partitioned as
\[
B = A \cup A^c B,
\]
which produces
\[
P(B) = P(A) + P(A^c B) \geq P(A).
\]
Note that we used the nonnegativity of \( P(A^c B) \).

EXAMPLE 3

Serial system: A certain control system is composed of \( m \) component systems. Suppose that the full system functions if and only if each of its \( m \) components functions; such a system is called a serial system. Let \( A \) be the event "the system functions," and let \( A_j \) be the event "the \( j \)th component system functions." The serial property can then be expressed as
\[
A = \cap_j A_j.
\]
Thus \( A \subseteq A_j \); and from the proposition it follows that \( P(A) \leq P(A_j) \). This gives the rather obvious result,
\[
P(\text{system functions}) \leq P(\text{\( j \)th component functions})
\]
for arbitrary \( j \).

D MONOTONE CONTINUITY

A sequence of events \( A_1, A_2, \ldots \) is called a monotone sequence if
\[
A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots \quad \text{(monotone increasing)}
\]
or if
\[
A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots \quad \text{(monotone decreasing)}.
\]

For a monotone sequence we can define the limit set in a straightforward manner:
\[
\lim_{i \to \infty} A_i = \bigcup_i A_i \quad \text{if monotone increasing}
\]
If $A_1, A_2, \ldots$ is a monotone sequence, then

$$P(\lim_{i\to\infty} A_i) = \lim_{i\to\infty} P(A_i).$$

**Proof**  This proposition is just $\sigma$-additivity slightly disguised. First consider the increasing case. Let $B_i = A_i - A_{i-1}$ be the difference set $A_i \setminus A_{i-1}$:

- $B_1 = A_1$
- $B_2 = A_2 - A_1$
- $B_3 = A_3 - A_2$.

Note that in going from $A_{i-1}$ to $A_i$ the set of "new" points is given by $B_i$; it then follows by induction that

$$\bigcup_{i=1}^n B_i = A_n = \bigcup_{i=1}^n A_i; \quad \bigcup_{i=1}^\infty B_i = \bigcup_{i=1}^\infty A_i = \lim_{i\to\infty} A_i.$$

The sets $B_1, B_2, \ldots$, however, form a partition of $\bigcup_{i=1}^\infty B_i = \bigcup_{i=1}^\infty A_i$; thus

$$P(\bigcup_{i=1}^\infty A_i) = P(\bigcup_{i=1}^\infty B_i) = \sum_{i=1}^\infty P(B_i) = \lim_{n\to\infty} \sum_{i=1}^n P(B_i)$$

$$= \lim_{n\to\infty} P(\bigcup_{i=1}^n B_i) = \lim_{n\to\infty} P(A_n),$$

which proves the proposition for the increasing case. The proof for the decreasing case follows by complementation: if $A_1, A_2, \ldots$ is monotone decreasing, then $A_{i}', A_{i-1}', A_{i-2}', \ldots$ is monotone increasing.

**Example 4**

**Endless tossing:** A symmetric coin is tossed until the first head appears; what is the probability that the first head appears on an even-numbered toss? Let $C_i$ be the

![Figure 1.10](image.png)

Monotone-increasing sequence with limit set $A$. 

occurrence of the first head on the \(i\)th toss; let \(A_{2n} = C_2 \cup C_4 \cup \ldots \cup C_{2n}\) be the occurrence of the first head on an even-numbered toss in the first \(2n\) tosses; then the event of interest is \(\lim_{n \to \infty} A_{2n}\) and we can use Proposition 6.

To cover all possible outcomes we must consider a countable number of tosses, and to have symmetry we must consider each possibility for each toss in the sequence. For this, let \(i_j = 0(1)\) if the \(j\)th toss is tails (heads); then the sample space is

\[ S = \{ (i_1, i_2, \ldots) : i_j = 0, 1 \}; \]

note that \(i_j\) is effectively the indicator function for heads on the \(j\)th toss. This is a complicated sample space, and we do not yet have the background for describing all the subsets or being sure there is a probability for each subset.

To see how big the sample space is, consider the following. A sample point \((i_1, i_2, \ldots)\) can be represented as a binary expansion, \(0.i_1i_2 \ldots\). This maps \(S\) into \([0, 1]\) in a way that is almost one to one; the exceptions are pairs such as 0.1011 and 0.11000 \ldots, which describe the same point just as 0.599 and 0.7000 \ldots in decimal expansion describe the same point.

In spite of not yet being able to determine the full probability model, we can, however, calculate the probability for the event \(C_2\) using symmetry for the first two tosses, for \(C_4\) using symmetry for the first four tosses, and so on:

\[ P(A_{2n}) = P(C_2) + P(C_4) + \ldots + P(C_{2n}) \]

\[ = \frac{1}{2^2} + \frac{1}{2^4} + \ldots + \frac{1}{2^{2n}}. \]

\[ \lim_{n \to \infty} P(A_{2n}) = \frac{1}{4} \cdot \frac{1}{1 - 1/4} = \frac{1}{3}. \]

There are some interesting aspects to the mapping of \(S\) into \([0, 1]\). The probability is 1/2 for each of the values 0, 1 in the first position; thus the probability is 1/2 for each of the intervals \([0, 1/2], [1/2, 1]\). The probability is 1/4 for each of 00, 01, 10, 11 in the first two positions; thus the probability is 1/4 for each of the intervals \([0, 1/4], [1/4, 1/2], [1/2, 3/4], [3/4, 1]\). And so on. We thus obtain a uniform distribution of probability on the interval \([0, 1]\).

---

**E ALTERNATIVE DEFINITION OF PROBABILITY**

The continuity proposition (Proposition 6) has an important special case: if \(A_1, A_2, \ldots\) is monotone decreasing with \(\lim_{i \to \infty} A_i = \emptyset\), then

\[ \lim_{i \to \infty} P(A_i) = 0. \]  

A function \(P\) with the preceding property is called *continuous at \(\emptyset\)*.

Definition 1.3.6 for a probability measure can now be presented in an alternative form that is often convenient:
DEFINITION 7

A probability measure \( P \) is a nonnegative function on a \( \sigma \)-algebra \( \mathcal{G} \) such that

(a) Additive: if \( A_1 \) and \( A_2 \) are disjoint, then

\[
P(A_1 \cup A_2) = P(A_1) + P(A_2).
\]

(b) Continuous at \( \varnothing \): if \( A_1, A_2, \ldots \) is monotone decreasing to \( \varnothing \), then

\[
\lim_{i \to \infty} P(A_i) = 0.
\]

(c) Normalized: \( P(\mathcal{G}) = 1 \).

The results in Section D establish Definition 7 from Definition 1.3.6; Problem 17 is to establish the reverse.

Consider a monotone-decreasing sequence \( A_1, A_2, \ldots \) with \( \lim A_i = A \). Sometimes we will find that

\[
\lim_{i \to \infty} P(A_i) = 0,
\]

and yet \( A \) is not empty. Of course, by (6) we have that \( P(A) = 0 \). Such a set \( A \) may present the points at which some condition or proposition does not hold. We shall then speak of the condition holding except on a set of probability zero. Or we shall speak of the condition holding with probability 1; or holding almost surely, abbreviated a.s.

F EXERCISES: GAMES OF CHANCE

1. Consider the tossing of two symmetric dice. Calculate the probability that at least one die shows an even number; use the expression for \( P(A \cup B) \).

2. Four cards are numbered 1, 2, 3, and 4. The cards are shuffled and dealt in succession while the dealer calls 1, 2, 3, and 4 in correspondence. Calculate the probability that at least one card is called incorrectly; see Exercises 1.2.10 and 1.3.6.

3. (continuation) Calculate the probability that at least 2 cards are called incorrectly.

4. A deck of 52 playing cards is thoroughly shuffled and 2 cards are dealt in succession. Let \( A_j \) be the occurrence of an ace on the \( j \)th draw; and let \( E_j \) be the occurrence of exactly \( j \) aces. Express \( E_j \) in terms of the \( A_j \) and calculate \( P(E_0), P(E_1), \) and \( P(E_2) \); see Exercise 1.2.8.

5. A deck of 52 playing cards is shuffled; a card is dealt, observed, and returned to the deck; the deck is shuffled and a card is dealt. Let \( A_j \) be the occurrence of an ace on the \( j \)th draw; let \( E_j \) be the occurrence of exactly \( j \) aces. Express \( E_j \) in terms of the \( A_j \) and calculate \( P(E_0), P(E_1), \) and \( P(E_2) \); see Exercise 1.2.9.

6. A deck of 52 playing cards is thoroughly shuffled and dealt in succession while the dealer calls AS, KS, \ldots, 2C in succession. Calculate the probability that at least one card is called incorrectly; that exactly one card is called incorrectly.

7. Each of two players, \( A \) and \( B \), tosses a symmetric coin; a player wins if he gets heads and his opponent gets tails; if neither wins the double toss is called a draw. Two players toss repeatedly until there is a win. Calculate the probability that player \( A \) wins:

(a) Use the method in Example 4.

(b) Define a very elementary probability space and use symmetry.

8. A player tosses a symmetric coin until he obtains the first head. Calculate the probability that the first head occurs on a toss with number divisible by 3; see Example 4.
9 Each of two players tosses a symmetric coin; they are interested in a match: both show heads or both show tails. For a single trial, calculate the probability for a match. The players decide to toss repetitively until they obtain a match. Calculate the probability that a match occurs on an even-numbered trial; assume the symmetry and the results of Example 4.

G EXERCISES: INDUSTRY AND SCIENCE

10 Suppose that a shipment of 4 calculators has one defective. Suppose the calculators are thoroughly shuffled. The buyer tests the calculators one at a time. Calculate the probability that
   (a) The last one tested is the defective calculator.
   (b) The defective calculator is found before the last test. Compare with Exercise 1.3.7.

11 The error \( e_1, e_2 \) in rounding off a pair of functions is found to behave as the outcome of a random system that produces values in \((-0.5, 0.5)^2\) and has a simple probability measure: \( P(A) \) is the area of \( A \) for any simple set \( A \subset (-0.5, 0.5)^2 \). Use the results calculated in Exercise 1.3.10 to obtain the probability that
   (a) At least one of \( |e_1|, |e_2| \) is \( \geq 0.1 \).
   (b) At least one of \( |e_1|, |e_2| \) is \( < 0.1 \).
   (c) \( e_1^2 + e_2^2 \) is \( \geq 0.01 \).

12 A certain blood smear is obtained on a grid \((-5, +5) \times (-5, +5)\) on the plane. One particular cell is known to have a different color under ultraviolet light. Available theory says that the probability that the cell is in the set \( A \) is given by \( \frac{\text{area } A}{100} \) for a set \( A \subset (-5, 5)^2 \). With the appropriate lighting the microscope makes a sweep covering one unit on either side of the first axis; calculate the probability of detection. Alternatively, the microscope makes a sweep within one unit of the second axis; calculate the probability of detection. The microscope makes a sweep within one unit of each of the axes; calculate the probability of detection.

H PROBLEMS

13 Use Proposition 3 to prove Proposition 4:

\[
7 \quad P(A_1 \cup A_2 \cup A_3) = \sum_i P(A_i) - \sum_{i<j} P(A_iA_j) + P(A_1A_2A_3).
\]

14 (continuation) Let \( A_1, A_2, \) and \( A_3 \) be subsets of the sample space \( S \) and let \( E_r \) be “the occurrence of exactly \( r \) of \( A_1, A_2, \) and \( A_3.\)“ Prove the following formulas in the order \( P(E_3), P(E_2), \) and \( P(E_1):\)

\[
P(E_3) = 1 - \sum_i P(A_i) + \sum_{i<j} P(A_iA_j) - P(A_1A_2A_3),
\]

\[
P(E_2) = \sum_i P(A_i) - 2 \sum_{i<j} P(A_iA_j) + 3P(A_1A_2A_3),
\]

\[
P(E_1) = \sum_i P(A_i) - 3P(A_1A_2A_3).
\]

15 Let \( E_{(r)} \) be the event “at least \( r \) of \( A_1, \ldots, A_3.\)“ Use Problem 14 to prove the following:

\[
P(E_{(3)}) = 1.
\]
\( P(E_{11}) = \sum_{i \leq j} P(A_i) - \sum_{i < j} P(A_i A_j) + P(A_1 A_2 A_3) \).

\( P(E_{12}) = \sum_{i < j} P(A_i A_j) - 2P(A_1 A_2 A_3) \).

\( P(E_{13}) = P(A_1 A_2 A_3) \).

16 Three cards are numbered 1, 2, and 3. The cards are shuffled and dealt in succession while the dealer calls 1, 2, and 3 in correspondence. Use the results in Problem 14 to calculate \( P(E_1), P(E_2), P(E_3), P(E_4) \), where \( E_i \) is the event "exactly \( i \) cards called correctly." Of course, these probabilities are easily calculated by counting. But the formulas do illustrate the method needed for \( n \) cards; see Problems 1.5.3 and 1.5.4 in Supplement Section 1.5.

17 Show that Definition 7 implies Definition 1.3.6. (Hint: To show that finite additivity plus continuity implies \( o \)-additivity, consider a finite union together with a "tail sum": i.e., \( A_1 \cup A_2 \cup \ldots \cup A_k \cup (\bigcup_{i=k}^{\infty} A_i) \) and let \( k \to \infty \).)

18 Generalization of Problem 13: Let \( A_1, \ldots, A_m \) be subsets of the sample space \( S \). Prove by induction that

\[ P(\bigcup_{i=1}^{m} A_i) = S_1 - S_2 + S_3 - \cdots - (-1)^{m-1} S_m, \]

where \( S_r = \sum_{1 \leq r \leq m} P(A_{r_1} \cdots A_{r_k}) \). A routine mechanical proof can be obtained from the results in Supplement Section 1.5.

19 For a monotone sequence \( A_1, A_2, \ldots \) of subsets of a sample space \( S \), show that

\[ \lim_{i \to \infty} A_i = \lim_{i \to \infty} \mu_i; \]

that is, show that

\[ \lim_{i \to \infty} \mu_i(s) = \lim_{i \to \infty} \mu_i(s) \quad \text{for each} \quad s \in S. \]

Recall the abbreviated method of proof in Section 1.2D.

20 (a) Prove: if \( A_1, A_2, \ldots \) is a monotone-increasing sequence, then

\[ \mu(\lim_{i \to \infty} A_i) = \lim_{i \to \infty} \mu(A_i). \]

(b) Prove: if \( A_1, A_2, \ldots \) is a monotone-decreasing sequence and if \( \mu(A_1) \), say, is finite, then

\[ \mu(\lim_{i \to \infty} A_i) = \lim_{i \to \infty} \mu(A_i). \]

21 (continuation) Prove: if \( A_1, A_2, \ldots \) is a monotone-decreasing sequence of sets having \( \mu(A_i) < \infty \) and if \( \lim_{i \to \infty} A_i = \emptyset \), then

\[ \lim_{i \to \infty} \mu(A_i) = 0. \]

Note: Consider a monotone-decreasing sequence \( A_1, A_2, \ldots \) with \( \lim A_i = A \). Sometimes we shall find that

\[ \lim_{i \to \infty} \mu(A_i) = 0, \]

and yet \( A \) is not empty. Of course, \( \mu(A) = 0 \). Such a set \( A \) may present the points at which some condition or proposition does not hold. We shall then speak of the condition holding except on a set of measure(\mu) zero; or holding almost everywhere(\mu), abbreviated a.e.(\mu).

SUPPLEMENTARY MATERIAL

Supplement Section 1.5 is on pages 511–519; it further develops some ideas in the preceding sections and helps toward making this text self-contained.
An indicator function takes the values 0 and 1; a simple generalization, a *simple function*, takes a finite number of values. The notion of *mean value* is introduced for such simple functions; this leads to easy proofs for *some probability formulas* recorded in the problems.

The extension theorem for measures is formally recorded. Among other things, it takes the notion of length for an interval and extends it to a *length measure* on the line, and similarly takes the notion of area for a rectangle and extends it to an *area measure* on the plane.

**NOTES AND REFERENCES**

The following are some textbooks that develop introductory probability theory.


Probability theory as a branch of mathematics has been developing since the seventeenth century. The current axiomatic formulation in terms of a probability measure is due to Kolmogorov (1933); this formulation brought the needed precision to probability theory. Some theory of Boolean algebras may be found in Halmos (1963). The extension theorem for measures may be found in advanced books on measure theory, for example, Ash (1972) and Halmos (1950).


In Chapter 1 we examined the probability model and then investigated some routine methods for calculating probabilities.

A central interest of probability theory is of course the calculation of probabilities for events of interest in the real world. To approach this in detail we need to examine rather closely the standard models encountered in science, engineering, and other applications.

Most data in science, engineering, and business come in the form of real numbers. Sometimes a particular piece of data will be a measurement or determination of a physical variable with possible values on an interval of the real line. As examples of such quantitative determinations, consider the weight increase of an animal under a diet treatment, the temperature at which a certain transducer fails, and the thickness of cold-rolled steel plate. At other times a particular piece of data will be a count of how many objects in a sample of objects under investigation have some particular quality or characteristic. As examples of such qualitative data, consider the number of rabbits surviving a certain drug treatment, the number of bearings passing a simple "go, no go" test of diameter, and the number of α particles received in a 7.5-second time interval with a radioactive source.

In Chapter 2 we discuss various ways of describing a response that involves one or several real numbers. In the first two sections we examine real-valued responses: first discrete responses, such as qualitative counts, and then continuous responses, such as quantitative determinations. The third section examines vector-valued responses involving several counts or determinations. The concluding section discusses changes of variable, the process of relabeling the response variable of a system under investigation.
2
PROBABILITY ON
THE LINE AND PLANE

2.1
PROBABILITY ON THE REAL LINE: DISCRETE

Consider a real-valued response from a random system, such as the measurement of a failure temperature for a transducer from a certain production run. By our discussions in Chapter 1 we know that we are interested in events such as the intervals on $R$. Closure under the standard operations then leads to the Borel class $\mathcal{B}^1$ of sets on $R$. For the full probability description we then need the function $P$ that gives the probability for each set in the class $\mathcal{B}^1$. Actually, however, we are interested only in intervals and perhaps unions of intervals; indeed, we have no way in general of presenting an arbitrary Borel set. We first examine an alternative and simple way of presenting probability, by a distribution function that gives probabilities only for very special intervals.

A DISTRIBUTION FUNCTION ON
THE REAL LINE

Consider a distribution of probability on the real line $R$ as given by a measure; for example, this could be the distribution describing breaking load for a new design of concrete beam. We now define a distribution function that presents probability only for the special subsets $(-\infty, y)$, that is, from $-\infty$ up to and including a point $y$:

DEFINITION 1

The distribution function $F$ is a real-valued function on $R$ such that

(1) $F(y) = P((-\infty, y])$

at each point $y$.

The distribution function is sometimes called the cumulative distribution function.

If we have the distribution function $F$, we can calculate the probability for any interval $(a, b]$:

$P([a, b]) = P((-\infty, b]) - P((-\infty, a])$

$= F(b) - F(a)$. 
For notation it is convenient to use \((a, b]\) as an operator that says "take a difference" in the obvious way; thus

\[(2) \quad P([a, b]) = (a, b]F = F(b) - F(a).\]

We can calculate the probability at a point \(b\) by first calculating it for an interval \((b - \delta, b]\) and then taking the limit as \(\delta\) goes to zero:

\[P((b - \delta, b]) = F(b) - F(b - \delta),\]

\[P([b]) = \lim_{\delta \to 0} P((b - \delta, b]) = F(b) - F(b - 0),\]

where we designate the limit from the left as

\[F(b - 0) = \lim_{\delta \to 0} F(b - \delta) \quad \text{with} \quad \delta > 0,
\]

and we have used the continuity of Proposition 1.4.6. Again for notation it is convenient to use \([b - 0, b]\) as an operator; thus

\[(3) \quad P([b]) = (b - 0, b]F = F(b) - F(b - 0).\]

**EXAMPLE 1**

**Heads with two symmetric coins:** Consider tossing two symmetric coins and suppose that only the number \(y\) of heads is recorded: see Example 1.3.5. The possible outcomes are 0, 1, and 2 with corresponding probabilities 1/4, 1/2, and 1/4. The distribution function can be calculated easily:

\[F(y) = \begin{cases} 
0 & \text{if } -\infty < y < 0, \\
1/4 & 0 \leq y < 1, \\
3/4 & 1 \leq y < 2, \\
1 & 2 \leq y < \infty;
\end{cases}\]

see Figure 2.1. The probability on the interval \((0.5, 1.3]\) is

\[P((0.5, 1.3]) = F(1.3) - F(0.5) = 3/4 - 1/4 = 1/2 = P(\{1\}).\]

This example has probability at the discrete points 0, 1, and 2. A natural

![Figure 2.1](image-url)

**FIGURE 2.1**

Distribution function for the number of heads when two symmetric coins are tossed.
generalization has probability \( p(y) \) at points \( y \) in a finite or countable set \( S_0 \); see the measure \( P_2 \) in Section 1.3D.

**EXAMPLE 2**

**Orientation with two symmetric coins:** Consider tossing two symmetric coins and recording the angle that the line joining them makes (measured positively) with the front edge of the table; any angle from \( 0^\circ \) to \( 180^\circ \) is possible. Symmetry gives, for example,

\[
P((0, 10]) = P((10, 20]) = \cdots = P((170, 180]) = 1/18,
\]

\[
P((0, 5]) = P((5, 10]) = \cdots = P((175, 180]) = 1/36;
\]

thus

\[
F(y) = P([\leq y])
\]

\[
= 0 \quad \text{if} \quad -\infty < y < 0,
\]

\[
= y/180 \quad \quad 0 \leq y < 180,
\]

\[
= 1 \quad \quad 180 \leq y < \infty;
\]

see Figure 2.2. The probability on the interval \((-10, 80]\) is

\[
P((-10, 80]) = F(80) - F(-10) = 80/180 - 0 = 4/9
\]

\[
= P((0, 80]).
\]

This example has probability spread uniformly over the interval \((0, 180]\) at the rate or density \(1/180\) per unit length. A natural generalization has probability on \( \mathbb{R} \) spread at the rate \( f(y) \) per unit length at the point \( y \); this gives probability \( f(y) \, dy \) for a small interval of length \( dy \) at the point \( y \), gives probability \( \int_c^d f(y) \, dy \), the area under \( f \) between \( c \) and \( d \), for the interval \((c, d)\), and, accordingly, gives probability 0 at a mathematical point.

**B PROPERTIES OF A DISTRIBUTION FUNCTION**

Consider a distribution of probability on the real line \( \mathbb{R} \) as given by the measure \( P \). Let \( F \) be the corresponding distribution function. We shall prove the following necessary properties of a distribution function \( F \):

![Figure 2.2](image)

**FIGURE 2.2**
Distribution function \( F \) for the angle when two symmetric coins are tossed.
1. Monotone: \( F(b) - F(a) = (a, b]F \geq 0 \) for \( b > a \).

2. Right continuous: \( F(y + 0) = F(y) \) for all \( y \).

3. Normed: \( F(-\infty) = 0, F(+\infty) = 1 \).

The proofs follow from the various properties that we have established for the probability measure \( P \). For property 1 we note that \( F(b) - F(a) \) is the probability for the interval \((a, b]\) and is thus \( \geq 0 \). For property 2 we consider the interval \((y, y + \delta]\) with \( \delta > 0 \) and note that \( \lim_{\delta \to 0} (y, y + \delta] = \emptyset \); thus \( F(y + 0) - F(y) \) is equal to \( P(\emptyset) = 0 \) by the continuity Proposition 1.4.6. And for property 3 we note that \((-\infty, y]\) has limit \( \emptyset \) as \( y \to -\infty \) and limit \( \mathbb{R} \) as \( y \to +\infty \); thus \( F(-\infty) = 0, F(\infty) = 1 \) by the same continuity property.

We now argue that properties 1–3 are also sufficient for a real-valued function \( F \) on \( \mathbb{R} \) to be the distribution function of some measure \( P \). We can define a function \( P \) for any interval \([a, b]\) by

\[
P([a, b]) = (a, b]F = F(b) - F(a),
\]

and property 1 ensures that this is \( \geq 0 \). And for a finite union of disjoint intervals, the measure would be the sum of such values. The Extension Theorem recorded in Supplement Section 1.5D then gives a unique extension of \( P \) to all the sets in \( \mathcal{B}^1 \). Thus the distribution function \( F \) uniquely determines the probability measure \( P \) on the Borel sets. This is the same kind of argument as is used to extend the notion of length from intervals to the Borel sets on \( \mathbb{R} \); see Supplement Section 1.5E.

We now have two ways of presenting a distribution of probability on the sample space \( \mathbb{R} \): by means of the measure \( P \) or by means of the distribution function \( F \). It is convenient, then, to have the general term distribution and to speak of a distribution with measure \( P \) or a distribution with distribution function \( F \).

For notation we shall usually use capital letters such as \( F, G, H, \ldots \) to designate distribution functions. Sometimes subscripts will be added: \( F_1, G_3, \ldots \). And sometimes, if the probability on \( \mathbb{R} \) has come from somewhere else by means of a real-valued function, say \( Y \), we will use that function as a subscript: \( F_Y \).

### C DISCRETE COMPONENT OF A DISTRIBUTION

In Example 1 we saw that the points 0, 1, and 2 received chunks of probability 1/4, 1/2, and 1/4. In Example 2 we saw that the probability was spread out continuously so that the probability at any mathematical point was zero. We now examine more formally the occurrence of chunks of probability.

Consider a distribution on the real line with measure \( P \) and distribution function \( F \). Let \( p \) be the function that gives probability at a point:

**DEFINITION 2**

The probability function \( p \) is a real-valued function on \( \mathbb{R} \) such that

\[
p(y) = P(\{y\}) = F(y) - F(y - 0)
\]

at each point \( y \).
The probability function \( p \) is, in effect, a restriction of \( P \) to certain special sets, the sets that contain a single point. In Example 1 the function \( p \) has the values 1/4, 1/2, and 1/4 at the points 0, 1, and 2 and the value 0 elsewhere. In Example 2, \( p \) has the value zero everywhere.

A point at which \( p(y) > 0 \) is called a discrete point or, more exactly, a point having discrete probability. Consider the set of points having discrete probability:

\[
S_d = \{ y : p(y) > 0 \}.
\]

It is straightforward to show that \( S_d \) has a finite or at most countable number of points; see Problem 16.

Now let \( F_d \) be a "distribution function" for the discrete probabilities; specifically, let \( F_d \) record the total discrete probability for any interval \((-\infty, y]\):

\[
F_d(y) = P((-\infty, y] \cap S_d) = \sum_{t \leq y} p(t).
\]

If all the probability is at discrete points, then \( P(S_d) = 1 \) and \( F_d(+\infty) = 1 \); and if all the probability is spread continuously, then \( P(S_d) = 0 \) and \( F_d(+\infty) = 0 \).

**D DISCRETE DISTRIBUTIONS**

A distribution is called "discrete" if all the probability is at discrete points in the manner just described. This is the kind of distribution that is useful for the qualitative counts mentioned in the preface to this section; for example, the distribution describing the number of \( \alpha \) particles in a 7.5-second interval.

**DEFINITION 3**

A distribution is **discrete** if its measure and distribution function have the forms

\[
P(A) = \sum_{y \in A} p(y), \quad F(y) = \sum_{t \leq y} p(t),
\]

where \( p(y) \) is given by Definition 2.

This is the measure \( P_2 \) in Section 1.3D, specialized to the real line \( R \).

With a discrete distribution it is straightforward to calculate the probability for a set—sum the probabilities at component points.

Consider a simple example.

**EXAMPLE 3**

**The Bernoulli**\( (p) \) **distribution**: For the tossing of a possibly biased coin, let \( x \) be the number of heads observed. The effective sample space for \( x \) is \( \{0, 1\} \), and all the probability must occur at the points 0 and 1 in the enlarged space \( S = R \). Let

\[ p(0) = P(\{0\}) = q, \quad p(1) = P(\{1\}) = p. \]
where \( p, q \geq 0 \) and \( p + q = 1 \), and let \( p(x) = 0 \) for \( x \neq 0, 1 \). This gives a very simple discrete distribution; see Figure 2.3.

The example presents the Bernoulli distribution in the context of a classical game of chance, coin tossing. The applications for this distribution are, however, wide-ranging. Consider the administration of a drug to a certain kind of animal and the observation of, say, survival for a specified time interval; the mass production of an electronic component and the observation for a single component of, say, good (as opposed to defective) quality; the production of \( \alpha \)-particles by a radioactive source and the observation of, say, "no particles" in a 7.5-second interval of time.

Now consider a discrete distribution with probability function \( p \). The following are necessary properties of a probability function \( p \); they follow from Definition 3.

1. Nonnegative: \( p(y) \geq 0 \) for all \( y \).
2. Discrete: \( p(y) = 0 \) except at a finite or countable number of points.
3. Normal: \( \sum p(y) = 1 \).

Property 2 follows in a sense from 3 because the summation is meaningful only if there are at most a countable number of points with \( p(y) > 0 \); see Problem 16.

We now note that the properties 1–3 are also sufficient for a real-valued function \( p \) on \( R \) to be the probability function of a discrete distribution on \( R \). This follows by noting that the properties give a function \( p \) as used in defining the discrete probability measure \( P_2 \) in Section 1.3D.

**EXAMPLE 4**

The Poisson(\( \lambda \)) distribution: Consider the Poisson distribution as described briefly in Example 1.3.6:
\[ p(y) = \frac{\lambda^y}{y!} e^{-\lambda} \quad \text{for } y = 0, 1, 2, \ldots, \]
\[ = 0 \quad \text{otherwise}; \]
\[ P(A) = P(y \in A) = \sum_{y \in A} p(y) = \sum_{y \in A} \frac{\lambda^y}{y!} e^{-\lambda} \quad \text{for nonnegative integers } y. \]

The Poisson probability function with \( \lambda = 2 \) is plotted in Figure 1.8; the values are best obtained by desk calculator or computer: \( p(0) = 0.1353, p(1) = 0.2707, p(2) = 0.2707, p(3) = 0.1804, \ldots \) For example,
\[ P(-\infty < y \leq 1.5) = 0.1353 + 0.2707 = 0.4060, \]
\[ P(0.5 < y \leq 2) = 0.2707 + 0.2707 = 0.5414, \]
\[ P(0.5 < y < 2) = 0.2707. \]

Recall from Example 1.3.6 that the Poisson distribution is useful for applications involving frequency counts of the occurrence of some event in an interval of space or time: blood cells on a slide, accidents in a week, and so on.

---

In the preceding example, note that we are using \( y \) to form events for a particular Poisson distribution; for example, the event \( 0.5 < y \leq 2 \) in the general sense of Section 1.1.4 or the event \( \{ y: 0.5 < y \leq 2 \} \) in the subset sense of Section 1.2.C; the corresponding probability is
\[ P(0.5 < y \leq 2) = P(\{ y: 0.5 < y \leq 2 \}) = 0.5414. \]

This is notationally very convenient—to use a particular letter, typically lowercase, for constructing sets on a designated space with a designated distribution. Note that when the letter \( is \) used, it is used only to present an event or to construct a subset as part of a probability calculation.

---

**E LARGE-SAMPLE DATA FROM A DISCRETE DISTRIBUTION**

Rutherford and Geiger took 2608 independent observations on the number \( y \) of \( \alpha \)-particles in a 7.5-second interval; see Table 2.1. As scientists they were interested in many things but certainly in the pattern of variation for the response \( y \). There are some theoretical reasons for thinking that the Poisson distribution would properly describe this variation. Accordingly, they would be interested in the appropriate Poisson distribution, that is, in estimating the true value for the parameter \( \lambda \).

The column \( f_i \) records the frequency for each value \( y_i \) of the number of \( \alpha \)-particles in a 7.5-second interval. The column \( \hat{p}_y \) records the corresponding proportion or estimated probability. Thus, without any assumption concerning a pattern among the probabilities, we would estimate the probability for, say, \( y = 2 \) as 0.1469.

Now suppose that we accept the Poisson distribution for the number \( y \); the theoretical grounds for this will be examined in Supplement Section 4.3. We are then faced with choosing a value for \( \lambda \) that makes the Poisson probabilities close in some
TABLE 2.1
Observations of $\alpha$-particles in 7.5 second interval

<table>
<thead>
<tr>
<th>Sample point $y_i$</th>
<th>Frequency $f_i$</th>
<th>Estimated probability $\hat{p}_i$</th>
<th>Estimated Poisson probability $\hat{\lambda}_i$</th>
<th>Expected Poisson frequency $e_i = 2608\hat{\lambda}_i$</th>
<th>Standardized residual $d_i = 2(\sqrt{f_i} - \sqrt{e_i})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>57</td>
<td>0.0219</td>
<td>0.0209</td>
<td>64.4</td>
<td>0.35</td>
</tr>
<tr>
<td>1</td>
<td>203</td>
<td>0.0778</td>
<td>0.0807</td>
<td>210.5</td>
<td>-0.52</td>
</tr>
<tr>
<td>2</td>
<td>383</td>
<td>0.1469</td>
<td>0.1562</td>
<td>407.3</td>
<td>-1.23</td>
</tr>
<tr>
<td>3</td>
<td>525</td>
<td>0.2013</td>
<td>0.2015</td>
<td>525.4</td>
<td>-0.02</td>
</tr>
<tr>
<td>4</td>
<td>532</td>
<td>0.2040</td>
<td>0.1950</td>
<td>508.4</td>
<td>1.03</td>
</tr>
<tr>
<td>5</td>
<td>408</td>
<td>0.1564</td>
<td>0.1509</td>
<td>393.5</td>
<td>0.72</td>
</tr>
<tr>
<td>6</td>
<td>273</td>
<td>0.1047</td>
<td>0.0973</td>
<td>253.9</td>
<td>1.18</td>
</tr>
<tr>
<td>7</td>
<td>139</td>
<td>0.0533</td>
<td>0.0538</td>
<td>140.4</td>
<td>-0.11</td>
</tr>
<tr>
<td>8</td>
<td>45</td>
<td>0.0172</td>
<td>0.0260</td>
<td>67.9</td>
<td>-3.06</td>
</tr>
<tr>
<td>9</td>
<td>27</td>
<td>0.0104</td>
<td>0.0112</td>
<td>29.2</td>
<td>0.42</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>0.0038</td>
<td>0.0043</td>
<td>11.3</td>
<td>-0.40</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>0.0015</td>
<td>0.0015</td>
<td>4.0</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>0.0008</td>
<td>0.0005</td>
<td>1.3</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0.0000</td>
<td>0.0002</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>11-14\textsuperscript{a}</td>
<td>(6)</td>
<td>(0.0023)</td>
<td>(0.0022)</td>
<td>(5.8)</td>
<td>(0.08)</td>
</tr>
<tr>
<td>2608</td>
<td>1.0000</td>
<td>1.0000</td>
<td>2608.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\textsuperscript{a}For comparing frequencies with expected frequencies it is conventional to group together cells with 'expected' frequencies less than 5, thus the combined range 11-14.

Sense to the observed proportions; this is essentially a statistical problem and the theoretical aspects will be examined in Chapter 9. For the present, however, we can argue rather directly to an appropriate value.

The mean or long-run average value of $\gamma$ for the Poisson($\lambda$) distribution is just $\lambda$; see Problem 20 in Supplement Section 1.5. The Rutherford and Geiger data have a total of

$$0 \times 57 + 1 \times 203 + 2 \times 383 + \ldots = 10,094$$

$\alpha$-particles for a total of 2608 time intervals. The average number of $\alpha$-particles for a 7.5-second interval is thus

$$\frac{10,094}{2,608} = 3.87.$$  

This is an average value for 2608 repetitions; as such it should provide a reasonable approximation for the mean or long-run average value $\lambda$. Accordingly, we will use $\hat{\lambda} = 3.87$ as an estimate of the true value $\lambda$. The estimated Poisson probabilities based on $\hat{\lambda} = 3.87$ are recorded in the column $\hat{p}_i$ in Table 2.1. Thus on the basis of the Poisson pattern we would estimate the probability for, say, $\gamma = 2$ as 0.1562. Note, of course, that this estimate is based not just on the frequency at the value 2 but on all the frequencies in the table.

Parenthetically, we can note that the average can be calculated for other time intervals. For example, the total count 10,094 occurred with an overall time interval of 2608 $\times$ 7.5 = 19,560.0 seconds. The average value for a 1-second interval is thus...
\[
\frac{10,094}{19,560} = \frac{3.87}{7.5} = 0.516,
\]
and for a 2-second time interval is 1.032.

Now suppose that we want to check whether the Poisson pattern is appropriate. In part, this is a question of fit and is examined in detail in Section 10.4. For the present we record the calculations and defer the justifications until Chapter 10. We can compare the observed frequencies \( f_i \) with "expected" frequencies \( e_i = 2608 \hat{p}_i \)

obtained by apportioning 2608 according to the proportions \( \hat{p}_i \). The comparison with expectation is made in suitable units by calculating the standardized residual

\[
d_i = 2(\sqrt{f_i} - \sqrt{e_i});
\]

some justifications for this will be examined in Chapter 6. The values of \( d_i \) are recorded in the right-hand column. If a Poisson distribution describes the counting process, then, subject to minor qualifications, the deviations will fall in the intervals \((-1, +1)\), \((-2, +2)\), and \((-3, +3)\) with probabilities 68\% 95\% and 99\% percent approximately. The observed values seem moderately in accord with this distribution, thus lending some reassurance to the use of the Poisson pattern.

An overall test of the residuals is also available from Section 10.4; the details would go as follows. The test is made by calculating

\[
\chi^2 = \sum d_i^2 = 14.60
\]

and comparing with a chi-square distribution (Appendix II) with 10 degrees of freedom (12 residuals, less 1 because \( \Sigma e_i = \Sigma f_i \), less 1 because the parameter \( \lambda \) has been fitted to the data). The observed value 14.60 is a reasonable value for this chi-square distribution, thus giving some overall assurance to the use of the Poisson pattern.

**F Exercises**

1. A symmetric die is tossed and the number \( y \) of points is recorded. Let \( F \) be the distribution function for \( y \).
   
   (a) Present the distribution function \( F \) and make a sketch of it.
   
   (b) Use the distribution function to calculate the probabilities for the following events:
   
   \( 3 < y \leq 5 \); and \( 4 < y \leq 7 \).

2. (continuation) For a suitable price a gambling house will pay \( x = y^2 \) on a roll of the die. Let \( G \) be the distribution function for \( x \).

   (a) Present the distribution function \( G \) and make a sketch of it.

   (b) Calculate the probabilities for the following events: \( 9 < x \leq 25 \); and \( 16 < x \leq 49 \).

3. For the Poisson(1) distribution, calculate the values of \( p(y) \) successively; use \( e^{-1} = 0.3679 \).

   Calculate \( P(y < 1) \); \( P(y = 1) \); and \( P(y > 1) \).

4. At a certain time of day the number of incoming calls for a 1-minute interval is approximately Poisson(2). Calculate the probability that less than 2 calls occur in a 1-minute interval; that exactly 2 occur in a 1-minute interval; and that more than 2 occur in a 1-minute interval.

5. In a certain city the number of fatal accidents is approximately Poisson with an average of 1 per week. For a 3-week period, calculate the probability for zero fatal accidents; for less than 3; and for 3 or more.

6. The number of fish caught by a fisherman at a certain location is approximately Poisson, with an average of 1 per hour. Calculate the probability that a 3-hour fishing period produces zero fish.
7 The number of defects of a certain kind in steel cable is approximately Poisson, with an average of 2 per 1000 feet. Calculate the probability of zero defects in a 250-foot coil; and 2 or more defects in a 250-foot coil.

**G EXERCISES**

Some common discrete distributions are recorded in these exercises together with an indication of various applications. The reader should become familiar with general characteristics of each. The grounds for using some of these in applications will be examined in Section 3.5.

8 The uniform \(\{1, 2, \ldots, N\}\) distribution has probability function

\[ p(y) = \frac{1}{N} \quad \text{if } y = 1, 2, \ldots, N \]

and zero otherwise; see Figure 2.4.

\[ \begin{align*}
    & 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \\
    \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6
\end{align*} \]

**FIGURE 2.4**

Uniform distribution on \(\{1, 2, \ldots, 6\}\).

*Applications:* A symmetric die produces a score \(y\); \(N = 6\). A computer can generate random bits, 0, 1, with equal probability \(1/2\); \(N = 2\), and the value 2 is identified with 0. A computer can generate random digits, 0, \ldots, 9, with equal probability \(1/10\); \(N = 10\), and the value 10 is identified with 0. Computer generation uses arithmetic operations and cyclic effects can occur, thus voiding the constant conditions presupposed for a random system.

(a) Present the distribution function \(F\).
(b) Calculate the probability for the following intervals: \((-\infty, N/2]\) and \((N/4, 3N/4]\).

9 The geometric\((p)\) distribution has probability function

\[ p(y) = pq^y \quad \text{if } y = 0, 1, 2, \ldots \]

and zero otherwise, where \(p > 0\), \(q \geq 0\) and \(p + q = 1\); see Figure 2.5.

\[ \begin{align*}
    & 0 \quad 1 \quad 2 \quad 3 \quad 4 \\
    \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
0 & 1 & 2 & 3 & 4
\end{align*} \]

**FIGURE 2.5**

Geometric\((1/2)\) distribution: number of tails until the first head.
Applications: A coin with probability \(p\) for heads is tossed until a head occurs; \(y\) is the number of tails obtained. A subject tries a test having probability \(p\) of success; without learning, \(y\) is the number of failures until the first success.
(a) Show that \(\rho(y)\) is a probability function for a discrete distribution.
(b) Calculate the distribution function and sketch it for \(p = 2/3\).

The **binomial** \((n, p)\) distribution has probability function

\[
p(y) = \frac{n!}{y!(n-y)!} p^y q^{n-y} \quad \text{if } y = 0, 1, \ldots, n
\]

and zero otherwise, where \(p, q \geq 0\) and \(p + q = 1\); note that \(\rho(0) = 1\) for all values of \(p\); see Figure 2.6.

**FIGURE 2.6**

Binomial\((4, 1/2)\) and \((4, 3/4)\) distributions.

Applications: A symmetric coin is tossed \(n\) times and the number \(y\) of heads is counted; \(p = 1/2\). Each of \(n\) rabbits is separately subjected to a drug treatment having probability \(p\) of success; \(y\) is the number of successes. Each of \(n\) electronic systems has a probability \(p\) of failure; \(y\) is the number of failures.
(a) Tabulate \(\rho(y)\) for the case \(n = 4, p = 1/2\); calculate the probability for the event \(1 \leq y \leq 3\).
(b) Tabulate \(\rho(y)\) for the case \(n = 4, p = 3/4\); calculate the probability for the event \(1 \leq y \leq 3\).

11 The **negative binomial** \((r, p)\) distribution has probability function

\[
p(y) = \frac{(y + r - 1)!}{y!(r - 1)!} p^r q^y \quad \text{if } y = 0, 1, \ldots
\]

and zero otherwise, where \(r\) is a positive integer and \(p > 0, q \geq 0, p + q = 1\); see Figure 2.7.

**FIGURE 2.7**

Negative binomial\((2, 1/2)\) distribution.

Applications: A die is tossed until the \(r\) aces occur; \(y\) is the number of nonaces obtained \(\{p = 1/6\}\). Separate components have a probability \(p\) of being good; \(y\) is the number of defectives until \(r\) good components have been obtained.
(a) Show that \( r = 1 \) gives the geometric(\( p \)) distribution.
(b) Tabulate the probability function for the case \( r = 2, p = 1/2 \).

The **hypergeometric**(\( N, n, D/N \)) **distribution** has probability function

\[
p(y) = \frac{n!}{y!(n-y)!} \frac{D!}{(D-y)!} \frac{(N-D)!}{(N-D-n+y)!} \frac{(N-n)!}{N!}
\]

for integer values such that \( \max (0, n - N + D) \leq y \leq \min (n, D) \) and zero otherwise; see Figure 2.8.

**FIGURE 2.8**

Hypergeometric(8, 4, 1/2) and (8, 4, 3/4) distributions.

**Applications:** An urn contains \( D \) black balls and \( N - D \) white balls; the balls are mixed and \( n \) balls are drawn; \( y \) is the number of black balls. A lot contains \( D \) defectives and \( N - D \) good items; the items are mixed and \( n \) items are drawn; \( y \) is the number of defectives.

(a) Tabulate \( p(\cdot) \) for \( N = 8, n = 4, \) and \( D/N = 4/8 \); calculate the probability that \( 1 \leq y \leq 3 \).

(b) Tabulate \( p(\cdot) \) for \( N = 8, n = 4, \) and \( D/N = 6/8 \); calculate the probability that \( 1 \leq y \leq 3 \).

### PROBLEMS

13 Define \( G_d \) by \( G_d(y) = F_d(y)/p_d \) provided that \( p_d = F_d(\pm \infty) \neq 0 \). Show that \( G_d \) is a distribution function for a discrete distribution.

14 Let \( S_c = \{ y : F \text{ is continuous at } y \} \); and define \( F_c \) by \( F_c(y) = P(S_c \cap (-\infty, y)) \). Show that \( F_c \) is a continuous function.

15 (continuation) Define \( G_c(y) = F_c(y)/p_c \) provided that \( p_c = F_c(\infty) \neq 0 \). Show that \( G_c \) is a distribution function. Note that \( F(y) = p_d G_d(y) + p_c G_c(y) \); thus the distribution function has been decomposed into a mixture of discrete and continuous components.

16 Show that the set \( S_d \) in Section C is finite or countable. For this, let \( A_n = \{ y : p(y) > 1/n \} \) and prove that \( S_d = \bigcup_n A_n \).

### 2.2

**PROBABILITY ON THE REAL LINE:**

**ABSOLUTELY CONTINUOUS**

We now discuss the distribution for a response that can take any value over a continuous range. Recall the quantitative determinations mentioned in the preface to this chapter; as an example, consider the distribution that describes the weight increase of an animal under a diet treatment.
A ABSOLUTELY CONTINUOUS
COMPONENT OF A DISTRIBUTION

First, consider in general a distribution on the real line with measure \( P \) and
distribution function \( F \). Let \( f \) be a function that gives probability density at a point:

DEFINITION 1

The probability density function \( f \) is a real-valued function on \( R \) defined by

\[
f(y) = \frac{dF(y)}{dy} = F'(y) \quad \text{if } F'(y) \text{ exists}
\]

and is equal to zero otherwise.

For the number of heads when two coins are tossed (Example 2.1.1), we have \( f(y) = 0 \)
everywhere. And for the orientation when two coins are tossed (Example 2.1.2), we
have \( f(y) = 1/180 \) on \((0, 180)\) and = 0 elsewhere.

A point at which \( F'(y) \) exists is called a point having probability density.

Consider the set of all points having probability density:

\[
S_a = \{ y : F'(y) \text{ exists} \},
\]

where the subscript \( a \) denotes absolutely continuous. The set \( S_a \) is most of the real line
\( R \). The complementary set \( R - S_a \), where \( F'(y) \) does not exist, contains, of course, the
set \( S_a \) of discrete points and, indeed, it may contain other points, but it is not large;
advanced analysis shows that it has length measure zero (for some details, see
Supplement Sections 1.5E and 1.5F).

Now let \( F_a \) be a "distribution function" for the absolutely continuous probability;
specifically, let \( F_a \) record the total absolutely continuous probability for any interval
\((-\infty, y]\):

\[
F_a(y) = P((-\infty, y] \cap S_a) = \int_{-\infty}^{y} f(t) \, dt.
\]

We can justify the expression in terms of an integral by using ordinary calculus
provided that \( F'(y) \) is continuous. In general, however, we would need the concept of
the general integral and some related analysis. Supplement Section 5.8 contains a
brief introduction to the general integral. For all our examples, however, we will be
able to use the ordinary integral.

If all the probability is at points having probability density, then \( P(S_a) = 1 \) and
\( F_a(+\infty) = 1 \); in this case the absolutely continuous component accounts for the full
distribution.

B ABSOLUTELY CONTINUOUS
DISTRIBUTIONS

A distribution is called "absolutely continuous" if all the probability is at points
having probability density. This is the kind of distribution that is useful for the
quantitative determinations discussed in the preface to this chapter.
DEFINITION 2

A distribution is **absolutely continuous** if its measure and distribution function have the forms

\[ P(A) = \int_A f(y) \, dy, \quad F(y) = \int_{-\infty}^y f(t) \, dt, \]

where \( f(y) \) is given by Definition 1.

With an absolutely continuous distribution it is straightforward to calculate the probability for a set—integrate over the set. For example, to calculate the probability for an interval \((c, d)\), we determine the area \( \int_c^d f(y) \, dy \) over the interval \((c, d)\) and under the density \( f \); see Figure 2.9. And we would obtain the same value for each of the intervals \([c, d)\), \((c, d]\), and \([c, d]\), for clearly the area over a mathematical point such as \(c\) or \(d\) is zero.

EXAMPLE 1

**Uniform\((a, b)\) distribution:** The uniform distribution on the interval \((a, b)\) is defined by

\[ f(y) = \frac{1}{b - a} \quad \text{if} \quad a < y < b \]

and zero otherwise; see Figure 2.10. Note that \( f \) is a multiple, \( 1/(b - a) \), of the indicator function \( I_{(a,b)} \) of the interval \((a, b)\). The distribution function \( F \) is

\[ F(y) = \int_{-\infty}^y f(t) \, dt = 0 \quad \text{if} \quad -\infty < y \leq a, \]

\[ = \begin{cases} \frac{y - a}{b - a} & a < y < b, \\ 1 & b \leq y < \infty. \end{cases} \]

Probabilities are easily calculated; for example, with an interval \((c, d)\) contained in \((a, b)\),
\begin{equation*}
P((c, d)) = \int_c^d f(y) \, dy = \int_c^d \frac{1}{b - a} \, dy = \frac{d - c}{b - a},
\end{equation*}

which is, of course, just the length of the interval \((c, d)\) as a proportion of the length of the full interval; see Exercises 1.3.8 and 1.3.9.

For applications of this distribution, consider the following. The orientation when two coins are tossed [Example 2.1.2 with \((a, b) = (0, 180)\)]. The round-off error with certain functions (Exercise 1.3.9); with round-off to the nearest integer, \((a, b) = (-0.5, 0.5)\). Certain exponential expressions for lifetime in the no-aging case; Exercise 1.3.8 with \(u = e^{-u/5}\) and \((a, b) = (0, 1)\). The binary expansion \(0, i_1i_2i_3\ldots = \Sigma_1^\infty i_j2^{-j}\); where \(i_1, i_2, \ldots\) are the indicators for heads in a succession of tosses with a symmetric coin (Example 1.4.4); or where \(i_1, i_2, \ldots\) are random bits (Exercise 2.1.8); here \((a, b) = (0, 1)\); the distribution is approximate for a finite expansion such as \(0, i_1i_2i_3i_4i_5\). The decimal expansion \(0, i_1i_2i_3\ldots = \Sigma_1^\infty i_j \cdot 10^{-j}\), where \(i_1, i_2, \ldots\) are random digits (Exercise 2.1.8). The theoretical basis for some of these applications will be available from Section 3.3.

Now consider an absolutely continuous distribution with probability density function \(f\). The following are necessary properties of a density function \(f\); they follow from Definitions 1 and 2.

1. Nonnegative: \(f(y) \geq 0\) for all \(y\).
2. Measurable: \(\{y : f(y) \leq c\} \in \mathcal{B}^1\) for all \(c\).
3. Normed: \(\int_{-\infty}^{\infty} f(y) \, dy = 1\).

Property 1 follows from Definition 1. Property 3 says that \(F_a(+\infty) = 1\). These are the basic conditions for an absolutely continuous distribution. Property 2 says that an interval condition on \(f\) must be an event in the Borel class \(\mathcal{B}^1\); the proof is beyond our
present scope; throughout the text and problems we will assume that this property holds.

Properties 1–3 are also sufficient for a function $f$ to define an absolutely continuous distribution. This follows easily by noting that

$$\int_{-\infty}^{y} f(t) \, dt$$

satisfies the conditions for a distribution function as recorded in Section 2.1B. The expressions for $P(A)$ and $F(y)$ in Definition 2 are not affected if we change $f$ at a point, or on a countable set of points, or even on a more general set of length measure zero; formally this needs the theory of the general integral. For our purposes we assume that $f$ is a function that does give probability density, as discussed at the beginning of this section.

C NORMAL DISTRIBUTION

Many quantitative responses produce values that concentrate around a general level for the variable and drop off rather sharply in a bell-shaped curve such as in Figure 2.11. The normal distribution has traditionally been used as an approximation for such a pattern of variation. For example, it seems to approximate very closely the controlled laboratory measurement of a physical constant, say the gravitational constant. If coordinates are taken relative to the general location and are appropriately scaled, then the normal distribution has the density function

$$g(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2};$$

this is the standard normal distribution, and its density function $g$ is plotted in Figure 2.11. Property 1 for a density is immediate; property 3 is somewhat complicated but follows easily from calculations in Section 2.3E.

Probabilities for the standard normal distribution can be obtained from the standard normal distribution function $G$, which is tabulated in Appendix II:

$$P((-1, 1)) = \int_{-1}^{1} g(z) \, dz = 2(G(1) - G(0)) = 68.26\%,$$

$$P((-2, 2)) = \int_{-2}^{2} g(z) \, dz = 2(G(2) - G(0)) = 95.45\%,$$

$$P((-3, 3)) = \int_{-3}^{3} g(z) \, dz = 2(G(3) - G(0)) = 99.73\%;$$

these calculations use the symmetry of $g$ about the origin.

Now suppose that we relabel the variation $z$ by the transformation

$$y = \mu + \sigma z,$$

which scales it by the factor $\sigma$ $(>0)$ and locates it at the general level $\mu$. The standard normal then becomes the general normal, which can describe a response with values located at a general level $\mu$ and scaled about $\mu$ by the factor $\sigma$; see Figure 2.11.

Let $f$ be the density function and $F$ the distribution function for this relabeled response variable $y$. The event that the original variable is less than or equal to a value
$g = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$

$G(z) = F(y)$

\[
\frac{1}{\sqrt{2\pi}} e^{-z^2/2} \, dz = \frac{1}{\sqrt{2\pi \sigma}} e^{-(y-\mu)^2/2\sigma^2} \, dy
\]

\[
f = \frac{1}{\sqrt{2\pi \sigma}} e^{-(y-\mu)^2/2\sigma^2}
\]

\[
\frac{1}{\sigma}
\]

**FIGURE 2.11**

Standard normal density function $g$ and density function $f$ for the general normal located at $\mu$ and scaled by $\sigma$.

$z$ is the same as the event that the new variable is less than or equal to the value $y = \mu + az$. Solving for $z$ gives $z = (y - \mu)/\sigma$. Thus we obtain the distribution function

\[
F(y) = G\left(\frac{y - \mu}{\sigma}\right) = \int_{-\infty}^{(y-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \, dt
\]

\[
= \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi \sigma}} e^{-(s-\mu)^2/2\sigma^2} \, ds,
\]

where $z = (y - \mu)/\sigma$ becomes $t = (s - \mu)/\sigma$ for the variables in the integration. Then by differentiation we obtain the density function

\[
f(y) = \frac{1}{\sqrt{2\pi \sigma}} e^{-(y-\mu)^2/2\sigma^2} = \frac{1}{\sqrt{2\pi \sigma}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}.
\]

This is called the *normal* $(\mu, \sigma)$ distribution, the normal located at $\mu$ and scaled by $\sigma$ or, as we shall see later, the normal with *mean* $\mu$ and *standard deviation* $\sigma$.

Perhaps the most convenient way to remember an absolutely continuous distribution is in terms of the *probability differential,*
\[
\frac{1}{\sqrt{2\pi}} e^{-z^2/2} \, dz = \frac{1}{\sqrt{2\pi} \sigma} e^{-(y-\mu)^2/2\sigma^2} \, dy;
\]

this represents the probability in an increment \( dz \) for the original variable \( z \) and in the corresponding increment \( dy \) for the relabeled variable \( y \); see Figure 2.11. Note that the density changes by the factor \( 1/\sigma \) and length changes by the factor \( \sigma \); these compensate each other, giving the same area above corresponding increments. The differential can be interpreted as a linear form; or it can be viewed as a formal piece of an integral used for calculating probabilities.

Probabilities for the general normal are easily obtained from the tables for the standard normal; consider an example.

**EXAMPLE 2**

A psychologist's measuring instrument produces a response \( y \) that is normal(500, 100) for a certain population of students. Find the probability that a measurement \( y \) is less than or equal to 600:

\[
P(y \leq 600) = P\left( \frac{y - 500}{100} \leq \frac{600 - 500}{100} \right)
\]

\[
= P(z \leq 1) = G(1) = 0.8413.
\]

Note that \( y \) is used to form events for the measurement variable and \( z \) is used to form events for the standardized variable describing the variation.

**D LARGE-SAMPLE DATA FROM AN ABSOLUTELY CONTINUOUS DISTRIBUTION**

In an application we can always think of taking a large number of repetitions, recording the frequency for some event of interest, and then dividing by the number of repetitions to obtain the proportion or estimated probability for the event. However, if we are interested in the density function at some point \( y \), we cannot fruitfully follow the preceding procedure because the probability at a point is zero. But we can take small intervals and use the procedure to estimate the corresponding probabilities.

In practice, values will occur to three decimal places, say; or to five decimal places; or within intervals if the real line has been partitioned into short intervals or cells. Of course, then the distribution is technically a discrete distribution. But it is usually more convenient to remain with the continuous distribution as a reasonable approximation and to acknowledge that there may be effects due to round-off to the cell midpoints.

Grumell and Dunningham have examined the percent ash in 250 samples of coal from a particular source. The values are recorded in Table 2.2 in terms of frequencies for various cells on the range of the variable; some cells with low frequencies are combined and recorded at the beginning and at the end of the table.

The column \( \hat{p} \) records the proportion or estimated probability for each of the given cells. These estimated probabilities are recorded in Figure 2.12 as rectangular areas over the cells; the resulting figure is called a histogram. Thus, without any assumption concerning a pattern among the probabilities, we would estimate the probability for the interval (16.00, 16.99) as 0.156. Note that this is not an estimate of the density
<table>
<thead>
<tr>
<th>Interval or cell</th>
<th>Frequency</th>
<th>Proportion or estimated probability $\hat{p}$</th>
<th>Estimated normal probability $\tilde{p}$</th>
<th>Expected normal frequency $e = 250\tilde{p}$</th>
<th>Standardized residual $2(\sqrt{e} - \sqrt{\hat{p}})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(9.00–12.99)</td>
<td>(16)</td>
<td>0.064</td>
<td>0.064</td>
<td>16.01</td>
<td>-0.00</td>
</tr>
<tr>
<td>9.00–9.99</td>
<td>1</td>
<td>0.004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.00–10.99</td>
<td>3</td>
<td>0.012</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.00–11.99</td>
<td>3</td>
<td>0.012</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12.00–12.99</td>
<td>9</td>
<td>0.036</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13.00–13.99</td>
<td>13</td>
<td>0.052</td>
<td>0.062</td>
<td>15.64</td>
<td>-0.70</td>
</tr>
<tr>
<td>14.00–14.99</td>
<td>27</td>
<td>0.108</td>
<td>0.096</td>
<td>24.62</td>
<td>0.59</td>
</tr>
<tr>
<td>15.00–15.99</td>
<td>28</td>
<td>0.112</td>
<td>0.128</td>
<td>31.88</td>
<td>-0.71</td>
</tr>
<tr>
<td>16.00–16.99</td>
<td>39</td>
<td>0.156</td>
<td>0.147</td>
<td>36.85</td>
<td>0.35</td>
</tr>
<tr>
<td>17.00–17.99</td>
<td>42</td>
<td>0.168</td>
<td>0.148</td>
<td>37.02</td>
<td>0.79</td>
</tr>
<tr>
<td>18.00–18.99</td>
<td>34</td>
<td>0.136</td>
<td>0.128</td>
<td>32.08</td>
<td>0.33</td>
</tr>
<tr>
<td>19.00–19.99</td>
<td>19</td>
<td>0.076</td>
<td>0.097</td>
<td>24.28</td>
<td>-1.14</td>
</tr>
<tr>
<td>20.00–20.99</td>
<td>14</td>
<td>0.056</td>
<td>0.064</td>
<td>15.88</td>
<td>-0.49</td>
</tr>
<tr>
<td>21.00–21.99</td>
<td>10</td>
<td>0.040</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22.00–22.99</td>
<td>4</td>
<td>0.016</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23.00–23.99</td>
<td>3</td>
<td>0.012</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.00–24.99</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25.00–25.99</td>
<td>1</td>
<td>0.004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(21.00–25.99)</td>
<td>(18)</td>
<td>0.072</td>
<td>0.065</td>
<td>16.34</td>
<td>0.40</td>
</tr>
<tr>
<td>250</td>
<td>1.00</td>
<td></td>
<td>250.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The function itself but rather of the integral of the density over the particular interval. Typically, with a larger number of repetitions it would be possible to slightly reduce the cell width. Continuing in this way it is possible in a limiting sense to estimate the density function itself, assuming continuity.

Now suppose that there is some background information to support a normal distribution for the response $y$. We are then faced with choosing values for $\mu$ and $\sigma$ that make the normal probabilities close to the observed proportions; this is essentially a statistical problem, and the theoretical aspects will be examined in Chapter 10. For the present, however, we record some informal discussion in support of the appropriate calculations.

First we remark that $\mu$ can be interpreted as the long-run average for the normal variable $y$ and $\sigma^2$ as the long-run average of $(y - \mu)^2$; details will be given in Chapter 5. For the $n = 250$ repetitions, the average value (based on cell midpoints) is $\bar{y} = \frac{\sum y_i}{n} = 17.015$ and the average squared deviation is

$$s_y^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2 = 6.97 = (2.64)^2;$$

the use of $n - 1$ rather than $n$ compensates for the use of $\bar{y}$ rather than the true (unknown) $\mu$; details will be given in Chapters 5 and 7. The normal probabilities based on $\mu = 17.015$ and $\sigma = 2.84$ are recorded as $\hat{p}$ in the table. Thus, on the basis of the
normal pattern of variation, we would estimate the probability for the interval (16.00, 16.99) as 0.147. Note that this estimate is based not just on the frequency for the interval but on all the frequencies in the table.

Now suppose that we want to check whether the normal pattern is appropriate. This is a question of fit and the details are examined in Section 10.4. For the present we record the calculations and defer the justifications until Chapter 10. We can compare the observed frequencies $f$ with the expected frequencies

$$e = 250\hat{\rho}$$

by calculating the standardized residuals

$$d = 2(\sqrt{f} - \sqrt{e}).$$

The values of $d$ are recorded in the right-hand column of Table 2.2. If we have fitted the true distribution for variation, the residuals will behave approximately as values for the standard normal distribution. The observed values seem quite reasonable for the standard normal and we thus have some reassurance in the use of the normal distribution for the response $y$.

An overall test of the residuals is also available from Section 10.4; the details would go as follows. The test is made by calculating

$$\chi^2 = \sum d_i^2 = 3.89,$$

and comparing with a chi-square distribution (Appendix II) with 7 degrees of freedom (10 deviations, less 1 because $\sum e_i = \sum f_i$, less 2 because two parameters $\mu$ and $\sigma^2$ have been fitted to the data). The observed value 3.89 is a reasonable value for this chi-square distribution, thus giving some overall assurance to the use of the normal pattern.
E EXERCISES

1 Consider the function $G$ given by

$$G(y) = \begin{cases} 
0 & \text{if } -\infty < y < 0, \\
y^2 & 0 \leq y < 1, \\
1 & 1 \leq y < \infty.
\end{cases}$$

(a) Show that $G$ is a distribution function and sketch it.
(b) Calculate the probability density function $g$; it is sometimes called the triangular distribution.
(c) Calculate the median $\xi$, the value that has $1/2$ the probability on each side of it.

2 A psychologist’s measuring procedure gives a response that is normal $(500, 100)$.
(a) Calculate the probabilities: $P(350 < y < 650)$; $P(y < 500)$; and $P(y < 400)$.
(b) Use the differential to calculate $P(599 < y < 600)$.\)

3 A manufacturer produces incandescent bulbs that have a normal $(1100, 100)$ lifetime. He guarantees that his bulbs have a lifetime of 1000 hours. Calculate the proportion of bulbs that will fail the guarantee.

4 A biologist has found that a certain diet treatment gives a weight increase in pounds that is normal $(5, 2.2)$. Calculate the probability that the increase is more than 6 pounds; that the increase is negative.

5 The brain weight in grams for a certain racial group is approximately normal $(1400, 125)$. Calculate the probability that an individual chosen at random has a brain weight between 1300 and 1500; and between 1200 and 1600.

F EXERCISES

Some of the common distributions are recorded in these exercises and the subsequent problems. The reader should become familiar with general characteristics of each. Some interrelations will be examined later. In each case assume the measurability property 2 for a density function.

6 Standard exponential distribution has density function

$$g(z) = e^{-z} \quad \text{if } 0 < z < \infty$$

and zero otherwise; see Figure 2.13 with $\theta = 1$.
(a) Verify that $g$ is a density function.
(b) Determine the distribution function $G$.
(c) Calculate $\bar{z}$ such that $G(\bar{z}) = 1/2$. This value $G^{-1}(1/2)$ is called the median or half-life.

7 The exponential($\theta$) distribution has density function

$$f(y) = \theta^{-1}e^{-y/\theta} \quad \text{if } 0 < y < \infty$$

and zero otherwise; see Figure 2.13.

Applications: Some electronic and industrial components show no signs of aging yet have a constant failure rate; the lifetime is exponential($\theta$), where $\theta$ is the average lifetime. The time interval between successive $\alpha$-particles in a radioactive disintegration is exponential($\theta$), where $\theta$ is the average interval.
(a) Verify that $f$ is a density function.
(b) Determine the distribution function $F$.
(c) Calculate the half-life $\bar{\gamma}$; see Exercise 6.
(d) Show that the exponential (1) distribution is obtained if the standard exponential $z$ is relabeled as $y = \theta z$; use the approach involving events as in Section C.
8  The standard Cauchy distribution (the Witch of Agnesi) has density function

\[ g(z) = \frac{1}{\pi(1 + z^2)}; \]

see Figure 2.14.

\[ g \]

\[ .5 \]

\[ .3 \]

\[ .1 \]

\[ 0 \quad 5 \]

\[ -5 \quad -4 \quad -3 \quad -2 \quad -1 \quad 0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \]

FIGURE 2.14
Standard Cauchy distribution.

Applications: The Cauchy distribution for variation has a bell-shaped curve like the normal but has more probability in the tails of the distribution. Some response variables produce extreme values much more frequently than the normal; the Cauchy provides an extreme pattern for such variation.

(a) Verify that \( g \) is a density function.

(b) Calculate the distribution function \( G \).

(c) The general Cauchy \( (\mu, \alpha) \) distribution is obtained by the relabeling \( y = \mu + \alpha z \).
Determine the density function \( f \) for \( y \); use the approach involving events as in Section C.
The Laplace or double exponential distribution has density function
\[ g(z) = \frac{1}{2} \exp \left( -\frac{|z|}{\lambda} \right); \]
see Figure 2.15.

**FIGURE 2.15**
Laplace distribution.

*Applications:* This distribution is primarily of mathematical interest. However, it can provide a tentative distribution for variation with more probability in the tails than the normal, but the cusp at the origin makes it somewhat unnatural.

(a) Verify that \( g \) is a density function.
(b) Calculate the distribution function \( G \).

10 The *Pareto(\alpha) distribution* has the density function
\[ g(z) = \alpha(1 + z)^{-\alpha - 1} \quad \text{if} \quad 0 < z < \infty \]
and zero otherwise; see Figure 2.16.

**FIGURE 2.16**
Pareto(\( \alpha \)) distribution: \( \alpha = 1, 2 \).
Applications: In a population the distribution of incomes above a threshold value. The distribution of the amounts of insurance claims above a threshold value.
(a) Verify that $g$ is a density function; $a > 0$.
(b) Calculate the distribution function $G$.
(c) Determine the median $\bar{z}$.

The **standard logistic distribution** has density function

$$g(z) = e^{-z}(1 + e^{-z})^{-2} = (1/4) \text{sech}^2(\frac{z}{2});$$

see Figures 2.17 and 2.25.

![Figure 2.17](image)

Standard logistic distribution.

**Applications:** The logistic has a bell-shaped curve like the normal but has slightly more probability in the tails; often used in place of the normal if a simple distribution function is needed. The logistic distribution function is often used to describe growth: $z$ is time in appropriate units and $G(z)$ is attained height. Or to describe chemical reactions: $z$ is time and $G(z)$ is the amount of a product, a product that catalyzes the reaction.
(a) Verify that $g$ is a density function.
(b) Calculate the distribution function $G$.
(c) Determine the median $\bar{z}$.

The **standard lognormal($\tau$) distribution** has density function

$$h(z) = \frac{1}{\sqrt{2\pi} \tau} \exp\left(-\frac{[\ln(z)]^2}{2\tau^2}\right) \text{ if } 0 < z < \infty$$

and zero otherwise; see Figure 2.18.

**Applications:** The particle size in naturally occurring mixtures. In certain cases, the critical dose of a drug just causing a reaction.
(a) Show that $h$ is a density function; $\tau > 0$.
(b) Express the distribution function in terms of the distribution function $G$ for the standard normal.
(c) Determine the median $\bar{z}$.

The **standard Weibull($\beta$) distribution** has density function

$$g(z) = \beta z^{\beta-1} \exp(-z^\beta) \text{ if } 0 < z < \infty$$

and zero otherwise; see Figure 2.19.

**Applications:** The breaking strength of certain materials. The lifetime of certain components when the conditions for the exponential are not fulfilled.
(a) Show that $g$ is a probability density function; $\beta > 0$.
(b) Calculate the distribution function $G$.
(c) Determine the median $\bar{z}$. 
The **standard extreme value distribution** has density function

\[ g(z) = e^{-z} \exp \left( -e^{-z} \right), \quad -\infty < z < \infty; \]

see Figure 2.20.

**Applications:** The largest value of a response over a period of time: rainfall, earthquakes, flood flows.

(a) Show that \( g \) is a probability density function.

(b) Calculate the distribution function.

(c) Determine the median \( \bar{Z} \).
Sec. 2.2: Probability on the Real Line: Absolutely Continuous

**FIGURE 2.20**
Standard extreme-value distribution.

**PROBLEMS: DISTRIBUTIONS RELATED TO THE NORMAL**

15 The *gamma*(*$\rho$*) distribution has density function

\[ h(u) = \Gamma^{-1}(\rho) e^{-u^2/2} \quad \text{if} \quad 0 < u < \infty \]

and zero otherwise. The gamma function $\Gamma(\cdot)$ is defined in Problem 21; see Figure 2.21.

**FIGURE 2.21**
Gamma($\rho$) distribution for $u$ with $\rho = 1.5, 2, 3$ and chi-square($m$) distribution for $\chi^2$ with $m = 3, 4, 6$. 
Applications: The lifetime of certain components when the conditions for the exponential are not fulfilled. With repetitions $z_1, \ldots, z_m$ from the standard normal we will see that $u = \chi^2 / 2 = (z_1^2 + \cdots + z_m^2) / 2$ is gamma$(m/2)$. Use the properties of the gamma function (Problem 21) to show that $g$ is a probability density function; $\rho > 0$.

16 The chi-square$(m)$ distribution has density function

$$g(\chi^2) = \Gamma^{-1} \left( \frac{m}{2} \right) e^{-\chi^2 / 2} \left( \frac{\chi^2}{2} \right)^{(m/2) - 1} \cdot \frac{1}{2} \quad \text{if } 0 < \chi^2 < \infty$$

and zero otherwise, where $\chi^2$ designates the variable and $m$ is usually a positive integer; see Figure 2.21. Percentage points are available from Appendix II.

Applications: With repetitions $z_1, \ldots, z_m$ from the standard normal, we will see that $\chi^2 = z_1^2 + \cdots + z_m^2$ is chi-square$(m)$.

(a) Use properties of the gamma function (Problem 21) to show that $g$ is a probability density function; $m > 0$.

(b) Use the approach involving events in Section C to show that $u = \chi^2 / 2$ is gamma$(m/2)$.

17 The beta$(\rho, q)$ distribution has density function

$$g(u) = \frac{\Gamma(\rho + q)}{\Gamma(\rho)\Gamma(q)} u^{\rho - 1}(1 - u)^{q - 1} \quad \text{if } 0 < u < 1$$

and zero otherwise; see Figure 2.22.

**FIGURE 2.22**
Beta$(\rho, q)$ distribution: $(\rho, q) = (2, 2), (2, 4), (4, 2), (4, 4)$.

Applications: With repetitions $z_1, z_2, \ldots$ from the standard normal, we will see that $u = \Sigma_n^m z_i^2 / \Sigma_i^{m+n} z_i^2$ is beta$(m/2, n/2)$. Use properties of the beta function (Problem 22) to show that $g$ is a probability density function; $\rho, q > 0$.

18 The canonical $F(m, n)$ distribution has density function

$$h(G) = \frac{\Gamma \left( \frac{m + n}{2} \right)}{\Gamma \left( \frac{m}{2} \right) \Gamma \left( \frac{n}{2} \right)} \frac{G^{(m/2) - 1}}{1 + G^{(m+n)/2}} \quad \text{if } 0 < G < \infty$$

and zero otherwise; see Figure 2.23.
Applications: With repetitions $z_1, z_2, \ldots$ from the standard normal, we will see that $G = \frac{\sum z_i^2}{\sum z_i^2/n} z_i^2$ is canonical $F(m, n)$, and $F = (\sum z_i^2/m) / (\sum z_i^2/n)$ is standard $F(m, n)$ described below.

Show that $h$ is a probability density function; $m, n > 0$. Use the transformation $G = u/(1 - u)$ and results from Problem 17. Note that the approach involving events in Section C shows that $F = (n/m)G$ has the standard $F(m, n)$ distribution with density function

$$g(F) = \frac{\Gamma\left(\frac{m + n}{2}\right)}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{m}{n} F\right)^{\left(m+n\right)/2} F_{\alpha} \quad \text{if } 0 < F < \infty$$

and zero otherwise. Percentage points are recorded in Appendix II.

The canonical Student$(n)$ distribution has density function

$$h(T) = \frac{\Gamma\left(\frac{n + 1}{2}\right)}{\Gamma\left(\frac{1}{2}\right) \Gamma\left(\frac{n}{2}\right)} (1 + T^2)^{-\left(n+1\right)/2} \quad -\infty < T < \infty.$$

Applications: With repetitions $z, z_1, \ldots, z_n$ from the standard normal, we will see that $T = z / (\sum z_i^2/n)^{1/2}$ is canonical Student$(n)$ and $t = z / (\sum z_i^2/n)^{1/2}$ is standard Student$(n)$ as defined below. The Student$(n)$ also provides a pattern for variation with more probability in the tails than the normal.

(a) Show that $\int_0^\infty h(T) \, dT = \int_0^\infty h(T) \, dT$.

(b) Show that $h$ is a probability density function; $n > 0$. Use the transformation $G = T^2$ and the results from Problem 18. Note that the approach involving events in Section C shows that $t = \sqrt{n}T$ has the standard Student$(n)$ distribution with density function

$$g(t) = \frac{\Gamma\left(\frac{n + 1}{2}\right)}{\Gamma\left(\frac{1}{2}\right) \Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2} \frac{1}{\sqrt{n}}.$$
see Figure 2.24. Percentage points are recorded in Appendix II. The Student\( (n) \) distribution is symmetrical and bell-shaped like the normal, but it has more probability in the tails; it ranges from the Cauchy for \( n = 1 \) to the standard normal as \( n \to \infty \); see Figure 2.25.

**FIGURE 2.24**
Student\( (n) \) distribution with \( n = 1, 3, 5, 10, 20, 50, \infty \).

**FIGURE 2.25**
Normal\((0, 1)\), Student\((6)\), Student\((1)\), and standard logistic distributions.

**PROBLEMS**

20. The failure or mortality rate for a distribution is defined by \( r(t) = f(t)/(1 - F(t)) \).
   (a) Give an explanation for the name.
Sec. 2.3: Probability on the Plane

(b) Show that the failure rate for the standard exponential is 1.
(c) Show that the failure rate for the standard Weibull(\(\beta\)) distribution is \(\beta t^{\beta-1}\).

21 The gamma function \(\Gamma\) is defined by

\[
\Gamma(p) = \int_0^\infty e^{-t} t^{p-1} \, dt
\]

for \(p > 0\). Show that

(a) \(\Gamma(1) = 1\).
(b) \(\Gamma(p + 1) = p\Gamma(p)\) by integrations by parts.
(c) \(\Gamma(n + 1) = n!\) for \(n = 1, 2, \ldots\); thus the \(\Gamma\) function provides a continuous version of the factorial function.
(d) Use the normed property of the normal to show that \(\Gamma(1/2) = \sqrt{\pi}\).

22 The beta function \(B\) is defined by

\[
B(p, q) = \int_0^1 u^{p-1}(1 - u)^{q-1} \, du
\]

for \(p, q > 0\). Show that \(B(p, q) = \Gamma(p)\Gamma(q)/\Gamma(p + q)\) by making the change of variable \(t = x + y, u = x/(x + y)\) in the double integral

\[
\Gamma(p)\Gamma(q) = \int_0^\infty \int_0^\infty e^{-x-y} x^{p-1}y^{q-1} \, dx \, dy.
\]

2.3 PROBABILITY ON THE PLANE

With many random systems we are interested in responses that have several real-valued components, say \(y_1, \ldots, y_k\). For example, with a diet treatment we could measure \(y_1\), the weight increase, and \(y_2\), the blood cholesterol level. Or with a psychology investigation we could measure \(y_1\), the time to reaction, and \(y_2\), the intensity of the reaction.

The obvious sample space for a \(k\)-dimensional response \(\mathbf{y} = (y_1, \ldots, y_k)^T\) is

\[R^k = \{(y_1, \ldots, y_k) : y_i \in \mathbb{R}\},\]

the \(k\)-fold Cartesian product of the reals.

On the space \(R^k\) we are clearly interested in rectangles such as

\[(a_1, b_1) \times (a_2, b_2) \times \cdots \times (a_k, b_k),\]

which would represent the intersection of the propositions \(a_1 < y_1 \leq b_1, \ldots, a_k < y_k \leq b_k\). The smallest class of sets containing the rectangles and closed under the standard operations is the Borel class \((\mathcal{B})^k\). Now consider a distribution on \(R^k\) as given by a measure \(\mathcal{P}\) on the Borel sets. In this section we examine some alternative and simpler ways of describing the distribution on \(R^k\).

A DISTRIBUTION FUNCTION ON \(R^k\)

We first define a distribution function that presents probability only for certain special subsets, the lower corners or lower rectangles of the form
(-\infty, y_1] \times \cdots \times (-\infty, y_k].

For this it is convenient to use vector notation \( y' = (y_1, \ldots, y_k) \) and write
\[ (-\infty, y] = (-\infty, y_1] \times \cdots \times (-\infty, y_k]. \]

The distribution function is effectively the restriction of \( P \) to subsets of the preceding form:

**DEFINITION 1**

The distribution function \( F \) is a real-valued function on \( R^k \) such that
\[
F(y) = P((-\infty, y_1] \times \cdots \times (-\infty, y_k]) = P((-\infty, y]).
\]

Consider two simple examples.

**EXAMPLE 1**

**Heads with two symmetric coins:** Consider tossing two symmetric coins and observing \((x_1, x_2)\), where \(x_i\) is the indicator for heads on the \(i\)th coin. From symmetry we have probability \(1/4\) at each of the four possible sample points; the distribution function can then easily be calculated:
\[
F(x_1, x_2) = \begin{cases} 
0 & -\infty < x_1 < 0 \quad \text{or} \quad -\infty < x_2 < 0, \\
1/4 & 0 \leq x_1 < 1 \quad \text{and} \quad 0 \leq x_2 < 1, \\
1/2 & 0 \leq x_1 < 1 \quad 1 \leq x_2 < \infty, \\
1/2 & 0 \leq x_2 < 1 \quad 1 \leq x_1 < \infty, \\
1 & 1 \leq x_1 < \infty \quad 1 \leq x_2 < \infty.
\end{cases}
\]

This is plotted in Figure 2.26.

**EXAMPLE 2**

**Orientation with two symmetric coins:** Consider a first and second tossing of two symmetric coins and recording of \((y_1, y_2)\), where \(y_i\) is the angle that the line joining them makes with the front edge of the table on the \(i\)th toss; see Example 2.1.2. Symmetry gives, for example,
\[
P([0, 10] \times (0, 10]) = P([0, 10] \times (10, 20]) = \cdots.
\]

and so on with finer partitions into squares. It follows, then, that
\[
F(y_1, y_2) = \begin{cases} 
0 & -\infty < y_1 < 0 \quad \text{or} \quad -\infty < y_2 < 0, \\
\frac{y_1 y_2}{180^2} & 0 \leq y_1 < 180 \quad \text{and} \quad 0 \leq y_2 < 180.
\end{cases}
\]
FIGURE 2.26
Distribution function $F$ and probability function $p$ for Example 1.

\[
= \frac{y_1}{180} \quad 0 \leq y_1 < 180 \quad 180 \leq y_2 < \infty,
\]

\[
= \frac{y_2}{180} \quad 180 \leq y_1 < \infty \quad 0 \leq y_2 < 180,
\]

\[
= 1 \quad 180 \leq y_1 < \infty \quad 180 \leq y_2 < \infty.
\]

This distribution has probability spread uniformly over the square $(0, 180] \times (0, 180]$ at the rate of $1/180^2$ per unit area; see Figure 2.27.

Now consider how the distribution function can be used to calculate some simple probabilities. For this we let $(a_1, b_1]_1$ be the difference operator in Section 2.1A as applied to the first argument of $F$:

\[
(a_1, b_1]_1 F = F(b_1, y_2, \ldots, y_k) - F(a_1, y_2, \ldots, y_k)
\]

\[
= P((a_1, b_1] \times (-\infty, y_2] \times \cdots \times (-\infty, y_k]).
\]

Similarly, let $(a_2, b_2]_2$ be the difference operator $(a_2, b_2]$ as applied to the second argument of $F$: 
FIGURE 2.27
Distribution function $F$ and probability density function $f$ for Example 2.

\[(a_1, b_1)\cap (a_2, b_2) \subset F(b_1, b_2, y_3, \ldots, y_k) - F(a_1, b_2, y_3, \ldots, y_k)
- F(b_1, a_2, y_3, \ldots, y_k) + F(a_1, a_2, y_3, \ldots, y_k);
= P([a_1, b_1] \times [a_2, b_2] \times (-\infty, y_3] \times \cdots \times (-\infty, y_k]);

the probability expression at this second stage follows directly from the probability expression at the first stage. Note that the order in which the operators are applied is irrelevant. Similarly, let $(a_i, b_i)_i$ be the corresponding operator for the $i$th argument. Then using the notation

\[(a, b) = (a_1, b_1) \cdots (a_k, b_k),

we can write

\[ (a, b)F = \sum_{y=a,b} (-1)^{n(y)}F(y_1, \ldots, y_k)
= P([a, b]),

where the summation is over the $2^k$ terms obtained by replacing each $y$ by an $a$ or a $b$, and $n(y)$ is the number of $a$'s in the argument. Thus by differencing $F$ we can obtain the probability for any left-open right-closed rectangle.
B CHARACTERIZATION OF A DISTRIBUTION FUNCTION

The following are necessary properties of a distribution function \( F \); they can be proved in a straightforward manner.

1. Generalized monotone: \((a, b]F \geq 0\) for all rectangles \((a, b]\).
2. Right continuous: \(F(y + 0) = F(y)\) for all \(y\).
3. Normed: \(F(-\infty, y_2, \ldots, y_k) = \cdots = F(y_1, \ldots, y_{k-1}, -\infty) = 0\)
   \(F(\infty, +\infty, \ldots, +\infty) = 1\).

Concerning the proofs, note the following: for property 1 the difference \((a, b]F\) denotes the probability in \((a, b]\). For property 2,

\[
F(y + 0) = \lim_{\delta \to 0} F(y_1 + \delta_1, \ldots, y_k + \delta_k),
\]

where \(\delta_1, \ldots, \delta_k > 0\); and the set \((-\infty, y + \delta] - (-\infty, y]\) is a surface shell for the rectangle \((-\infty, y]\) and is monotone decreasing to \(\emptyset\) as \(\delta\) decreases to 0. And for property 3 the set \((-\infty, y_1] \times (-\infty, y_2] \times \cdots \times (-\infty, y_k]\) is monotone decreasing to \(\emptyset\) as, say, \(y_1 \to -\infty\), and is monotone increasing to \(\mathbb{R}^k\) as all the \(y_i \to \infty\).

The properties 1–3 are also sufficient for a function \(F\) to be the distribution function of a probability measure \(P\). For certainly we can calculate the "probability" \(P\) for any left-open right-closed rectangle. The Extension Theorem recorded in Supplement Section 1.5D then gives a unique extension of \(P\) to all the sets in \(\mathcal{B}^k\); compare with Section 2.1B.

C DISCRETE DISTRIBUTIONS

Consider a distribution over \(\mathbb{R}^k\) with measure \(P\) and distribution function \(F\):

DEFINITION 2

The probability function \(p\) is a real-valued function on \(\mathbb{R}^k\) such that

\[
(3) \quad p(y) = P(\{y\}) = (y - 0, y]F.
\]

Thus in Example 1 we have \(p(y) = 1/4\) at the four possible outcomes and zero elsewhere; and in Example 2 we have \(p(y) = 0\) everywhere. A point having \(p(y) > 0\) is called a point with discrete probability.

We now define a discrete distribution to be one that has all its probability at points with discrete probabilities:

DEFINITION 3

A distribution is a discrete distribution if its measure and distribution function have the form
(4) \[ P(A) = \sum_{y \in A} p(y), \quad F(y) = \sum_{t_1 \leq y_1, \ldots, t_k \leq y_k} p(t_1, \ldots, t_k) \]

The necessary and sufficient conditions for a real-valued function \( p \) on \( R^k \) to be the probability function for a discrete distribution are

1. Nonnegative: \( p(y) \geq 0 \) for all \( y \) in \( R^k \).
2. Discrete: \( p(y) = 0 \) except at a finite or countable number of points.
3. Normed: \( \sum p(y) = 1 \).

**EXAMPLE 3**

Tossing two symmetric dice (Example 1.3.7): Two symmetric dice are tossed and \( (y_1, y_2) \) recorded, where \( y_i \) is the number of points on the \( i \)th die. Then

\[
p(y_1, y_2) = \frac{1}{36}, \quad y_1, y_2 = 1, 2, \ldots, 6,
\]

\[ = 0, \quad \text{elsewhere}.\]

See Figure 2.26 and envisage \( \{0, 1\}^2 \) replaced by \( \{1, 2, \ldots, 6\}^2 \) and the value 1/4 replaced by 1/36.

---

**D ABSOLUTELY CONTINUOUS DISTRIBUTIONS**

Consider a distribution over \( R^k \) with measure \( P \) and distribution function \( F \). We can check for a probability density limit at the point \( y \) by examining an open cube \( C_\delta \) of side length \( \delta \) containing the point \( y \), calculating the density per unit volume

\[
\frac{P(C_\delta)}{\delta^k},
\]

and then seeing if this approaches a fixed value for any sequence of cubes shrinking to \( \emptyset \). We then define \( f(y) \) to be this limit if it exists and to be equal to zero otherwise. A distribution is called "absolutely continuous" if all its probability is at points at which the probability density limit exists.

**DEFINITION 4**

A distribution is an **absolutely continuous distribution** if its measure and distribution function have the form

\[
P(A) = \int_A f(y_1, \ldots, y_k) \, dy_1 \cdots dy_k = \int_A f(y) \, dy,
\]

\[
F(y) = \int_{(-\infty, y]} f(t) \, dt = \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_k} f(t_1, \ldots, t_k) \, dt_1 \cdots dt_k,
\]

where \( f \) is the probability density function defined above.
With an absolutely continuous distribution, probability for a set \( A \) is given by the volume over the set \( A \) beneath the function \( f \) and is obtained by integrating \( f \) over the set. Thus for the orientation of two symmetric coins as described in Example 2, we have

\[
P([0, 10] \times (0, 10]) = \int_0^{10} \int_0^{10} \frac{1}{180^2} \, dy_1 \, dy_2 = \frac{10^2}{180^2}.
\]

The following properties are necessary for the probability density of an absolutely continuous distribution:

1. Nonnegative: \( f(y) \geq 0 \) for all \( y \) in \( \mathbb{R}^k \).
2. Measurable: \( \{ y : f(y) \leq c \} \in \mathcal{B}^k \) for all \( c \).
3. Normed: \( \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(y) \, dy = 1 \).

Properties 1 and 3 are basic; property 2 will be taken for granted.

The properties are also sufficient for a function \( f \) to define an absolutely continuous distribution. This follows easily by noting that

\[
\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(t_1, \ldots, t_k) \, dt_1 \cdots dt_k
\]
satisfies the conditions for a distribution function as recorded in Section 2.3B. In peculiar cases a small technicality may arise—that such an \( f \) may differ from the probability density (as defined earlier) on a set having volume zero. For practical purposes we shall assume that \( f \) does give probability density.

If the distribution function \( F \) is a very nice regular function, then \( f \) can be calculated as

\[
f(y) = \frac{\partial}{\partial y_k} \cdots \frac{\partial}{\partial y_1} F(y_1, \ldots, y_k)
\]

\[
= \frac{\partial}{\partial y_k} \cdots \frac{\partial}{\partial y_1} \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_k} f(t) \, dt_1 \cdots dt_k
\]

by differentiating successively with respect to the coordinates. Note that if some function \( F \) has a derivative \( f \) as just defined, then the generalized monotone property 1 for a distribution function in Section B can be replaced by \( f(y) \geq 0 \).

**EXAMPLE 2 (continuation)**

For points in the rectangle \((0, 180)^2\) we have

\[
F(y_1, y_2) = \frac{y_1 y_2}{180^2},
\]

\[
f(y_1, y_2) = \frac{\partial}{\partial y_2} \frac{\partial}{\partial y_1} \frac{y_1 y_2}{180^2} = \frac{\partial}{\partial y_2} \frac{y_2}{180^2} = \frac{1}{180^2}.
\]

and for other points we have \( f(y_1, y_2) = 0 \). Now consider the probability for the event \( y_1^2 + y_2^2 \leq c^2 \) with \( c < 180 \):

\[
P(y_1^2 + y_2^2 \leq c^2) = \int_{y_1^2 + y_2^2 \leq c^2} f(t) \, dt_1 \, dt_2 = \int_{A_c} \frac{1}{180^2} \, dy_1 \, dy_2,
\]
where \( A_c = \{(y_1, y_2): 0 < y_1 < 180, 0 < y_2 < 180, y_1^2 + y_2^2 < c^2\} \) is the intersection of a disk and a square:

\[
P(y_1^2 + y_2^2 \leq c^2) = \frac{\pi c^2/4}{180^2},
\]

which for this uniform distribution is the area of the possible set for the event divided by the area of the possible set for the distribution; compare with Exercise 1.3.10.

**E MULTIVARIATE NORMAL DISTRIBUTION**

Many response distributions on \( R^2 \) and \( R^3 \) are concentrated around a central value and drop off sharply in each direction in the bell-shaped pattern mentioned in Section 2.2C; consider the example for \( (y_1, y_2) \) mentioned at the beginning of this section. Sometimes different coordinates tend to move together in such a way that the probability is piled up in a ridge in some direction. The multivariate normal distribution can sometimes provide an approximation for such a pattern of variation. In the simplest rotationally symmetric case, we have the standard normal distribution on \( R^k \), given by

\[
g(z_1, \ldots, z_k) = (2\pi)^{-k/2}e^{-z_1^2/2 - \cdots - z_k^2/2};
\]

see Figure 2.28.

We verify property 3 for a density function. The integral

\[
\int_{R^k} g(z) \, dz = \int_{R^k} (2\pi)^{-1/2}e^{-z_1^2/2} \cdots (2\pi)^{-1/2}e^{-z_k^2/2} \, dz_1 \cdots dz_k
\]

![Figure 2.28](image)

**Density function for the standard bivariate normal.**
has an integrand that factors so the variables separate and the range \( R \times \cdots \times R \) factors as a product set. We can then integrate factor by factor:

\[
\int_{\mathbb{R}^k} g(z) \, dz = \int_{-\infty}^{\infty} (2\pi)^{-1/2} e^{-z^2/2} \, dz \cdot \cdots \cdot \int_{-\infty}^{\infty} (2\pi)^{-1/2} e^{-z^2/2} \, dz;
\]

and we are thus led back to the unproved property 3 for the standard normal on \( R \). But consider the case with \( k = 2 \),

\[
\int_{\mathbb{R}^2} \frac{1}{2\pi} e^{-\left(x_1^2 + x_2^2\right)/2} \, dx_1 \, dx_2 = \int_0^\infty \int_0^{2\pi} \frac{1}{2\pi} e^{-r^2/2} \, dr \, da,
\]

where

\[

c_1 = r \cos \theta, \quad r^2 = z_1^2 + z_2^2,
\]

\[

c_2 = r \sin \theta, \quad \theta = \tan^{-1} \frac{z_2}{z_1},
\]

and where \( dr \) measures length radially at the point \((c_1, c_2)\) and \( r \, da \) measures length at right angles along the circle through \((c_1, c_2)\); see Figure 2.29. Thus

\[
\int_{\mathbb{R}^2} \frac{1}{2\pi} e^{-\left(x_1^2 + x_2^2\right)/2} \, dx_1 \, dx_2 = \int_0^\infty \int_0^{2\pi} \frac{1}{2\pi} e^{-r^2/2} \, dr \, da = 1.
\]

And it follows that property 3 holds for \( k = 2 \), then \( k = 1 \), then all \( k \).

The distribution function \( G \) for the standard normal on \( \mathbb{R}^k \) is given by

\[
G(z_1, \ldots, z_k) = \int_{-\infty}^{z_1} \cdots \int_{-\infty}^{z_k} \frac{1}{\sqrt{2\pi}} e^{-t_1^2/2} \cdots \frac{1}{\sqrt{2\pi}} e^{-t_k^2/2} \, dt_1 \cdots \, dt_k
\]

\[
= \int_{-\infty}^{z_1} \frac{1}{\sqrt{2\pi}} e^{-t_1^2/2} \, dt_1 \cdots \int_{-\infty}^{z_k} \frac{1}{\sqrt{2\pi}} e^{-t_k^2/2} \, dt_k
\]

\[
= G(z_1) \cdots G(z_k)
\]

in terms of the distribution function for the standard normal on \( R \).

**FIGURE 2.29**

Area change under the transformation \( z_1 = r \cos \theta, \, z_2 = r \sin \theta \).
F EXERCISES

1. Let $F(y_1, y_2) = F_1(y_1)F_2(y_2)$, where $F_1$ and $F_2$ are distribution functions on the real line.
   (a) Calculate $\{a_1, b_1, [a_2, b_2]\}$ if in terms of the component functions.
   (b) Show that $F$ is a distribution function.

2. A distribution on $R^2$ has probability function $p(y_1, y_2)$ with nonzero values as follows:

<table>
<thead>
<tr>
<th>$y_1$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0.05</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>2</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.05</td>
</tr>
<tr>
<td>3</td>
<td>0.05</td>
<td>0.10</td>
<td>0.05</td>
<td>0</td>
</tr>
</tbody>
</table>

Calculate the values of the distribution function $F(y_1, y_2)$ at the points given in the table.

3. A function $G$ on $R^2$ has the following form:

$$G(z_1, z_2) = 1 - (1 + z_1)^{-1} - (1 + z_2)^{-1} + (1 + z_1 + z_2)^{-1}, \quad z_1, z_2 > 0$$

and zero otherwise.
   (a) Calculate the function $g$ of formula (6).
   (b) Show that $G$ is a distribution function; note the remark following formula (6).

4. Consider a uniform distribution for $(y_1, y_2)$ over the unit disk centered at the origin.
   (a) Calculate the probability $P(y_1^2 + y_2^2 < c^2)$.
   (b) Plot the preceding probability as a function of $c$ on the range $(0, \infty)$; and as a function of $c^2$ on the range $(0, \infty)$.

5. Consider the following function $f$ on $R^2$: $f(y_1, y_2) = 4y_1y_2$ on $(0, 1)^2$ and zero elsewhere.
   (a) Show that $f$ is a probability density function.
   (b) Calculate the probabilities: $P(y_1 \leq c_1); P(y_2 \leq c_2); P(y_1 \leq c_1, y_2 \leq c_2)$. 

6. A transistor and a condenser have lifetimes $(y_1, y_2)$ with density function

$$f(y_1, y_2) = \frac{1}{6} \exp \left( -\frac{y_1}{6} - y_2 \right), \quad y_1, y_2 > 0,$$

and zero otherwise. Verify that $f$ is a probability density function.

7. (continuation) Calculate $P(y_1 > t, y_2 > t)$ for a positive number $t$. If the transistor and condenser are the only items subject to failure in a simple electronic system, calculate the probability that the system survives until time $t$.

8. Consider a uniform distribution for $(y_1, y_2)$ on the square $(-1, 1)^2$. Calculate the probabilities (a) $P(y_1^2 \leq c^2)$, and (b) $P(y_1^2 + y_2^2 \leq c^2)$, both with $0 < c < 1$.

9. Consider a uniform distribution for $(y_1, y_2, y_3)$ on the cube $(-1, 1)^3$. Calculate the following probabilities for $0 \leq c \leq 1$: (a) $P(y_1^2 \leq c^2)$; (b) $P(y_1^2 + y_2^2 \leq c^2)$; (c) $P(y_1^2 + y_2^2 + y_3^2 \leq c^2)$.

10. A coin of radius $r$ is tossed on a smooth floor marked with east–west lines and with north–south lines at unit spacings. Let $(y_1, y_2)$ with $(0 \leq y_1 < 1)$ give the coordinates of its center in the square in which it falls.
   (a) Record the probability density function based on symmetry.
   (b) Calculate the probability that it will intersect an E-W line.
   (c) Calculate the probability that it will intersect both an E-W line and a N-S line.
   (d) Calculate the probability that it will not intersect any line.

11. A needle of length $l$ is tossed onto the smooth floor described in Exercise 10. Let $(y_1, y_2)$ give
the coordinates of its center (Exercise 10) and \( \gamma_3 \) be its angle measured positively from E–W 
\( 0 < \gamma_3 < \pi \).
(a) Record the probability density for \( (\gamma_1, \gamma_2, \gamma_3) \) based on symmetry.
(b) Calculate the probability that it does not intersect an E–W line; for simplicity assume that 
\( f < 1 \).

**G PROBLEMS**

12 The **Pareto(a)** distribution on \( \mathbb{R}^2 \) has density function
\[
g(x_1, x_2) = \alpha(a + 1)(1 + x_1 + x_2)^{-\alpha+2}, \quad x_1, x_2 > 0,
\]
\[
= 0, \quad \text{otherwise}.
\]
(a) Show that \( g \) is a probability density function; \( \alpha > 0 \).
(b) Calculate the distribution function \( G \); compare with Exercise 3.

13 The **canonical Student(n)** distribution on \( \mathbb{R}^2 \) has density function
\[
g(x_1, x_2) = \frac{\Gamma\left(\frac{n + 2}{2}\right)}{\pi \Gamma\left(\frac{n}{2}\right)} \left(1 + x_1^2 + x_2^2\right)^{-\frac{n+2}{2}}.
\]

Show that \( g \) is a probability density function; \( n > 0 \). Use the integration results from
Exercise 2.2.19. 

14 The **Pareto(a)** distribution on \( \mathbb{R}^k \) has density function
\[
g(x_1, \ldots, x_k) = \alpha(a + 1) \cdots (\alpha + k - 1)(1 + x_1 + \cdots + x_k)^{-\alpha+k}
\]
for \( x_1, \ldots, x_k > 0 \), and zero otherwise. Show that \( g \) is a probability density function;
\( \alpha > 0 \).

15 The **canonical Student(n)** distribution on \( \mathbb{R}^k \) has density function
\[
g(x_1, \ldots, x_k) = \frac{\Gamma\left(\frac{n + k}{2}\right)}{\pi^{k/2} \Gamma\left(\frac{n}{2}\right)} \left(1 + x_1^2 + \cdots + x_k^2\right)^{-\frac{n+k}{2}}.
\]

Show that \( g \) is a probability density function; \( n > 0 \). Use the integration results of Exercise
2.2.19 and Problem 13.

16 The **Dirichlet(r_1, \ldots, r_{k+1})** distribution on \( \mathbb{R}^k \) has density
\[
f(y_1, \ldots, y_k) = \frac{1}{\Gamma(r_1) \cdots \Gamma(r_{k+1})} y_1^{r_1-1} \cdots y_k^{r_k-1} \left(1 - \sum_{1}^{k} y_j\right)^{r_{k+1}-1}
\]
on the set \( \{ y_1 > 0, \sum_{1}^{k} y_j < 1 \} \) and zero elsewhere. Show that \( f \) is a probability density
function; \( r_j > 0 \). This is a generalized beta distribution; use Exercise 2.2.17.

17 The norming constant for the normal density can be verified without recourse to the bivariate
and multivariate case. With
\[
F(x) = \left(\int_{0}^{x} e^{-t^2/2} dt\right)^2, \quad G(x) = \int_{0}^{x} \frac{2e^{-s^2(1+s^2)/2}}{1 + s^2} ds
\]
(a) Show that \( F'(x) + G'(x) = 0 \).
(b) Then show that \( F(x) + G(x) = \pi/2 \).
(c) Then consider \( \lim_{x \to \infty} (F(x) + G(x)) \) and thus obtain the norming constant.
2.4

CHANGE OF VARIABLES

There is often a considerable degree of arbitrariness in how we present the response variable of a random system. For example, in measuring the weight increase of an animal under diet therapy, we could use pounds or grams or some other units. In measuring the noise level at which hearing just occurs in a psychology test, we could use a convenient energy unit or we could take 10 times the logarithm of energy and use decibels. And in a catalyzed chemical reaction, we could measure the amount of converted material or we could use the ratio of converted material to unconverted material.

Often there is a preferred or natural way of presenting the response. This can relate to the simplicity with which response effects are determined by input variables. And it can relate to having a simple scaling for the variation, independent of the input variables.

In this section we examine the mathematics of changing the way we present a response, that is, of making a change of variable or of relabeling the response. The question of a natural labeling for the response will be mentioned briefly in Chapter 12.

A CHANGE OF VARIABLE FOR A REAL-VALUED RESPONSE

In Section 2.2C we introduced the standard normal density $g(z)$ for $z$ and then obtained the general normal$(\mu, \sigma)$ density for $y = \mu + oz$ by a simple argument in terms of events; see Figure 2.30. Now consider this change of variable more generally.

Let $G$ be the distribution function for an initial response variable $x$. And suppose that $y = Y(x)$ is a one-to-one monotone-increasing function from $R$ onto $R$. We investigate the distribution for this new or relabeled response $y = Y(x)$; let $F_y$ designate the distribution function.

For notation let $x = X(y)$ be the inverse function. We can then note that the event $(-\infty, y]$ for the new response is the same as the event $(-\infty, X(y)]$ for the original response; see Figure 2.31. Thus we obtain

$$F_y(y) = G(X(y)),$$

which expresses $F_y$ in terms of the original $G$. In this formula note that we merely substitute, expressing the original variable in terms of the new variable.

Formula (1) holds more generally. Suppose that $y = Y(x)$ is a one-to-one

![Figure 2.30](image)

Transformation $y = \mu + oz$ and inverse $z = (y - \mu)/\sigma$. 
monotone-increasing function from an interval for the initial response to an interval for the new response. And suppose that all the probability for \( G \) is on the initial interval. Then formula (1) gives the new distribution function \( F_y \) throughout the new interval.

**EXAMPLE 1**

**Logistic and exponential growth:** Consider the standard logistic distribution (Exercise 2.2.11) for an initial response \( z \); recall Exercise 2.2.11 and Figure 2.17. The distribution function is easily derived and has the simple form

\[
G(z) = \frac{1}{1 + e^{-z}} = \frac{e^z}{1 + e^z}, \quad -\infty < z < \infty.
\]

Now suppose that we take an exponential transformation \( y = e^z \) of the initial response; this maps \((-\infty, \infty)\) into \((0, \infty)\). The inverse transformation is \( z = \ln y \). Thus the distribution function for the new response is

\[
F_y(y) = \frac{e^{\ln y}}{1 + e^{\ln y}} = \frac{y}{1 + y}, \quad 0 < y < \infty.
\]

This is easily seen from Problem 2.2.18 to be canonical \( F(2, 2) \). The logistic is often used to represent growth when there is an inhibiting upper limit; see Figure 2.32.

**B IN TERMS OF THE PROBABILITY FUNCTION**

For a discrete distribution the effects of relabeling are very simple if we use the probability function. Consider a discrete distribution for \( x \) with probability function \( p \). And suppose that \( y = Y(x) \) is a one-to-one transformation of \( R \) onto \( R \) (it suffices to have it one to one from possible values of \( x \) to possible values of \( y \)). Let \( p_y \) be the probability function for the new \( y = Y(x) \).

The probability function for the new response at the value \( y \) is, of course, the same as for the original response at \( x = X(y) \); thus

\[
(2) \quad p_y(y) = p(X(y)).
\]
We merely substitute expressing the original variable in terms of the new variable.

**EXAMPLE 2**

**Heads with two symmetric coins:** The distribution for the number, say $x$, of heads is available from Example 1.3.5:

$p(0) = 1/4, \quad p(1) = 1/2, \quad p(2) = 1/4.$

The probability function for the proportion $\hat{p} = Y(x) = x/2$ of heads is available immediately:

$p_{Y}(0) = 1/4, \quad p_{Y}(1/2) = 1/2, \quad p(1) = 1/4;$

see Figure 2.33.

**FIGURE 2.33**

Probability function plotted against number of heads and against proportion of heads.
C IN TERMS OF THE PROBABILITY DENSITY FUNCTION

For an absolutely continuous distribution the effects of relabeling are quite simple if we use the probability density function $g$. Consider an absolutely continuous distribution for $x$ with probability density function $g$. And suppose that $y = Y(x)$ is a one-to-one transformation continuously differentiable each way from $R$ onto $R$. Let $f_y$ be the density function for the new $y = Y(x)$ and $F_y$ be the corresponding distribution function.

Then in the monotone-increasing case we have

$$F_y(y) = G(X(y)) = \int_{-\infty}^{X(y)} g(x) \, dx;$$

and differentiation with respect to $y$ (using the chain rule) gives

$$(3) \quad f_y(y) = g(X(y)) \cdot \left| \frac{dx}{dy} \right|;$$

see Figure 2.34. It is easily seen that with the absolute-value sign the formula also holds in the monotone-decreasing case.

Formula (3), in fact, holds more generally. Suppose that $y = Y(x)$ has the continuity properties and maps an interval for the initial response into an interval for the new response. And suppose that all the probability for $g$ is on the initial interval, then formula (3) gives the density function $f_y$ through the new interval.

Intuitively, it is more convenient to think of a chunk of probability $g(x) \, dx$

![Diagram of change of variables](image)

**FIGURE 2.34**

Density $g$ for $x$ transformed to the density $f_y$ for $Y$. 
associated with the interval \((x, x + dx)\) for the initial variable. This same chunk of probability of course refers to the image \((y, y + dy)\) obtained from the mapping \(y = Y(x)\). If the length of the increment \(dy\) is different from the length of \(dx\), then the density must change inversely so that the area representing probability remains the same. This can be presented conveniently as follows:

\[
(4) \quad g(x) \, dx = g(X(y)) \, dX(y) = g(X(y)) \cdot \left| \frac{dX(y)}{dy} \right| \, dy = f_y(y) \, dy;
\]

see Figure 2.34. More formally, this is an abbreviated way of presenting the essentials of the change-of-variable formula for an integral on the real line. Note the simplicity: The density function is expressed in terms of the new variable and the differential is expressed in terms of the corresponding differential for the new variable.

EXAMPLE 3

Consider a triangular distribution for the variable \(x\):

\[
g(x) \, dx = 2x \, dx \quad \text{on} \quad (0, 1)
\]

and zero elsewhere. And suppose that we are interested in \(y = Y(x) = x^3\); this maps the region \((0, 1)\), where the probability is, into the region \((0, 1)\); see Figure 2.35. The effect of the transformation on the differential is easily calculated:

\[
\begin{align*}
\text{FIGURE 2.35} & \\
\text{Triangular distribution for } x, \text{ the transformation } y = x^3, \text{ and the distribution } f_y \text{ for } y.
\end{align*}
\]
Sec. 2.4: Change of Variables  

\[ x = X(y) = y^{1/3}, \quad \frac{dx}{dy} = (1/3)y^{-2/3} \quad \text{on } (0, 1) \]

The density function for \( y \) is then obtained from (3):

\[ g(X(y)) \cdot \left| \frac{dx}{dy} \right| = 2y^{1/3} \cdot (1/3) \cdot (2/3) = (2/3)y^{-1/3} \quad \text{on } (0, 1) \]

\[ = \frac{\Gamma(5/3)}{\Gamma(2/3)\Gamma(1)} \cdot y^{(12/3)-1}(1 - y)^{1-1} \quad \text{on } (0, 1), \]

which is the beta(2/3, 1) distribution in Problem 2.2.17.

EXAMPLE 4

Consider a gamma(1) distribution for the variable \( x \):

\[ g(x) \, dx = e^{-x} \, dx \quad \text{on } (0, \infty) \]

and zero elsewhere. And suppose that we are interested in \( u = U(x) = e^{-x} \); this maps the region \((0, \infty)\) into \((0, 1)\); see Figure 2.36.

The effect of the transformation on the differential is easily calculated:

\[ x = X(u) = -\ln u, \quad \frac{dx}{du} = -\frac{1}{u}. \]

The density function for \( u \) is then obtained from (3):

\[ g(X(u)) \cdot \left| \frac{dx}{du} \right| = u \cdot \frac{1}{u} = 1 \]

on the effective range \((0, 1)\); thus \( u \) has the uniform(0, 1) distribution. The details can be expressed more compactly in terms of differentials:

\[ \frac{du}{x} = \frac{1}{u} \quad \text{and} \quad dx = -\frac{1}{u} \, du \]

FIGURE 2.36

Gamma(1) distribution for \( x \), the transformation \( u = e^{-x} \), and the uniform(0, 1) distribution for \( u \).
\[ e^{-x} \, dx = u \left| \frac{d}{du} (-\ln u) \right| \, du = u \frac{1}{u} \, du = du \]
on the effective ranges \((0, \infty)\) and \((0, 1)\).

To illustrate the use of differentials to approximate probabilities, consider the interval \((1, 1.01)\) for \(x\) and the corresponding interval \((0.36422, 0.36788)\) for \(u\). Then
\[
\begin{align*}
P((1.00, 1.01)) &= e^{-1} \cdot 0.01 = 0.00368, \\
P_u((0.36422, 0.36788)) &= 0.00366.
\end{align*}
\]

**D PROBABILITY INTEGRAL TRANSFORMATION**

Consider some continuous distribution function \(G\). And for simplicity in the discussion to follow suppose that \(G\) is one to one from an open interval for the initial variable, say \(x\), to the interval \((0, 1)\) for values, say \(u\), of the function \(G\); this avoids flat spots on the function \(G\). We are going to consider a special transformation from \(x\) to \(u\), the transformation \(u = U(x) = G(x)\), and then the inverse transformation from \(u\) to \(x\), say \(x = X(u) = G^{-1}(u)\).

First consider an initial response \(x\) with distribution function \(G\). And suppose that we define a new response \(u = U(x) = G(x)\) by means of the given distribution function \(G\); this is called the **probability integral transformation** (integral of density in the absolutely continuous case). In an experimental investigation, a certain theory may specify a distribution \(G\). Then with an observed value \(x\) we could be interested in \(u = G(x)\), the probability of a value \(x\) or smaller. The distribution describing possible \(u\) values under the theory would be of particular interest for assessing the theory; we now derive this distribution for \(u\).

By formula (1) the distribution function for \(u\) is
\[
F_u(u) = G(X(u)) = G(G^{-1}(u)) = u, \quad 0 < u < 1.
\]
Thus \(u = G(x)\) has the uniform\((0, 1)\) distribution; see Figure 2.37. For convenience, let \(H\) be the distribution function for this uniform distribution: that is, \(H(u) = u\) on \((0, 1)\).

![FIGURE 2.37]

Probability integral transformation \(u = G(x)\) gives the uniform\((0, 1)\) distribution, and inverse probability integral transformation \(X = G^{-1}(u)\) carries the uniform\((0, 1)\) into the distribution with distribution function \(G\).
Second, consider an initial response \( u \) that has a uniform(0, 1) distribution. And suppose that we define a new response \( x = X(u) = G^{-1}(u) \), where \( G \) is a distribution function as described in the first paragraph. By formula (1) the distribution function for \( x \) is

\[
F_x(x) = H(U(x)) = U(x) = G(x).
\]

Thus \( x = G^{-1}(u) \) has a distribution with distribution function \( G \); see Figure 2.37. This is called the inverse probability integral transformation. It provides a means of carrying an initial variable with a uniform distribution into a new variable with a special distribution of interest. Values from a uniform distribution are easily produced on a computer; by an appropriate transformation, then, we can obtain values from any particular distribution of interest.

**EXAMPLE 5**

In Example 2.2.1 we noted how sample values can be obtained from the uniform(0, 1) distribution. Suppose that we are using three-figure accuracy and obtain the sequence

7 9 3 5 9 2 2 3 1

from a computer programmed to produce random digits or from a table of randomly produced digits. Then (the fourth-place 5 is added to center the values)

0.7935, 0.5925, 0.2315, . . .

are sample values (approximately) from the uniform(0, 1). For the logistic in Example 1, we can calculate the inverse function \( G^{-1}(u) = \ln \frac{u}{1 - u} \) and from formula (6) obtain the sample values

\[
\begin{align*}
\ln \frac{0.7935}{1 - 0.7935} & = \ln 3.843, \\
\ln \frac{0.5925}{1 - 0.5925} & = \ln 1.454, \\
\ln \frac{0.2315}{1 - 0.2315} & = \ln 0.3012, \\
= & 1.346, \\
= & 0.374, \\
= & -1.200,
\end{align*}
\]

these are sample values from the logistic distribution. For an initial sample value near 0 or near 1 we should add more digits to maintain accuracy under the transformation.

---

**CHANGE OF VARIABLE ON THE PLANE**

Consider a distribution for an initial response \((x_1, x_2)\). And suppose that \((y_1, y_2) = (Y_1(x_1, x_2), Y_2(x_1, x_2))\) is a one-to-one mapping of \(R^2\) onto \(R^2\). We investigate the distribution for the new response \((y_1, y_2)\). For notation let \((x_1, x_2) = (X_1(y_1, y_2), X_2(y_1, y_2))\) be the inverse function. As an example, \(y_1 \) could be the average yield \((x_1 + x_2)/2\) on two plots of land and \(y_2 \) could be the difference, \(x_2 - x_1\), in yields.

For the case of a discrete distribution, the values of the probability function are reassigned to the corresponding image points under the mapping; see Section B.

For the absolutely continuous case, let \(g\) be the density function for \((x_1, x_2)\). And suppose that the mapping is continuously differentiable each way. Let \(f_{Y_1,Y_2}\) be the density function for the new response \((y_1, y_2)\).
See Figure 2.38. The inverse mapping gives \((x_1, x_2)\) from \((y_1, y_2)\), and gives \((x_1 + dx_1, x_2 + dx_2)\) from \((y_1 + dy_1, y_2 + dy_2)\). By partial differentiation we can obtain the relationship between the increments \(dx_1, dx_2\) and the increments \(dy_1, dy_2\):

\[
\begin{align*}
dx_1 &= \frac{\partial X_1}{\partial y_1} \, dy_1 + \frac{\partial X_1}{\partial y_2} \, dy_2, \\
dx_2 &= \frac{\partial X_2}{\partial y_1} \, dy_1 + \frac{\partial X_2}{\partial y_2} \, dy_2.
\end{align*}
\]

This can be rewritten in vector and matrix notation as

\[
\begin{pmatrix} dx_1 \\ dx_2 \end{pmatrix} = \begin{pmatrix} \frac{\partial X_1}{\partial y_1} & \frac{\partial X_1}{\partial y_2} \\ \frac{\partial X_2}{\partial y_1} & \frac{\partial X_2}{\partial y_2} \end{pmatrix} \begin{pmatrix} dy_1 \\ dy_2 \end{pmatrix}.
\]

The matrix is called the Jacobian matrix. The preceding equation shows how the inverse mapping produces the increment vector \((dx_1, dx_2)\) from the increment vector \((dy_1, dy_2)\); see Figure 2.38.

In particular, we record in Table 2.3 the increment vectors \((dx_1, dx_2)\) that are obtained from the increment vectors \((dy_1, dy_2)\) = (1, 0) and = (0, 1) (these correspond to the two axes on the \(y\) space). From this we see that a parallelogram in the initial space is obtained from a square in the new space; see Figure 2.38. The area of the parallelogram is obtained as the positive value of the determinant of the two vectors

\[
\begin{vmatrix}
\frac{\partial X_1}{\partial y_1} & \frac{\partial X_1}{\partial y_2} \\
\frac{\partial X_2}{\partial y_1} & \frac{\partial X_2}{\partial y_2}
\end{vmatrix} = \begin{vmatrix}
\frac{\partial Y_1}{\partial x_1} & \frac{\partial Y_1}{\partial x_2} \\
\frac{\partial Y_2}{\partial x_1} & \frac{\partial Y_2}{\partial x_2}
\end{vmatrix}^{-1}
\]

which is, of course, the positive value of the Jacobian determinant; the second expression is available from advanced calculus. The area of the square is unity. Thus we obtain

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{Figure2.38}
\caption{Inverse mapping \((X_1, X_2)\) gives \((x_1, x_2)\) from \((y_1, y_2)\) and gives the parallelogram with vertex \((x_1, x_2)\) from the square with vertex \((y_1, y_2)\).}
\end{figure}
TABLE 2.3
Increment vectors

<table>
<thead>
<tr>
<th>Initial space increment</th>
<th>inverse mapping</th>
<th>New space increment</th>
</tr>
</thead>
</table>
| \[
\begin{pmatrix}
\frac{\partial X_1}{\partial y_1} \\
\frac{\partial X_2}{\partial y_1} \\
\frac{\partial X_1}{\partial y_2} \\
\frac{\partial X_2}{\partial y_2}
\end{pmatrix}
\] | \[
\begin{pmatrix}
1 \\
0 \\
0 \\
1
\end{pmatrix}
\] | \[
\begin{pmatrix}
\frac{\partial X_1}{\partial y_1}, X_2 \\
\frac{\partial X_2}{\partial y_1}, y_2
\end{pmatrix}
\] |

the ratio of areas as expressed by

(7) \[ dx = \begin{vmatrix} \frac{\partial X_1}{\partial y_1} & \frac{\partial X_1}{\partial y_2} \\ \frac{\partial X_2}{\partial y_1} & \frac{\partial X_2}{\partial y_2} \end{vmatrix}, \quad dy = \begin{vmatrix} \frac{\partial X_1}{\partial y_1}, X_2 \\ \frac{\partial X_2}{\partial y_1}, y_2 \end{vmatrix} \ dy. \]

This intuitive description is presented to motivate the standard change-of-variable formula for a two-dimensional integral.

Let \( B \) be a set for the new response and \( A \) be the corresponding event for the original response. Then

\[
P(\{y_1, y_2\} \in B) = \int_A g(x_1, x_2) \, dx_1 \, dx_2
\]

\[
= \int_B g(X_1(y_1, y_2), X_2(y_1, y_2)) \begin{vmatrix} \frac{\partial X_1}{\partial y_1}, X_2 \\ \frac{\partial X_2}{\partial y_1}, y_2 \end{vmatrix} \, dy_1 \, dy_2.
\]

Thus the density function for \((y_1, y_2)\) is

(8) \[
f_{y_1, y_2}(y_1, y_2) = g(X_1(y_1, y_2), X_2(y_1, y_2)) \begin{vmatrix} \frac{\partial X_1}{\partial y_1}, X_2 \\ \frac{\partial X_2}{\partial y_1}, y_2 \end{vmatrix}.
\]

Formula (8), in fact, holds more generally. Suppose that the mapping is one to one from a region for the initial response to a region for the new response. And suppose that all the probability for \(g\) is on the initial region. Then formula (8) gives the new density throughout the new region; it is zero otherwise.

EXAMPLE 6

Consider a uniform distribution for \((x_1, x_2)\) over the unit square \((0, 1)^2\) and suppose that we are interested in a new variable \((y_1, y_2)\) obtained by the affine transformation

\[
y_1 = 1 + x_1 - x_2, \quad \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} + \begin{pmatrix} 1 & -1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.
\]
FIGURE 2.39
Uniform distribution on the unit square \((0, 1)^2\) is mapped into a uniform distribution for \((y_1, y_2)\) on the parallelogram.

Note that this maps the square \((0, 1)^2\) on the initial space into the parallelogram with vertices \((1, 2), (2, 4), (0, 3), (1, 5)\); see Figure 2.39. The probability density function \(g\) for the initial variable \((x_1, x_2)\) is

\[ g(x_1, x_2) = 1 \quad \text{on} \quad (0, 1)^2 \]

and zero otherwise. For this example we do not need to calculate the inverse transformation. We can calculate the Jacobian for the given transformation

\[
\begin{pmatrix}
\frac{\partial Y_1}{\partial x_1} & \frac{\partial Y_1}{\partial x_2} \\
\frac{\partial Y_2}{\partial x_1} & \frac{\partial Y_2}{\partial x_2}
\end{pmatrix} = \begin{pmatrix}
1 & -1 \\
2 & 1
\end{pmatrix},
\]

which has determinant 3. Thus from formula (8) we obtain

\[ f_{Y_1, Y_2}(y_1, y_2) = 1 \cdot 3^{-1} = 1/3 \]

on the parallelogram formed by \((1, 2), (2, 4), (0, 3), (1, 5)\) and zero otherwise. This is a uniform distribution on that parallelogram.

EXAMPLE 7

Consider a distribution for \((x_1, x_2)\) with density function \(g\). And suppose that we are interested in a new variable \((r, \theta)\) given by

\[ r = (x_1^2 + x_2^2)^{1/2} \quad x_1 = r \cos \theta, \]
\[ \theta = \tan^{-1} \left( \frac{x_2}{x_1} \right) \quad x_2 = r \sin \theta; \]

the inverse mapping is recorded on the right (see Figure 2.29). Note that the transformation carries \(R^2\) less the origin into \((0, \infty) \times [0, 2\pi]\). The Jacobian determinant of the inverse mapping is easily calculated:
\[
\begin{bmatrix}
\frac{\partial x_1}{\partial r} & \frac{\partial x_1}{\partial \theta} \\
\frac{\partial x_2}{\partial r} & \frac{\partial x_2}{\partial \theta}
\end{bmatrix} = \begin{bmatrix}
\cos \theta & -r \sin \theta \\
\sin \theta & r \cos \theta
\end{bmatrix}
\]
\[
= r(\cos^2 \theta + \sin^2 \theta) = r.
\]

Thus from formula (8) we obtain
\[f_{R\theta}(r, \theta) = g(r \cos \theta, r \sin \theta) r\]
on \((0, \infty) \times [0, 2\pi)\) and zero otherwise. In the special case that \(g\) is rotationally symmetric, \(g(x_1, x_2) = h(\sqrt{x_1^2 + x_2^2})\), then
\[f_{R\theta}(r, \theta) = h(r) r\]
on \((0, \infty) \times [0, 2\pi)\) and zero otherwise. And in the particular case that \(g\) is the standard normal density (Section 2.3E),
\[f_{R\theta}(r, \theta) = \frac{1}{2\pi} e^{-r^2/2r}\]
on \((0, \infty) \times [0, 2\pi)\) and zero otherwise. The simplicity of these last two cases will be examined under independence in Section 3.3.

### F CHANGE OF VARIABLE FOR VECTOR RESPONSES

Now consider the general case of a distribution on \(R^k\) with density function \(g\). And suppose that we have a function \(Y' = (Y_1, \ldots, Y_k)\) which maps \(R^k\) onto \(R^k\) and is one to one and continuously differentiable each way. We investigate the distribution for the new response \(y = Y(x)\). For notation let \(x = X(y)\) designate the inverse function.

Then in the pattern in the preceding section, we have
\[(9) \quad f_y(y) = g(X(y)) \left| \frac{\partial x}{\partial y} \right|_+\]
where
\[
\left| \frac{\partial x}{\partial y} \right| = \begin{vmatrix}
\frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_k} \\
\vdots & \ddots & \vdots \\
\frac{\partial x_k}{\partial y_1} & \cdots & \frac{\partial x_k}{\partial y_k}
\end{vmatrix}
\]
is the Jacobian determinant. This can be presented more conveniently in terms of differentials:
\[(10) \quad g(x) \, dx = g(X(y)) \cdot \left| \frac{\partial x}{\partial y} \right|_+ \, dy = f_y(y) \, dy.
\]

Formula (9) holds more generally. Suppose that the mapping is one to one from a region for the initial response to a region for the new response. And suppose that all the probability for \(g\) is in the initial region. Then formula (9) gives the new density function throughout the new region. (For details, see a text on advanced calculus.)
EXAMPLE 8

Consider a distribution for \( x' = (x_1, \ldots, x_k) \) with density function

\[
g(x) = \exp\left\{ -\sum x_i \right\} \quad \text{if each } x_i > 0
\]

and zero otherwise. And suppose that we are interested in \( y' = (y_1, \ldots, y_k) \), given by

\[
\begin{align*}
y_1 &= \frac{x_1}{\sum x_i} & x_1 &= y_1 y_k \\
\vdots & & \vdots \\
y_{k-1} &= \frac{x_{k-1}}{\sum x_i} & x_{k-1} &= y_{k-1} y_k \\
y_k &= \sum x_i & x_k &= (1 - y_1 - \cdots - y_{k-1}) y_k;
\end{align*}
\]

the inverse transformation is recorded on the right. Note that the transformation carries the region \((0, \infty)^k\), where the probability for \( x \) is onto the set

\[
S_k = \{ (y_1, \ldots, y_{k-1}) : 0 < y_i, \sum y_i^k < 1 \} \times \{ y_k : 0 < y_k < \infty \}.
\]

The Jacobian determinant is easily calculated:

\[
\begin{vmatrix}
\frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_k} \\
\vdots & \ddots & \vdots \\
\frac{\partial x_k}{\partial y_1} & \cdots & \frac{\partial x_k}{\partial y_k}
\end{vmatrix}
= \begin{vmatrix}
y_k & 0 & y_1 \\
0 & \ddots & \vdots \\
y_{k-1} & y_k & \cdots & y_k & (1 - \sum y_i^{k-1}) y_k
\end{vmatrix}
= y_k^{k-1}.
\]

Thus from formula (9) we obtain

\[
f_{y'}(y) = \exp\left\{ -y_k \right\} y_k^{k-1}
\]

on the set \( S_k \) in \( R^k \) and zero elsewhere. The simplicity of this last formula will become transparent in Section 3.3.

\[\text{G MULTIVARIATE NORMAL}\]

Consider the standard normal distribution for \( z = (z_1, \ldots, z_k)' \) on \( R^k \):

\[
g(z_1, \ldots, z_k) = (2\pi)^{-k/2} e^{-z_1^2 + \cdots + z_k^2/2},
\]

as recorded in Section 2.3E. Suppose that we picture \( z \) as describing in standard form the variation for a \( k \)-dimensional response. And suppose the response itself is obtained by the relabeling,

\[
\begin{align*}
y_1 &= \mu_1 + \gamma_{11} z_1 + \cdots + \gamma_{1k} z_k \\
\vdots \\
y_k &= \mu_k + \gamma_{k1} z_1 + \cdots + \gamma_{kk} z_k
\end{align*}
\]

or more briefly in vector and matrix notation,
\[ y = \mu + \Gamma z \]

where \( \mu = (\mu_1, \ldots, \mu_k)' \) is the column vector giving the relocation and

\[
\Gamma = \begin{pmatrix}
\gamma_{11} & \cdots & \gamma_{1k} \\
\vdots & \ddots & \vdots \\
\gamma_{k1} & \cdots & \gamma_{kk}
\end{pmatrix}
\]

is a matrix giving the linear scaling and combining of components; see Figure 2.40.

For a one-to-one transformation we assume that \( |\Gamma| \neq 0 \). The Jacobian matrix of \( y \) with respect to \( z \) is just \( \Gamma \). Thus

\[
dy = |\Gamma|_+ dz; \quad dz = |\Gamma|_+^{-1} dy.
\]

The probability at an element \( dz \) for the initial variable is given by

\[
(2\pi)^{-k/2} \exp \left\{ -(1/2)(z_1^2 + \cdots + z_k^2) \right\} dz = (2\pi)^{-k/2} \exp \left\{ -(1/2)z'z \right\} dz,
\]

where we use the matrix product of the row vector \( z' \) and the column vector \( z \) to give \( \Sigma z_i^2 \). By expressing this in terms of the new labels, we obtain

**FIGURE 2.40**

Standard normal \( g \). The general normal \( N(\mu; \Sigma) \) located at \( \mu \) and scaled by \( \Sigma \). The volume of the rectangular cylinders is the same; if the base area is smaller, the height is proportionately greater.
\[ f(y) \, dy = (2\pi)^{-k/2} \exp \left\{ -(1/2)(y - \mu) \Gamma^{-1/2}(y - \mu) \right\} |\Gamma|_{+1}^{-1} \, dy \]
\[ = (2\pi)^{-k/2} \exp \left\{ -(1/2)(y - \mu) \Sigma^{-1/2}(y - \mu) \right\} |\Sigma|_{-1/2} \, dy, \]

where the matrix
\[ \Sigma = \Gamma \Gamma' = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1k} \\ \vdots & \ddots & \vdots \\ \sigma_{k1} & \cdots & \sigma_{kk} \end{pmatrix} \]

has typical component \( \sigma_{ij} = \gamma_{1i} \gamma_{j1} + \cdots + \gamma_{1k} \gamma_{jk} \), which is the inner product of the \( i \)th row vector and the \( j \)th row vector of \( \Gamma \); for some matrix details, see Supplement Section 2.5. The probability differential for \( y \) can be expressed more fully as
\[ f(y) \, dy = (2\pi)^{-k/2} |\Sigma|^{-1/2} \exp \left\{ -(1/2) \sum_{i,j=1}^{k} \sigma^{ij}(y_i - \mu_i)(y_j - \mu_j) \right\} \, dy, \]

where the inverse matrix
\[ \Sigma^{-1} = \begin{pmatrix} \sigma^{11} & \cdots & \sigma^{1k} \\ \vdots & \ddots & \vdots \\ \sigma^{k1} & \cdots & \sigma^{kk} \end{pmatrix} \]

has components \( \sigma^{ij} \). This is the multivariate normal(\( \mu; \Sigma \)) distribution, normal with mean \( \mu \) and variance \( \Sigma \), as we shall see later. The multivariate normal is used in archeology, medicine, education—wherever there is a need to describe reasonably behaved multiple responses.

### H Exercises

1. The distribution function \( G \) is equal to \( y^2 \) on \((0, 1)\) (Exercise 2.2.1). Find the distribution function of \( w = y^2 \).
2. A distribution function \( H \) is equal to \( x^2/4 \) on \((0, 2)\). Find the distribution function for \( w = x^3 \).
3. Show that the distribution of \( y = e^x \) in Example 1 is canonical \( F(2, 2) \); see Problem 2.2.18.
4. A symmetrical die is tossed, yielding \( y \) points. Record the probability function for \( w = y^2 \).
5. The distribution of cells on a slide has \( \lambda \) per unit area. It can be shown that the distance \( d \) from a cell to its nearest neighbor has density
\[ g(d) = 2\pi \lambda d \exp \left\{ -\lambda \pi d^2 \right\}. \quad d > 0. \]

The area of the largest free disk centered on a cell has area \( A = \pi d^2 \). Find the density function for \( A \).
6. Let \( G \) be the distribution function for \( x \), and let \( y = Y(x) \) with inverse \( x = X(y) \) be one to one monotone decreasing from \( R \) to \( R \). Show that the distribution function of \( y \) is such that
\[ F_y(y - 0) = 1 - G(X(y)). \]

(continuation) Consider the gamma(1) distribution for \( x \). Find the distribution function for \( y = -x \).
8. (continuation) Show that \( u = e^{-x} \) is the probability integral transformation for the modified response \(-x\); compare with Example 4.
9. Let \( z \) be Pareto(\( \alpha \)) (Exercise 2.2.10). Show that \( y = (1 + z)^p - 1 \) with \( p > 0 \) is Pareto(\( \alpha/p \)).
10. Let \( z \) be Weibull(\( \beta \)) (Exercise 2.2.13). Show that \( y = z^p \) is exponential(1); and that \( y = z^p \) is Weibull(\( \beta/p \)) with \( p > 0 \).
11. Let \( z \) be extreme value (Exercise 2.2.14). Show that \( y = e^{-z} \) is exponential(1).
I EXERCISES: INVERSE PROBABILITY INTEGRAL TRANSFORMATION

12 Show that the inverse probability integral transformation (IPIT) for the exponential(\(\theta\)) distribution (Exercise 2.2.7) is
\[ y = -\theta \ln(1 - u). \]

13 (continuation) If \(u\) is uniform(0, 1), deduce that \(-\ln u\) is exponential(1) or gamma(1); and that \(-2\ln u\) is chi square(2); see Exercises 2.2.15 and 2.2.16.

14 Show that the IPIT for the Cauchy distribution (Exercise 2.2.8) is
\[ z = \tan\left(\pi u - \frac{\pi}{2}\right). \]

15 Show that the IPIT for the Pareto(\(\alpha\)) distribution (Exercise 2.2.10) is
\[ z = (1 - u)^{-1/\alpha} - 1. \]

16 Show that IPIT for the Weibull(\(\beta\)) distribution (Exercise 2.2.13) is
\[ z = [-\ln(1 - u)]^{1/\beta}. \]

17 Show that IPIT for the extreme-value distribution (Exercise 2.2.14) is
\[ z = -\ln(-\ln u). \]

J EXERCISES: COMPUTER USAGE

Find a computer program that gives reliable random digits (Exercise 2.1.8) or a table that contains randomly obtained digits. Use the inverse probability integral transformation IPIT to obtain a sample of 25 from each of the following distributions and see where the values fall in relation to the density function.

18 Exponential(3) distribution; Exercise 12.
19 Chi-square(2) distribution; Exercise 13.
20 Cauchy distribution; Exercise 14.
21 Pareto(2) distribution; Exercise 15.
22 Weibull(2) distribution; Exercise 16.
23 Extreme-value distribution; Exercise 17.

K PROBLEMS

Many problems involving changes of variable will be found in Sections 3.2 and 3.3, where some additional concepts will make the calculations very fruitful.

24 Consider the standard normal distribution for \((z_1, z_2)\) and define \((y_1, y_2)\) by
\[
\begin{pmatrix}
  y_1 \\
  y_2
\end{pmatrix}
= \begin{pmatrix}
  \mu_1 \\
  \mu_2
\end{pmatrix}
+ \begin{pmatrix}
  \sigma_1 \\
  \rho \sigma_2
\end{pmatrix}
\begin{pmatrix}
  z_1 \\
  z_2
\end{pmatrix},
\]
where \(-1 < \rho < 1\).

(a) Determine the elements of the matrix \(\Sigma\).

(b) Determine the elements of the inverse matrix \(\Sigma^{-1}\).

(c) Show that \((y_1, y_2)\) has the bivariate normal density \((\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)\) as given by
\[
f(y_1, y_2) = \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1 - \rho^2}} \exp\left\{ \frac{-1}{2(1 - \rho^2)} \left[ \frac{(y_2 - \mu_2)^2}{\sigma_2^2} \right. \right.
\]
\[-2\rho \frac{(y_1 - \mu_1)(y_2 - \mu_2)}{\sigma_1 \sigma_2} + \left. \frac{(y_2 - \mu_2)^2}{\sigma_2^2} \right]\right].\]
25 *Rotationally symmetrical distribution:* Let $g(r^2)$ be the density function for $(z_1, \ldots, z_k)$, where for brevity $r^2 = (z_1^2 + \cdots + z_k^2)$. Consider the transformation $y = \mu + \Gamma z$, where $|\Gamma| 
eq 0$ and $\Sigma = \Gamma \Gamma^\top$.

(a) Show that the density function $f$ for $y$ is given by

$$f(y) = |\Sigma|^{-1/2} g((y - \mu)^\top \Sigma^{-1}(y - \mu)).$$

(b) Record the density function for the case that $g$ describes the canonical Student$(n)$ distribution.

**SUPPLEMENTARY MATERIAL**

Supplement Section 2.5 is on pages 520–525. A brief survey of vector algebra is given and some discussion of inner products is included. The continuous and discrete components of a distribution have been discussed rather fully in preceding sections; the remaining component, the singular component, is examined for some interesting details.

**NOTES AND REFERENCES**

For general material on probability theory, see the Notes and References in Chapter 1. Simple notions from vector algebra are used intermittently through the chapters on probability and statistics; results needed for this are collected in Supplement Section 2.5.

In this chapter we have discussed ways of presenting probability distributions on the line, plane, and higher-dimensional real spaces. A familiar approach to this topic is in terms of random variables, a random variable being a real-valued function on a probability space (a more general definition is given in Chapter 3). With this other approach, the distributions on the line, plane, and higher-dimensional real spaces are viewed as coming from an antecedent probability space by means of functions (or random variables), although the functions as such are almost never presented. As our subject matter is the distributions themselves, we have chosen to discuss them directly. In Chapter 3 we shall examine how probability is mapped from one space to another by functions—by functions on a probability space.

The use of lowercase letters such as $x$ and $y$ for convenient reference to distributions should not be taken as suggesting that they be treated as identity or projection functions so as to conform to the language of random variables. Rather, they should be treated as the small $\omega$ that is used with the probability space $(\Omega, \mathcal{A}, P)$, $\omega$ being thought of as an arbitrary point in the space $\Omega$. The substantive material discussed will, of course, always be events and probabilities for events; the use of the lowercase letters provides a notationally convenient way of presenting the events as subsets on appropriate spaces.
In Chapter 1 we developed the basic language for handling probabilities. Then in Chapter 2 we examined various distributions of probability on the line, plane, and higher dimensions. This gives us the wherewithal to describe the common real and vector response variables in applications—in science, engineering, and everyday life.

Now consider a large batch of data in an application. It is common practice to calculate one, two, or several statistics that represent salient features of the data. For example, a psychologist with 30 interaction time measurements might calculate the average time interval. Or a biologist with 125 lengths for 1-year-old fish might calculate the average length representing the location of the values and a standard deviation representing how spread out the values are. In the first sections of Chapter 3 we examine various probability aspects of such data reduction.

We will then be in a position to discuss the statistical independence of two different responses. This important concept of probability and statistical theory is examined in the third section of Chapter 3. We will also see that independence gives us the means for building larger models from small initial models; call this compounding.

With the concepts of reduction and compounding in hand we will have the needed theory to derive the common discrete distributions—the binomial, the multinomial, and so on. These distributions are then derived in the concluding sections of Chapter 3.
3
MARGINAL PROBABILITY

3.1
DATA REDUCTION

In the preface to this chapter we commented on the practice of simplifying a large batch of data by calculating one, two, or several statistics that represent salient features of the data. In this section we examine some general ideas connected with this data reduction.

The process of reduction is, in fact, the reverse of that of observing more, as discussed in Section 1.1C. In this section we examine reduction briefly for the probability model in terms of the pure Boolean events and then in more detail using a sample space.

A REDUCTION IN TERMS OF EVENTS: AN EXAMPLE

Consider the tossing of two symmetric coins and suppose that we are interested initially in the face showing for each coin. The standard operations in Section 1.1 give the class \( \mathcal{B} = \mathcal{B}_2 \) of 16 events and symmetry gives the corresponding 16 probabilities, shown in Table 3.1.

Now suppose that we restrict our interest to events connected with the number of heads. Let \( E_1 \) be the event "exactly 1 heads." We then have the obvious events

\[ \emptyset, \quad E_0 = T_1T_2, \quad E_1 = H_1T_2 \lor T_1H_2, \quad E_2 = H_1H_2, \quad \mathbb{U}, \]

and we obtain the events

\[ E_1 \lor E_2 = (T_1T_2)^c, \quad E_2 \lor E_0 = H_1H_2 \lor T_1T_2, \quad E_0 \lor E_1 = (H_1H_2)^c \]

by the standard operations. These 8 events form a class \( \mathcal{B}_* \) closed under the standard

\[ \begin{array}{cccccccc}
\mathcal{B} & \mathcal{B}_* & P & \mathcal{B} & \mathcal{B}_* & P & \mathcal{B} & \mathcal{B}_* & P \\
\emptyset & \emptyset & 0 & H_1 & 1/2 & (H_1H_2)^c & E_2 & 3/4 \\
H_1H_2 & E_2 & 1/4 & T_1 & 1/2 & (H_1T_2)^c & \mathbb{U} & \mathbb{U} & 1 \\
H_1T_2 & 1/4 & H_2 & 1/2 & (H_1H_2)^c & E_2 & 3/4 \\
T_1H_2 & 1/4 & T_2 & 1/2 & (T_1T_2)^c & E_2 & 3/4 \\
T_1T_2 & E_0 & 1/4 & H_1H_2 \lor T_1T_2 & E_1 & 1/2 & \mathbb{U} & \mathbb{U} & 1 \\
& & & H_1T_2 \lor T_1H_2 & E_1 & 1/2 & \mathbb{U} & \mathbb{U} & 1
\end{array} \]
operations. Thus our reduced interest has taken us from the original class $\mathcal{B}$ of 16
events to the subclass $\mathcal{B}_\star$ of the 8 events still of interest, and correspondingly it has
restricted our attention to the probabilities for those 8 events.

More formally, our reduced interest has taken us from the original model $(\mathcal{B}, P)$ to
the marginal model $(\mathcal{B}_\star, P)$, where the values are as given in Table 3.2. Note that $\mathcal{B}_\star$
is a Boolean $\sigma$-algebra using the standard operations, and $P$ is the original measure $P$
but now restricted to the subclass $\mathcal{B}_\star$ of $\mathcal{B}$.

**TABLE 3.2**
Probabilities for the special events
of interest

<table>
<thead>
<tr>
<th>$\mathcal{B}_\star$</th>
<th>$P$</th>
<th>$\mathcal{B}_\star$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\emptyset$</td>
<td>0</td>
<td>$\mathcal{E}_0$</td>
<td>3/4</td>
</tr>
<tr>
<td>$E_0$</td>
<td>1/4</td>
<td>$\mathcal{E}_1$</td>
<td>1/2</td>
</tr>
<tr>
<td>$E_1$</td>
<td>1/2</td>
<td>$\mathcal{E}_2$</td>
<td>3/4</td>
</tr>
<tr>
<td>$E_2$</td>
<td>1/4</td>
<td>$\mathcal{U}$</td>
<td>1</td>
</tr>
</tbody>
</table>

Consider the reduction process more generally. Let $(\mathcal{B}, P)$ be an initial probability
model with Boolean $\sigma$-algebra $\mathcal{B}$ and probability measure $P$. Suppose that we are
interested only in certain events that form a subclass $\mathcal{B}_\star$ closed under the standard
operations. Then our reduced interest takes us from the initial model $(\mathcal{B}, P)$ to the
marginal model $(\mathcal{B}_\star, P)$, where $P$ is the original measure but now restricted to the
subclass $\mathcal{B}_\star$.

**B REDUCTION IN TERMS OF
SAMPLE SPACES: AN EXAMPLE**

Consider the coin-tossing example again and let $s = (x_1, x_2)$ designate the outcome,
where $x_1$ is the number of heads for the $i$th coin. Then the sample space is given by

$$ S = \{(x_1, x_2) : x_1 = 0, 1 \}; $$

and the probability measure for sample points is recorded on the left of Table 3.3.

Now suppose that we restrict our attention to events connected with the number of
heads. Let $y$ designate the number of heads and $Y$ be the function that calculates the
number of heads,

$$ y = Y(x_1, x_2) = x_1 + x_2. $$

**TABLE 3.3**
Probabilities for sample points

<table>
<thead>
<tr>
<th>$s$</th>
<th>$P((s)) = p(s)$</th>
<th>$Y$</th>
<th>$P_Y((y)) = p_Y(y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0, 0)</td>
<td>1/4</td>
<td>0</td>
<td>1/4</td>
</tr>
<tr>
<td>(1, 0)</td>
<td>1/4</td>
<td>1</td>
<td>1/2</td>
</tr>
<tr>
<td>(0, 1)</td>
<td>1/4</td>
<td>1</td>
<td>1/2</td>
</tr>
<tr>
<td>(1, 1)</td>
<td>1/4</td>
<td>2</td>
<td>1/4</td>
</tr>
</tbody>
</table>
The sample space for $\gamma$ is

$$3 = \{0, 1, 2\}.$$  

Note that $\gamma$ is the sum function mapping $\mathcal{S}$ into $\mathcal{Z}$ or even $R^2$ into $R$; see Figure 3.1 and Table 3.3.

**Table 3.4**  
Some events for $s$ and $\gamma$

<table>
<thead>
<tr>
<th>Event for $s$</th>
<th>Event for $\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>${(0, 0)}$</td>
<td>$\leftrightarrow$</td>
</tr>
<tr>
<td>${(1, 0), (0, 1)}$</td>
<td>$\leftrightarrow$</td>
</tr>
<tr>
<td>${(1, 1)}$</td>
<td>$\leftrightarrow$</td>
</tr>
</tbody>
</table>

Table 3.4 records some events in terms of new sample points $\gamma$ and in terms of original sample points $s$. The probability for an event in terms of new sample points is, of course, the probability for the event as recorded in terms of original sample points. The probabilities for the new sample points are recorded on the right in Table 3.3. This is the marginal distribution for $Y$ or for $\gamma = Y(s)$; see Figure 3.1.

Reduction is more complicated when sample spaces are used. But, sample spaces are very useful. In the remainder of this section we examine reduction from one sample space to another.

### C Reduction

Consider a random system that produces an outcome $s$, say the vector of interaction times in the psychologist's investigation. And suppose that we have a probability model $(\mathcal{S}, \mathcal{G}, P)$, where $\mathcal{S}$ is the sample space, $\mathcal{G}$ the class of events expressed as subsets, and $P$ the probability measure.

Now for the random system, suppose that we are interested only in events based on a certain function $Y$. Let $\mathcal{Z}$ be the sample space for $\gamma = Y(s)$; then $Y$ is a function that maps $\mathcal{S}$ into $\mathcal{Z}$; see Figure 3.2.

The set of values for the function $Y$, the set $\{Y(s) : s \in \mathcal{S}\}$ may not be the whole of the space $\mathcal{Z}$: let $C$ be the remaining portion of $\mathcal{Z}$; then $Y^{-1}(C) = \emptyset$. For the present it is convenient to ignore the set $C$ and assume that $Y$ maps $\mathcal{S}$ onto $\mathcal{Z}$. 

FIGURE 3.2
Mapping \( Y \) from space \( \mathcal{S} \) into space \( \mathcal{J} \). The event 'y' in \( B \) is the same as 's' in \( Y^{-1}(B) \).''

Consider the assertion that the function \( Y \) has the value \( y \); this is equivalent to the assertion that the original sample point is in the set

\[
Y^{-1}(y) = \{ s : Y(s) = y \},
\]

called the preimage of \( y \). And consider the assertion that the function \( Y \) has value in \( B \); this is equivalent to the assertion that the original sample point is in the set

\[
Y^{-1}(B) = \{ s : Y(s) \in B \},
\]
called the preimage of \( B \). Thus the event expressed as \( B \) on the new space is equivalent to the event expressed by \( Y^{-1}(B) \) on the initial space; call it an event for the function \( Y \).

Now consider repetitions of the random system. Values of the function \( Y \) will fall in a set \( B \) exactly as original sample points fall in the preimage set \( Y^{-1}(B) \). Thus the probability that attaches to the set \( B \) on the new space must be the probability as given for the preimage set \( Y^{-1}(B) \) on the initial space. This leads us to the following definition:

**DEFINITION 1**

The marginal probability for the set \( B \) of values for the function \( Y \) is

\[
P_Y(B) = P(Y^{-1}(B)).
\]

We have neglected a small technical problem. We might be interested in a very reasonable set \( B \) on the new space \( \mathcal{J} \) and yet have a very peculiar function \( Y \) such that the preimage set \( Y^{-1}(B) \) was not a set for which a probability was given. Clearly we want our function \( Y \) to be respectable so that such an embarrassment does not occur. Accordingly, we require that the function \( Y \) be measurable:

**DEFINITION 2**

A function \( Y \) from the space \( \mathcal{S} \) with \( \sigma \)-algebra \( \mathcal{G} \) to the space \( \mathcal{J} \) with \( \sigma \)-algebra \( \mathcal{B} \) is measurable if for each \( B \) in \( \mathcal{B} \), \( Y^{-1}(B) \) is in \( \mathcal{G} \).
Such a function on a probability space \((S, \mathcal{G}, P)\) is also called a random variable. All ordinary continuous functions and derived functions obtained by limits are measurable functions. For some details, see Supplement Section 3.6G.

We can now present the marginal probability model for the measurable function \(Y\). The model is

\[
(6) \quad (\mathcal{B}, P_Y),
\]

where

\[
(7) \quad P_Y = PY^{-1}
\]

is given by Definition 1; this is the marginal model for \(Y\) or for \(y = Y(s)\). Of course we must be sure that \(P_Y\) is a probability measure; but this is primarily a notational matter and is covered by some details in Section D.

**EXAMPLE 1**

*Uniform distribution on \((0, 1)^2\):* Consider the uniform distribution for \((x_1, x_2)\) on the unit square \((0, 1)^2\); see Figure 3.3. Find the probability that \(y = x_1 + x_2 \leq 0.5\). The set \(\{y \leq 0.5\}\) on the image space is equivalent to the set \(\{(x_1, x_2): x_1 + x_2 \leq 0.5\}\) on the initial space. This is the set of points to the lower left from the line \(x_1 + x_2 = 0.5\), and within the square it is just the triangle formed by \((0, 0), (0.5, 0), (0, 0.5)\). With the uniform distribution the probability can be calculated as a ratio of areas: \(\frac{1}{2}(0.5)(0.5)/1\), which is 0.125. Find the probability that \(0.9 < y < 1.2\). The set \(\{0.9 < y < 1.2\}\) on the image space is equivalent to the set \(\{(x_1, x_2): 0.9 < x_1 + x_2 < 1.2\}\) on the initial space. Calculated as a ratio of areas we obtain

\[
(1 - \frac{1}{2}(0.9)(0.9) - \frac{1}{2}(0.8)(0.8))/1 = 0.275.
\]

**D SOME PROPERTIES OF A FUNCTION**

Throughout our applications we will be interested in many different kinds of functions; as examples, consider the average \(\bar{Y}\) and sample variance \(s_Y^2 = \Sigma (y_i - \bar{Y})^2/(n - 1)\). Consider a function \(Y\) from a space \(S\) into a space \(\mathbb{R}\); see Figure 3.2. The set \(\{Y(s): s \in S\}\) of values of the function \(Y\) may not be the whole of the

![Figure 3.3](image)

*Sample space \(S = (0, 1)^2\). The events \(y \leq 0.5, 0.9 < y < 1.2\), where \(y = x_1 + x_2\).*
Sec. 3.1: Data Reduction

Let \( C \) be the remaining portion of \( \mathcal{J} \); then \( Y^{-1}(C) = \emptyset \). Again it is convenient to ignore the set \( C \) and assume that \( Y \) maps \( \mathcal{S} \) onto \( \mathcal{J} \).

Consider points in the space \( \mathcal{J} \), and the corresponding preimage sets:

\[
Y^{-1}(y') \leftrightarrow y',
\]

\[
Y^{-1}(y'') \leftrightarrow y''.
\]

If \( y' \neq y'' \), then \( Y^{-1}(y') \) and \( Y^{-1}(y'') \) are clearly disjoint. Also, each point in \( \mathcal{S} \) lies in some set, \( Y^{-1}(y) \). Thus the different sets \( Y^{-1}(y) \) form a partition of \( \mathcal{S} \), called the partition induced by \( Y \). And the sets in this partition are in one-to-one correspondence with the points of \( \mathcal{J} \); see Figure 3.2.

As an example, consider the sum function given by \( y = x_1 + x_2 \) mapping \( \mathbb{R}^2 \) into \( \mathbb{R} \). A line \( x_1 + x_2 = c \) is mapped into the point \( c \). Thus \( \mathbb{R}^2 \) is partitioned into the class of dashed lines as indicated in Figure 3.1. Knowing the value of \( x_1 + x_2 \) is equivalent to knowing that \( (x_1, x_2) \) is somewhere on the corresponding line; thus calculating the value of the function is equivalent to losing the information as to where the point \( (x_1, x_2) \) is on the line and thus to making a reduction on the information contained in the original outcome.

Note that the average function \( \bar{x} = (x_1 + x_2)/2 \) induces the same partition as does the sum function \( x_1 + x_2 \). For clearly the values of the two functions can be put into one-to-one correspondence: \( \bar{x} = (1/2)(x_1 + x_2) \) and \( (x_1 + x_2) = 2\bar{x} \).

Now consider sets on the image space \( \mathcal{J} \) and the corresponding preimage sets on the original space. We have the one-to-one correspondence

\[
Y^{-1}(B) \leftrightarrow B.
\]

These sets are describing the same proposition; that is, \( Y(s) \in B \) is equivalent to \( s \in Y^{-1}(B) \). Thus it is not surprising to find that the following relations hold:

\[
Y^{-1}\left( \bigcap_i B_i \right) = \bigcap_i Y^{-1}(B_i),
\]

\[
Y^{-1}\left( \bigcup_i B_i \right) = \bigcup_i Y^{-1}(B_i),
\]

\[
Y^{-1}(B^c) = (Y^{-1}(B))^c;
\]

thus \( Y^{-1} \) commutes with the standard operations. The proofs are straightforward and can be put on the abbreviated form used in Section 1.2D; for example,

\[
s \in Y^{-1}\left( \bigcap_i B_i \right) \quad \leftrightarrow \quad Y(s) \in \bigcap_i B_i \quad \leftrightarrow \quad Y(s) \in B_i \text{ for each } i
\]

\[
\quad \leftrightarrow \quad s \in Y^{-1}(B_i) \text{ for each } i \quad \leftrightarrow \quad s \in \bigcap_i Y^{-1}(B_i).
\]

Now consider the sets \( B \) that form a \( \sigma \)-algebra \( \mathcal{B} \) on the image space \( \mathcal{J} \). The relations (9) show that the preimage sets form a class that is closed under countable union, countable intersection, and complementation. Designate the class of preimage sets by

\[
Y^{-1}(\mathcal{B}) = \{ Y^{-1}(B) : B \in \mathcal{B} \};
\]

then \( Y^{-1}(\mathcal{B}) \) is a \( \sigma \)-algebra on the original space. Indeed in terms of propositions, the two \( \sigma \)-algebras are equivalent; they are just expressed on different spaces by the one-to-one correspondence

\[
Y^{-1}(B) \leftrightarrow B.
\]

We can then rephrase Definition 2 as:
DEFINITION 3

A function \( Y \) from \( \mathcal{S} \) with \( \sigma \)-algebra \( \mathcal{G} \) to \( \mathcal{I} \) with \( \sigma \)-algebra \( \mathcal{B} \) is **measurable** if 
\[ Y^{-1}(\mathcal{B}) \subseteq \mathcal{G}. \]

In other words, \( Y \) is measurable if it does not introduce any new "events."

For a measurable function \( Y \) the events \( B \) or, equivalently, \( Y^{-1}(B) \) will be called the **events for the function** \( Y \).

We are now in a position to see that \( P_Y \) in Section C is a probability measure on \( \mathcal{B} \). We know that \( P \) is a probability measure on \( \mathcal{G} \) and hence on the smaller \( \sigma \)-algebra \( Y^{-1}(\mathcal{B}) \). But \( P \) on this smaller \( \sigma \)-algebra is equivalent,
\[ P(Y^{-1}(B)) = P_Y(B). \]
to \( P_Y \) on the \( \sigma \)-algebra \( \mathcal{B} \). Thus \( P_Y \) is a probability measure.

**E Exercises**

1. Three symmetric coins are tossed and \( s = (x_1, x_2, x_3) \) is observed, where \( x_i \) is the indicator for heads on the \( i \)th coin.
   (a) Tabulate \( s \) and \( p(s) = p(s) \).
   (b) Let \( Y \) give the number of heads: \( Y(s) = x_1 + x_2 + x_3 = y \). Tabulate \( y \) and 
       \[ P_Y\{y\} = p_Y(y). \]

2. Two symmetric dice are tossed and \( s = (x_1, x_2) \) is observed, where \( x_i \) is the number of points on the \( i \)th die.
   (a) Partially tabulate \( s \) and \( p(s) \) in a convenient order for part (b).
   (b) Let \( Y \) give the total points: \( Y(s) = x_1 + x_2 = y \). Tabulate \( y \) and 
       \[ P_Y\{y\} = p_Y(y). \]

3. Consider the uniform distribution for \( (x_1, x_2) \) on the unit square \((0, 1)^2\). Let \( Y \) be the sum function: \( Y(x_1, x_2) = x_1 + x_2 = y \).
   (a) Calculate the distribution function \( F_Y(y) \) for \( 0 \leq y \leq 1 \); and for \( 1 \leq y \leq 2 \).
   (b) Calculate the density function \( f_Y \).

4. A biologist makes two measurements on a growth rate and obtains \((y_1, y_2)\).
   (a) Determine the preimage of \( y_1 + y_2 = c \); of \( y_1^2 + y_2^2 = d \); of \( (\Sigma y_i, \Sigma y_i^2) = (c, d) \).
   (b) Let \( y_{(1)} \) be the smallest of \( (y_1, y_2) \), and \( y_{(2)} \) the second smallest. Determine the preimage of \( y_{(1)} = c' \); of \( y_{(2)} = d' \); and of \( (y_{(1)}, y_{(2)}) = (c', d') \).

5. (continuation) Argue that \( (\Sigma y, \Sigma y^2) \) is a function equivalent to \( (y_{(1)}, y_{(2)}) \).

6. A physicist makes \( n \) measurements on a current flow and obtains \((y_1, \ldots, y_n)\). The physicist is interested in the sample mean \( \bar{y} \) and sample variance \( s_y^2 \):

\[
\bar{y} = \frac{\Sigma y_i}{n}, \quad s_y^2 = \frac{1}{n-1} \Sigma (y_i - \bar{y})^2.
\]

Compare with Exercise 1.2.17. Note that \( \bar{y} (\bar{y}, \ldots, \bar{y}) \) is the projection of \( y \) on the line \( \mathcal{L}(1) = \{a \cdot a \in \mathbb{R}\} \) and that \( (n-1)s_y^2 \) is the squared distance of \( y \) from the line \( \mathcal{L}(1) \).
   (a) For \( n = 3 \) determine the preimage of \( \bar{y} = c \) and show it on a sketch.
   (b) For \( n = 3 \) determine the preimage of \( s_y^2 = d \) and show it on a sketch.
   (c) For \( n = 3 \) determine the preimage of \( (\bar{y}, s_y^2) = (c, d) \) and show it on a sketch.

**F Problems**

7. Let \( Y \) be a function from \( \mathcal{S} \) into \( \mathcal{I} \). Use the method in Section D to prove that 
   \[ Y^{-1}(\mathcal{B}^c) = (Y^{-1}(\mathcal{B}))^c. \]
8 (continuation) Prove that \( Y^{-1}(\bigcup A_j) = \bigcup Y^{-1}(A_j) \).

9 Consider the following three functions from \( \mathbb{R}^n \) to \( \mathbb{R}^m \): \( (y_1, \ldots, y_m) \), where \( y_i \) is the \( i \)th smallest of \( y_1, \ldots, y_n \); \( (\sum_i y_i, \sum_i y_i^2, \ldots, \sum_i y_i^2) \); and \( (\sum_i y_i, \sum_{i<j} y_i y_j, \sum_{i<j<k} y_i y_j y_k, \ldots, y_1, \ldots, y_n) \). The algebra of symmetric functions shows that the last two functions are equivalent: in fact, if we know the value for say \( (\sum_i y_i, \sum_i y_i^2) \), then we can calculate the value of \( (\sum_i y_i, \sum_{i<j} y_i y_j) \), and conversely. Use an algebraic result concerning the roots of a polynomial to show that the first and third are equivalent.

10 A chemist makes two determinations on the logarithm of a concentration and obtains \((y_1, y_2)\). For simplicity assume that \((y_1, y_2)\) has the standard bivariate normal. Let \( W \) be the sum function \( W(y_1, y_2) = y_1 + y_2 = w \).

(a) Determine the set \( W^{-1}((\infty, w]) \) and sketch it on the plane.

(b) Use the rotational symmetry of the normal to calculate \( P(W^{-1}((\infty, w]) \) and express it in terms of the standard normal distribution function \( G \).

(c) Determine the distribution function \( F_w \) for \( w = y_1 + y_2 \). What distribution is this?

11 The relationships(9) are not restricted to the case with a countable index set. Show that \( Y^{-1}(\bigcap A_\alpha) = \bigcap Y^{-1}(A_\alpha) \) and \( Y^{-1}(\bigcup A_\alpha) = \bigcup Y^{-1}(A_\alpha) \), where \( \alpha \) ranges over some index set \( I \).

### 3.2

**MARGINAL PROBABILITY ON THE LINE AND PLANE**

Most scientific and industrial data come in the form of real numbers, a moderate-to-large collection of real numbers. Analysis of such data often involves a reduction to one, two, or several real numbers that represent significant features of the full data. Correspondingly, a probability description of possibilities for the full data leads to a probability description of possibilities for the reduced data, the marginal distribution for such data. In this section we examine some notation and methods for marginal distributions on real spaces. We organize this in terms of distribution functions, probability functions, and probability density functions. Recall the basic formulas (3.1.5) and (3.1.7) for marginal probabilities: \( P_Y = P_{Y^{-1}} \).

### A MARGINAL DISTRIBUTION FUNCTIONS

Consider a basic distribution on a space \( \mathbb{S} \) with probabilities given by \( P \), say the distribution for \((x_1, \ldots, x_{10})\) on \( R^{10} \), where \( x_i \) is the yield of barley on an \( i \)th plot of land. Suppose first that we are interested in the marginal distribution for a real-valued function \( X \), say the sum function \( X(x_1, \ldots, x_{10}) = \sum_1^{10} x_i \). Let \( F_X \) be the distribution function:

\[
F_X(x) = P_X((-\infty, x]) = P(X^{-1}((-\infty, x]) = P\{ s : -\infty < X(s) \leq x \}.
\]

Thus to evaluate the distribution function we must calculate the probability for the set \( \{ s : -\infty < X(s) \leq x \} \), the preimage set of \((-\infty, x]\); this is usually accomplished by summing values of a probability function or integrating a probability density function.
Now suppose that we are interested in the marginal distribution for a pair of functions $(X, Y)$. Let $F_{XY}$ be the distribution function:

$$F_{XY}(x, y) = P_X([-\infty, x] \times (-\infty, y])$$

$$= P(X^{-1}(-\infty, x] \cap Y^{-1}(-\infty, y])$$

$$= P\{s : -\infty < X(s) \leq x, -\infty < Y(s) \leq y\}. $$

This is usually evaluated by summation or by integration.

**EXAMPLE 1**

**Uniform distribution on** $(0, 1)^2$: Consider a uniform distribution for $(x_1, x_2)$ over the unit square $(0, 1)^2$. And suppose that we want the distribution of $y = Y(x_1, x_2) = x_1 + x_2$. Let $F_Y$ be the marginal distribution function. Then following the pattern of calculation in Example 3.1.1, we obtain

$$F_Y(y) = 0 \quad \text{if} \quad y < 0,$$

$$= \frac{y^2}{2} \quad 0 \leq y < 1,$$

$$= 1 - \frac{(2 - y)^2}{2} \quad 1 \leq y < 2,$$

$$= 1 \quad 2 \leq y.$$

See Figure 3.4. The density function can be obtained by differentiation:

**FIGURE 3.4**

Probability in the square $(0, 1)^2$ is projected by $y = Y(x_1, x_2) = x_1 + x_2$ onto the real line. The distribution function at $y$, $F_Y(y)$, gives the probability for the pre-image set $Y^{-1}(-\infty, y)$ as marked.
Sec. 3.2: Marginal Probability on the Line and Plane 107

\[ f_Y(y) = \begin{cases} 
    y & 0 \leq y < 1, \\
    2 - y & 1 \leq y < 2, \\
    0 & \text{otherwise.}
\end{cases} \]

This is a triangular distribution on \((0, 2)\) with mode or peak probability at \(y = 1\).

**EXAMPLE 2**

**Square of a standard normal:** Consider a standard normal distribution for \(z\) on \(\mathbb{R}\). And suppose that we want the distribution of \(y = Y(z) = z^2\). Let \(F_Y\) be the marginal distribution function; then

\[ F_Y(y) = P_Y([-\infty, y]) = P(Y^{-1}([-\infty, y])) = P([-\sqrt{y}, \sqrt{y}]) \text{ if } y \geq 0. \]

See the function \(y = z^2\) in Figure 3.5 and note that the function is two to one generally and that the preimage of \([-\infty, y]\) is \([-\sqrt{y}, \sqrt{y}]\) for \(y \geq 0\). We can calculate directly from the standard normal distribution:

\[ F_Y(y) = \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = 2 \int_0^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \text{ if } y \geq 0. \]

The density function can be obtained by differentiation:

\[ f_Y(y) = 2 \frac{1}{2} y^{-1/2} \frac{1}{\sqrt{2\pi}} e^{-y/2} = (1/y) \Gamma\left(\frac{1}{2}\right) \left(\frac{y}{2}\right)^{(1/2)-1} e^{-y/2} \frac{1}{2} \text{ if } 0 < y, \]

\[ = 0 \text{ if } y \leq 0. \]

Thus \(y/2\) is gamma\((1/2)\) and \(y\) is chi square\((1)\); see Exercises 2.2.15 and 2.2.16.

---

Sometimes we are interested in very simple functions, coordinate projection functions. Consider a distribution on the plane with distribution function \(F\); then

\[ F(x, y) = P((-\infty, x] \times (-\infty, y]). \]

And suppose that we are interested in only the first coordinate, \(x\). As a function of \((x, y)\), \(x\) is the *projection* of \(\mathbb{R} \times \mathbb{R}\) onto the first space in the product; by restricting

![Figure 3.5](image)

Function \(y = Y(z) = z^2\). The preimage of \((-\infty, y]\) is \([-\sqrt{y}, +\sqrt{y}]\) for \(y \geq 0\).
attention to the first coordinate, \( x \), we are directly ignoring the second coordinate, \( y \). Correspondingly, the probability distribution on the plane is effectively collapsed or projected onto the first axis; see Figure 3.6. Let \( F_1 \) be the marginal distribution function for the first coordinate \( x \); then

\[
F_1(x) = P((-\infty, x] \times \mathbb{R}) = \lim_{y \to \infty} F(x, y) = F(x, \infty).
\]

Analogously, let \( F_2 \) be the marginal distribution function for the second coordinate \( y \); then

\[
F_2(y) = F(\infty, y).
\]

Now consider a distribution on \( \mathbb{R}^k \) with distribution function \( F \); then

\[
F(y_1, \ldots, y_k) = P((-\infty, y_1] \times \cdots \times (-\infty, y_k]).
\]

And suppose first that we are interested in just a single coordinate, say, the first, \( y_1 \). The first coordinate \( y_1 \) as a function of \( (y_1, \ldots, y_k) \) is the projection of \( \mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R} \) onto the first space in the product. Let \( F_1 \) be the marginal distribution function for the first coordinate \( y_1 \); then

\[
F_1(y_1) = P((-\infty, y_1] \times \mathbb{R} \times \cdots \times \mathbb{R}) = F(y_1, \infty, \ldots, \infty).
\]

And suppose now that we are interested in just two coordinates, say the first two, \( y_1 \) and \( y_2 \). The first coordinate pair \( (y_1, y_2) \) as a function of \( (y_1, \ldots, y_k) \) is the projection of \( \mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R} \) onto the first space \( \mathbb{R} \times \mathbb{R} \). Let \( F_{12} \) be the marginal distribution function for the first two coordinates \( (y_1, y_2) \); then

\[
F_{12}(y_1, y_2) = P((-\infty, y_1] \times (-\infty, y_2] \times \mathbb{R} \times \cdots \times \mathbb{R})
\]

\[
= F(y_1, y_2, \infty, \ldots, \infty).
\]

And analogously, we can obtain, for example, the marginal distribution function \( F_{137} \) for the coordinate triple \( (y_1, y_3, y_7) \).

**EXAMPLE 3**

**Orientation with two symmetric coins:** Consider the angles \((y_1, y_2)\) when a pair of symmetric coins are tossed twice (Example 2.3.2). The marginal distribution function \( F_1(y_1) \) can be obtained from \( F(y_1, y_2) \) by taking \( y_2 \) to \( \infty \); this gives

- \( F_1(y_1) = 0 \) if \( y_1 < 0 \),
- \( \frac{y_1}{180} \) if \( 0 \leq y_1 < 180 \),
- \( 1 \) if \( 180 \leq y_1 \).
B MARGINAL PROBABILITY FUNCTIONS

The most convenient way of handling discrete distributions is almost always in terms of the probability function that records the discrete chunks of probability. Consider a basic discrete distribution on a space $\mathcal{S}$ with probability measure $P$ and probability function $p$. Suppose first that we are interested in the marginal distribution for a real-valued function $X$. Let $p_X$ be the probability function

$$ p_X(x) = P_X(x) = P(X^{-1}(x)) = \sum_{x \in X(s)} p(s), \quad (7) $$

where the summation is, of course, over all points $s$ with the given value for $X(s)$.

Now suppose that we are interested in the marginal distribution for a pair of functions $(X, Y)$. Then

$$ p_{X,Y}(x, y) = P(X^{-1}(x) \cap Y^{-1}(y)) = \sum_{x \in X(s)} \sum_{y \in Y(s)} p(s). \quad (8) $$

EXAMPLE 4

Tossing two dice: Consider the tossing of two symmetric dice and the recording of $(y_1, y_2)$, where $y_i$ is the number of points on the $i$th die (Example 2.3.3). Then $p(y_1, y_2) = 1/36$ on $\{1, 2, 3, 4, 5, 6\}^2$ and zero elsewhere. And suppose that we are interested only in the total points $t = T(y_1, y_2) = y_1 + y_2$ for the two dice; then

$$ p_T(t) = p(T^{-1}(t)), $$

which is the summation of $1/36$ for each possible sample point having total $t$. As our initial distribution is symmetric, we can calculate by counting sample points:

$$ p_T(t) = \frac{\text{number of points } (y_1, y_2) \text{ having } y_1 + y_2 = t}{\text{number of points } (y_1, y_2)}, $$

where $y_1$ and $y_2$ take values $1, 2, \ldots, 6$. The probabilities are easily enumerated, giving the results shown in Table 3.5.

<table>
<thead>
<tr>
<th>$t$</th>
<th>$p_T(t)$</th>
<th>$t$</th>
<th>$p_T(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1/36</td>
<td>8</td>
<td>5/36</td>
</tr>
<tr>
<td>3</td>
<td>2/36</td>
<td>9</td>
<td>4/36</td>
</tr>
<tr>
<td>4</td>
<td>3/36</td>
<td>10</td>
<td>3/36</td>
</tr>
<tr>
<td>5</td>
<td>4/36</td>
<td>11</td>
<td>2/36</td>
</tr>
<tr>
<td>6</td>
<td>5/36</td>
<td>12</td>
<td>1/36</td>
</tr>
<tr>
<td>7</td>
<td>6/36</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The preceding is a special case of the convolution formula:

**EXAMPLE 5**

**Convolution formula:** Consider a discrete distribution on \( R^2 \) with probability function \( p \). The probability function for the sum \( t = T(y_1, y_2) = y_1 + y_2 \) is given by

\[
 p_T(t) = \sum_{t = y_1 + y_2} p(y_1, y_2) = \sum_{y_2} p(t - y_2, y_2).
\]

Sometimes we are interested in very simple functions—coordinate projection functions. Consider a discrete distribution on the plane with probability function \( p \) giving the probability \( p(x, y) \) at the point \( (x, y) \). The marginal probability functions are then given by

\[
 p_1(x) = \sum_y p(x, y),
\]

\[
 p_2(y) = \sum_x p(x, y);
\]

for example, \( p_1(x) \) is obtained by summing the probabilities for all the discrete points \( (x, y) \) having first coordinate \( x \); see Figure 3.6.

Now consider a discrete distribution on \( R^k \) with probability function \( p \). The marginal probability function for the first coordinate is then given by

\[
 p_1(y_1) = \sum_{y_2 \ldots y_k} p(y_1, \ldots, y_k)
\]

and for the first two coordinates by

\[
 p_{12}(y_1, y_2) = \sum_{y_3 \ldots y_k} p(y_1, \ldots, y_k).
\]

Thus to obtain the marginal probability function we merely sum out the coordinates that are not of interest.

**C MARGINAL PROBABILITY DENSITY FOR A PROJECTION**

Consider an absolutely continuous distribution on the plane with probability density function \( f \); then

\[
P(A) = \int_A f(x, y) \, dx \, dy.
\]

And suppose that we are interested in only the first coordinate, \( x \). The first coordinate, \( x \), is the projection of \( R \times R \) onto the first factor, \( R \); thus our interest is in the probability
on the plane as projected or collapsed onto the first coordinate axis; see Figure 3.6 again. Let \( F_1 \) be the distribution function and \( f_1 \) the density function; then

\[
F_1(x) = \int_{(\mathbb{R} \times \mathbb{R})} f(t, y) \, dt \, dy = \int_{\mathbb{R}} \left[ \int_{-\infty}^{\infty} f(t, y) \, dy \right] dt,
\]

and thus

\[
(13) \quad f_1(x) = \int_{-\infty}^{\infty} f(x, y) \, dy.
\]

With a continuous density function \( f \) we can think of the probability in the strip between \( x \) and \( x + dx \); see Figure 3.7. The volume above this strip can be integrated and represented as area above the interval \((x, x + dx)\); this gives

\[
(14) \quad f_1(x) \, dx = \int_{-\infty}^{\infty} f(x, y) \, dy \cdot dx.
\]

More formally, we take this to be a probability differential, the marginal probability differential for the first coordinate \( x \).

Analogously, let \( f_2 \) be the probability density function for the second coordinate \( y \). The probability density is given by

\[
(15) \quad f_2(y) = \int_{-\infty}^{\infty} f(x, y) \, dx,
\]

and the probability differential by

\[
(16) \quad f_2(y) \, dy = \int_{-\infty}^{\infty} f(x, y) \, dx \cdot dy.
\]

**FIGURE 3.7**

Joint probability density \( f \); marginal density \( f_1 \) for \( x \) by integrating out \( y \).
Now consider an absolutely continuous distribution on $R^k$ with probability density function $f$. Then the marginal density for the first coordinate is given by

$$f_1(y_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(y_1, y_2, \ldots, y_k) \, dy_2 \cdots dy_k$$

and for the first two coordinates, say, by

$$f_{12}(y_1, y_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(y_1, y_2, y_3, \ldots, y_k) \, dy_3 \cdots dy_k.$$  

**Example 6**

**Bivariate distribution:** Consider an absolutely continuous distribution for $(x, y)$ with density function

$$f(x, y) = 6x \quad \text{if } 0 < x < 1, \quad 0 < y < 1 - x,$$

and zero otherwise. This is a distribution effectively over the triangle $(0, 0), (1, 0), (0, 1)$; see Figure 3.8. The marginal densities are

$$f_1(x) = \int_0^{1-x} 6x \, dy = 6xy \bigg|_{y=0}^{1-x} = 6x(1 - x), \quad 0 < x < 1,$$

and zero otherwise [this is the beta(2, 2) distribution], and

**Figure 3.8**
Joint and marginal densities in Example 6.
(20) \[ f_2(y) = \int_0^{1-y} 6x \, dx = 3x^2 \bigg|_{x=0}^{1-y} = 3(1 - y)^2, \quad 0 < y < 1, \]
and zero otherwise [this is the \( \text{beta}(1, 3) \) distribution].

Now suppose that the distribution on \( R^2 \) has come from some initial space \( \mathbb{S} \) with probability measure \( P \). Let \( (X, Y) \) be the function that gives the distribution on \( R^2 \); then

\[ P_{XY}(A) = P\{s : (X(s), Y(s)) \in A \} = \int_A f_{XY}(x, y) \, dx \, dy. \]

If the marginal density \( f_{XY} \) is available, then the marginal densities for \( X \) and \( Y \) alone are

\[ f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) \, dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y) \, dx. \]

**D MARGINAL PROBABILITY DENSITY FOR A GENERAL FUNCTION**

We now examine probability densities for functions more complicated than the simple coordinate projections just discussed. First consider an absolutely continuous distribution for \( x \) on the real line \( R \) with density function \( g \). And suppose that we are interested in the distribution for \( y = Y(x) \) produced by the function \( Y \). If \( Y \) is a monotone, continuously differentiable one-to-one function, then we are in the change of variable (or relabeling) situation discussed in Section 2.4C and there is no reduction. Recall the density function:

(21) \[ f_Y(y) = g(X(y)) \left| \frac{dx}{dy} \right|_+, \]

where \( X \) is the inverse function \( Y^{-1} \).

If the function \( Y \) is a several-to-one mapping, then several different neighborhoods for \( x \) can map into a single neighborhood for \( y \); see Figure 3.9. The probability for the neighborhood of \( y \) is then calculated by adding the probabilities for the corresponding neighborhoods of \( x \); with differentiability, formula (21) can then be applied component by component. Rather than present here a general formula that can be somewhat

![Figure 3.9](image)

The probability for a neighborhood of \( y \) is the sum of the probabilities for the corresponding two neighborhoods for \( x \).
cumbersome, we shall present a simple example that illustrates the use of the principles from Section 3.1 and the change of variable in formula (21).

EXAMPLE 7

Square function: Consider a distribution for $x$ on the real line $R$ with density function $g$. And suppose that we want the distribution of $y = Y(x) = x^2$; compare with Example 2 and see Figure 3.9. The preimage of $y (> 0)$ consists of two points, $-\sqrt{y}$ and $+\sqrt{y}$. The probability at $y$ from the first of these points is

\[ g(-\sqrt{y}) \frac{d}{dy} -\sqrt{y} \, dy = g(-\sqrt{y})(1/2)y^{-1/2} \, dy, \]

and the probability at $y$ from the second of the points is

\[ g(\sqrt{y}) \frac{d}{dy} \sqrt{y} \, dy = g(\sqrt{y})(1/2)y^{-1/2} \, dy. \]

Thus the total probability at $y$ is

\[ f_Y(y) \, dy = (g(\sqrt{y}) + g(-\sqrt{y}))(1/2)y^{-1/2} \, dy. \]  

(22)

If $g$ is standard normal, then substitution gives the result obtained in Example 2 by the distribution function route.

Now consider an absolutely continuous distribution for a multiple response $(x_1, \ldots, x_k)$ on $R^k$ with density function $g$; for notational convenience we examine the case $k = 2$: for example, reaction time and reaction intensity in a psychology experiment.

Suppose that we are interested in the distribution for $(y_1, y_2) = (Y_1(x_1, x_2), Y_2(x_1, x_2))$ produced by the function $(Y_1, Y_2)$. If $(Y_1, Y_2)$ is a one-to-one continuously differentiable function, then we are in the change of variable (or relabeling) situation in Section 2.4E and there is no reduction. Recall the density function

\[ f_{Y_1,Y_2}(y_1, y_2) = g(X_1(y_1, y_2), X_2(y_1, y_2)) \left| \frac{\partial X_1, X_2}{\partial y_1, y_2} \right|, \]

(23)

where $(X_1, X_2)$ is the inverse function. If the function $(Y_1, Y_2)$ is a several-to-one mapping, then several neighborhoods for $(x_1, x_2)$ can map into a single neighborhood for $(y_1, y_2)$. The probability for this neighborhood is then calculated by adding the probabilities for the corresponding neighborhoods for $(x_1, x_2)$; the procedure follows as in Example 7.

Now suppose that we are interested in the distribution for $y = Y(x_1, x_2)$ produced by the function $Y$ from $R^2$ into $R$. To have the value $y$ for $Y$ is to know that $(x_1, x_2)$ is somewhere on the preimage contour $Y^{-1}(x_1, x_2)$ as indicated in Figure 3.10. In fact, the function $Y$ partitions the plane into the class of such preimage contours. If the function $Y$ has reasonable properties, then the contours will be continuous curves, as in Figure 3.10.

The probability for an interval $(y, y + dy)$ is given by the probability for the corresponding range of preimage contours, the curved band, as shown in Figure 3.10. To calculate this probability it is convenient to find a second function $Y_2$ whose preimage contours are cross sectional to those for $Y$. Thus we try to complete the initial
function \( Y \) so as to obtain a function \((Y, Y_2)\) with the one-to-one differentiable properties needed for formula (23). We then integrate out the unwanted coordinate as done in Section C:

\[
(24) \quad f_{Y_2}(y) = \int_{-\infty}^{\infty} g(X_1(y, y_2), X_2(y, y_2)) \left| \frac{\partial x_1}{\partial y}, \frac{\partial x_2}{\partial y_2} \right| \, dy_2.
\]

The method generalizes immediately to finding the density for several functions on \( \mathbb{R}^k \).

**EXAMPLE 8**

**Convolution formula**: Consider a distribution for \((x_1, x_2)\) with density function \( g \) on \( \mathbb{R}^2 \). And suppose that we want the distribution of \( t = T(x_1, x_2) = x_1 + x_2 \). A simple choice of second variable to complete this as a one-to-one transformation is \( u = x_2 \); thus

\[
t = x_1 + x_2, \quad x_1 = t - u, \quad u = x_2, \quad x_2 = u.
\]

The Jacobian matrix is

\[
\begin{pmatrix}
\frac{\partial x_1}{\partial t} & \frac{\partial x_1}{\partial u} \\
\frac{\partial x_2}{\partial t} & \frac{\partial x_2}{\partial u}
\end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix},
\]

which is a constant (the transformation is linear); the determinant is 1. Thus \( g(x_1, x_2) \, dx_1 \, dx_2 = g(t - u, u) \, 1 \, dt \, du \), and the marginal density for \( t = x_1 + x_2 \) is

\[
(25) \quad f_T(t) = \int_{-\infty}^{\infty} g(t - u, u) \, du.
\]

This is called the **convolution formula** and it gives the density for \( t = x_1 + x_2 \) in terms of the joint density for \((x_1, x_2)\).
EXAMPLE 9

Consider the standard normal distribution for \((z_1, z_2)\) on \(R^2\). And suppose that we are interested in \(r\) and also interested in \(a\), where

\[
    r = (z_1^2 + z_2^2)^{1/2}, \quad z_1 = r \cos a,
\]

\[
    a = \tan^{-1} \left( \frac{z_2}{z_1} \right), \quad z_2 = r \sin a
\]

for \(0 < r < \infty\) and \(0 < a < 2\pi\). The differential transformation is

\[
    \begin{pmatrix}
    dz_1 \\
    dz_2
    \end{pmatrix} = \begin{pmatrix}
    \cos a & -r \sin a \\
    \sin a & r \cos a
    \end{pmatrix} \begin{pmatrix}
    dr \\
    da
    \end{pmatrix},
\]

the Jacobian determinant is \(r(\cos^2 a + \sin^2 a) = r\), and thus

\[
    dz_1 \, dz_2 = r \, dr \, da;
\]

compare this with the geometrical derivation in Section 2.3E and recall Example 2.4.7. Thus

\[
    \frac{1}{2\pi} e^{-(z_1^2 + z_2^2)/2} \, dz_1 \, dz_2 = \frac{1}{2\pi} e^{-r^2/2r} \, dr \, da = e^{-r^2/2r} \, dr \cdot \frac{1}{2\pi} \, da
\]

on the product set \((0, \infty) \times (0, 2\pi)\) and zero elsewhere. The marginal distributions are then available immediately:

\[
    (26) \quad f_R(r) \, dr = e^{-r^2/2r} \, dr, \quad f_A(a) = \frac{1}{2\pi} \, da
\]

on the appropriate ranges. Note that \(r^2/2\) is exponential(1) (Exercise 2.2.7), that \(r^2\) is chi-square(2) (Exercise 2.2.16), and that \(a\) is uniform(0, 2\pi) (Example 2.2.1).

E EXERCISES: DISTRIBUTION FUNCTIONS

1. Let \(H_1\) and \(H_2\) be distribution functions on \(R\). Then by Exercise 2.3.1, \(F(y_1, y_2) = H_1(y_1)H_2(y_2)\) is a distribution function on \(R^2\). Calculate the marginal distribution functions \(F_1\) and \(F_2\).

2. A distribution on \(R^2\) has probability function \(p\) as given in Exercise 2.3.2. Calculate the marginal distribution function \(F_1\) at \(y_1 = 0, 1, 2, 3\). Calculate \(F_2\) at \(y_2 = 0, 1, 2, 3\).

3. A distribution on \(R^2\) has distribution function \(F\) given by

\[
    F(y_1, y_2) = 1 - (1 + y_1)^{-1} - (1 + y_2)^{-1} + (1 + y_1 + y_2)^{-1}, \quad y_1, y_2 \geq 0,
\]

and zero otherwise.

(a) Calculate the marginal distribution functions \(F_1\) and \(F_2\).

(b) Calculate the marginal density functions.

4. The standard normal distribution on \(R^k\) has distribution function \(F\) given by \(F(z_1, \ldots, z_k) = G(z_1) \cdots G(z_k)\), where \(G\) is the standard normal distribution function on \(R\).

(a) Calculate the marginal distribution function \(F_1\).

(b) Calculate the marginal distribution function \(F_{12}\).

5. A distribution on \(R^n\) has distribution function \(F\) given by \(F(y_1, \ldots, y_n) = H_1(y_1) \cdots H_k(y_k)\), where the \(H_i\) are distribution functions on \(R\). Calculate the marginal distribution functions \(F_1, F_{12}, \text{ and } F_{123}\).

6. A distribution on \(R^2\) has distribution function \(F\) given by
Sec. 3.2: Marginal Probability on the Line and Plane 117

\[ F(y_1, y_2) = 0, \quad y_2 < 0, \]
\[ = (1/2)G(y_1), \quad 0 \leq y_2 < 1, \]
\[ = (1/2)G(y_1) + (1/2)H(y_1), \quad 1 \leq y_2 < \infty, \]

where \( G \) and \( H \) are distribution functions on \( R \).

(a) Find the marginal distribution function \( F_1 \).
(b) Find the marginal distribution function \( F_2 \).

7 Let \( F \) be a distribution function on \( R^2 \). And let \( W \) be the function that carries \((y_1, y_2)\) into \( w = \max(y_1, y_2) \) and \( X \) be the function that carries \((y_1, y_2)\) into \( x = \min(y_1, y_2) \). Show that the marginal distribution functions \( F_W \) and \( F_X \) satisfy

\[ F_W(w) = F(w, w), \quad F_X(x) = F(x, \infty) + F(\infty, x) - F(x, x). \]

8 Let \( F \) be a distribution function on \( R^k \). And let \( W \) be the function that carries \((y_1, \ldots, y_k)\) into max \( y_j \). Show that the marginal distribution function \( F_W \) satisfies \( F_W(w) = F(w, \ldots, w) \).

A distribution on \( R^2 \) has \( f(y_1, y_2) = \pi^{-1} \) for \( y_1^2 + y_2^2 < 1 \) and zero otherwise.

(b) Calculate the distribution function \( H \) for \( r = (y_1^2 + y_2^2)^{1/2} \).

Consider a uniform distribution inside the sphere \( y_1^2 + y_2^2 + y_3^2 = 1 \).

(a) Calculate the distribution function \( G \) for \( w = y_1^2 + y_2^2 + y_3^2 \).
(b) Calculate the distribution function \( H \) for \( r = (y_1^2 + y_2^2 + y_3^2)^{1/2} \).

11 The Pareto(\( \alpha \)) distribution function on \( R^2 \) is given by \( \alpha > 0 \)

\[ F(y_1, y_2) = 1 - (1 + y_1)^{-\alpha} - (1 + y_2)^{-\alpha} + (1 + y_1 + y_2)^{-\alpha}, \quad y_1 \geq 0, \]

and zero otherwise. Determine the marginal distribution functions \( F_1 \) and \( F_2 \); verify that each is Pareto(\( \alpha \)) on the real line \( R \).

\[ F \] exercises: Probability functions

12 A symmetric die and a symmetric coin are tossed and \((x_1, x_2)\) is recorded, where \( x_i \) is the number of points on the die and \( x_2 \) is the number of heads showing. Record the symmetric probability function for \((x_1, x_2)\). Calculate the probability function \( p_Y \) for the total \( y = x_1 + x_2 \).

13 A distribution on \( R^2 \) has probability function \( p \) as given in Exercise 2.3.2. Calculate the marginal probability functions \( p_1 \) and \( p_2 \).

14 Suppose that \( f \) is a symmetric density on \( R^2 \): \( f(y_1, y_2) = f(y_2, y_1) \).
(a) Show that \( P(y_1 < y_2) = P(y_2 < y_1) = 1/2 \).
(b) Let \((r_1, r_2) = (1, 2)\) if \( y_1 < y_2 \) and \((2, 1)\) if \( y_1 > y_2 \). Show that the probability function \( p \) for \((r_1, r_2)\) is given by

\[ p(r_1, r_2) = 1/2 \quad \text{at (1, 2) and (2, 1)} \]

and zero otherwise.

15 Suppose that \( f \) is a density on \( R^k \) and is symmetric \( f(y_1, \ldots, y_k) = f(y_{i_1}, \ldots, y_{i_k}) \), where \((i_1, \ldots, i_k)\) is any permutation of \((1, \ldots, k)\).
(a) Show that \( P(y_{i_1} < \cdots < y_{i_k}) = 1/k! \).
(b) Let \( r(y) = (r_1, \ldots, r_k) \) be a function of \( y \) such that \( y_{i_h} \) is the \( r_{i_h} \)th smallest of \( y_1, \ldots, y_k \). The vector \( r \) is called the rank statistic; determine the probability function for \( r \).

\[ G \] exercises: Probability density functions

16 Consider a distribution on \( R^2 \) with density \( f \) given by
118 Chap. 3: Marginal Probability

\[ f(y_1, y_2) = 18e^{-6y_1 - 3y_2}, \quad y_1 > 0, y_2 > 0. \]

and zero otherwise. Determine the marginal densities \( f_1 \) and \( f_2 \).

17 Consider a distribution on \( R^2 \) with density \( f \) given by

\[ f(y_1, y_2) = 2e^{-y_1 - y_2}, \quad 0 < y_1 < y_2 < \infty, \]

and zero otherwise. Determine the marginal densities \( f_1 \) and \( f_2 \).

18 Let \( f \) be the density function for the uniform distribution over the disk \( y_1^2 + y_2^2 < 1 \) in \( R^2 \). Calculate the marginal densities \( f_1 \) and \( f_2 \).

19 Let \( f \) be the density function for the uniform distribution over the unit ball \( y_1^2 + y_2^2 + y_3^2 < 1 \) in \( R^3 \).

(a) Determine the marginal density \( f_{12} \).

(b) Determine the marginal density \( f_1 \).

20 Consider the density \( f \) on \( R^k \) given by \( f(y_1, \ldots, y_k) = f_1(y_1) \cdots f_k(y_k) \), where each \( f_i \) is a density on \( R \). Determine the marginal densities \( f_1, f_{12}, \) and \( f_{123} \).

21 Consider the Pareto(\( \alpha \)) distribution on \( R^2 \) with density \( f \) given by \( (\alpha > 0) \)

\[ f(y_1, y_2) = \alpha(\alpha + 1) (1 + y_1 + y_2)^{-\alpha - 2}, \quad y_1 > 0, y_2 > 0, \]

and zero otherwise. Show that each of the marginal densities \( f_1 \) and \( f_2 \) is Pareto(\( \alpha \)) on \( R \).

22 Consider the canonical Student(\( n \)) distribution for \( (y_1, y_2) \) on \( R^2 \) (Problem 2.3.13). Show that the marginal density \( f_1 \) is canonical Student(\( n \)) on \( R \).

23 Consider a uniform distribution for \( (u_1, u_2) \) over the unit square \( (0, 1)^2 \). Use the convolution formula to obtain the distribution of \( y = u_1 + u_2 \).

H PROBLEMS

24 Consider the Pareto(\( \alpha \)) distribution for \( (y_1, \ldots, y_k) \) on \( R^k \). Show that the marginal distribution of \( (y_1, \ldots, y_{k-1}) \) is Pareto(\( \alpha \)) on \( R^{k-1} \). Deduce that \( (y_1, y_2) \) is Pareto(\( \alpha \)) on \( R^2 \) and that \( y_1 \) is Pareto(\( \alpha \)) on \( R \). See Problem 2.3.14.

25 Consider the canonical Student(\( n \)) distribution for \( (y_1, \ldots, y_k) \) on \( R^k \). Show that the marginal distribution of \( (y_1, \ldots, y_{k-1}) \) is canonical Student(\( n \)) on \( R^{k-1} \). Deduce that \( (y_1, y_2) \) is canonical Student(\( n \)) on \( R^2 \) and that \( y_1 \) is canonical Student(\( n \)) on \( R \). See Problem 2.3.15.

26 Consider the Dirichlet(\( r_1, \ldots, r_{k+1} \)) distribution for \( (y_1, \ldots, y_k) \) on \( R^k \). Prove the convolution property: the marginal distribution of \( (y_1 + y_2, y_2 + \ldots, y_k) \) is Dirichlet(\( r_1 + r_2, r_2, \ldots, r_{k+1} \)).

27 Sum of chi-squares: Consider the distribution for \( (\chi_1^2, \chi_2^2) \) with density \( f \) on \( R^2 \) given by

\[ f(\chi_1^2, \chi_2^2) = \Gamma^{-1}(1/2) \left( \frac{\chi_1}{2} \right)^{(m-2)/2} \left( \frac{\chi_2}{2} \right)^{(n-2)/2} \left( \frac{1}{2} \right)^{(m+n-2)/2} \]

for \( \chi_1^2, \chi_2^2 > 0 \) and zero otherwise. Use the convolution formula and Problem 2.2.22 to show that the marginal distribution of \( \chi^2 = \chi_1^2 + \chi_2^2 \) is chi square(\( m + n \)).

28 Consider the uniform distribution over the unit ball \( y_1^2 + y_2^2 + y_3^2 < 1 \) in \( R^3 \). Let \( y_1 = r \cos a \sin b, y_2 = r \cos a \cos b, y_3 = r \sin a \) define distance \( r \) from the origin in \( (0, \infty) \), latitude \( a \) in \( (-\pi/2, \pi/2) \), and longitude \( b \) in \( (0, 2\pi) \).

(a) Determine the density function for \( (r, a, b) \).

(b) Determine the marginal density of \( r \), of \( a \), and of \( b \).

I PROBLEMS: OTHER CONVOLUTION FORMULAS

29 Convolution: a difference: Consider the distribution \( f(x, y) \) for \( (x, y) \) on \( R^2 \), and derive the following expression for the marginal density of \( x - y \):

\[ f_{x-y}(t) = \int_{-\infty}^{\infty} f(t + y, y) \, dy. \]
30 Convolution: a ratio: Consider a distribution with density $f$ on $\mathbb{R}^2$ with $f(x, y) = 0$ for $y \leq 0$. Show that the density function for $v = x/y$ is given by

$$f_v(v) = \int_0^\infty f(vy, y) y \, dy.$$  

31 Convolution: a proportion: Consider a distribution with density $f$ on $\mathbb{R}^2$ with $f(x, y) = 0$ unless $x > 0$, $y > 0$. Show that the density function for $u = x/(x + y)$ is given by

$$f_u(u) = \int_0^\infty f(ut, (1 - u)t) t \, dt$$
on $(0, 1)$ and zero otherwise.

32 Convolution: linear combination: Let $w = ax + by$ with $a \neq 0$. Derive the following expression for the marginal density of $w$:

$$f_w(w) = |a|^{-1} \int_{-\infty}^\infty f \left( \frac{w - by}{a}, y \right) \, dy.$$  

33 (continuation of Problem 27) Use Problem 30 to show that $G = \chi^2_1/\chi^2_2$ is canonical $F(m, n)$; recall Problem 2.2.21. Thus $F = (\chi^2_1/m)/(\chi^2_2/n)$ is standard $F(m, n)$.

34 (continuation of Problem 27) Use Problem 31 to show that $B = \chi^2_1/(\chi^2_1 + \chi^2_2)$ is beta$(m/2, n/2)$; recall Problem 2.2.21.

35 Consider the distribution for $(z, \chi)$ with density $f$ on $\mathbb{R}^2$ given by

$$f(z, \chi) = (2\pi)^{-1/2} e^{-z^2/2} \cdot |z|^{-1} \left( \frac{n}{2} \right) e^{-\chi^2/2} \left( \frac{\chi^2}{2} \right)^{(n/2) - 1} X$$

for $\chi > 0$ and zero otherwise. Use Problem 30 to show that $T = z/\chi$ is canonical Student$(n)$; recall Problem 2.2.21.

### 3.3

STATISTICAL INDEPENDENCE

In science and industry it is common practice to take repeated observations on a random system operating under essentially constant conditions. At one extreme this can involve duplicating measurements or determinations on some physical quantity; for example, measurements on the lifetime of high-pressure sodium light bulbs made by a particular industrial process; or measurements of the weight increase in rabbits under a certain diet therapy. At another extreme this can involve repeating or replicating a basic design of response determinations coming from specified input conditions: for example, a basic design has three fertilizers applied respectively to three plots of land in a block and the corresponding yields observed; this design pattern is then replicated on several blocks of land.

Consider the process of taking repeated observations on a random system operated under constant conditions. The probability pattern for a performance will depend on the conditions. With constant conditions the pattern will be independent of results from antecedent performances—or otherwise the conditions would not, in fact, be constant.

This can be examined empirically by considering a pair of separate performances and then repeating the pair. If the performances in a pair are kept separate, then under repetitions the performance pattern for the pair is found to exhibit a symmetry or
independence: the performance pattern for one element of the pair appears symmetrically with respect to possibilities for the other element.

In this section we examine this symmetry or independence for a pair of performances under constant conditions. And more generally we examine this symmetry or independence for a pair of different systems operated independently.

Independence is an important and fundamental property. It is useful for the probability analysis of given probability models; and it allows us to build large probability models from simple models that describe separate components of a system.

A INDEPENDENT EVENTS

Consider the tossing of a symmetric coin; let \( H \) be the event "'heads for the coin'": \( P(H) = 1/2 \). Also consider the tossing of a symmetric die; let \( A \) be the event "'ace'": \( P(A) = 1/6 \). Now consider the tossing of the coin and die in a thorough way to ensure symmetry among the pairings of a face for the coin with a face for the die. The symmetry among the \( 2 \times 6 \) pairings then gives

\[
P(H)P(A) = (1/2) \cdot (1/6) = 1/12,
\]

which we note is equal to

\[
P(HA) = 1/12.
\]

A basic way to ensure symmetry is to have the coin tossing physically separate or independent of the tossing of the die. Accordingly, with

\[
P(HA) = P(H)P(A),
\]

we say that the event \( H \) is "'statistically independent'" of the event \( A \).

Now consider a sample space \( \mathcal{S} \) with probabilities given by \( P \). Then we define a generalized symmetry relating an event \( A \) to an event \( B \) as follows:

**DEFINITION 1**

\( A \) is statistically independent of \( B \) if

\[
(1) \quad P(AB) = P(A)P(B).
\]

As empirical support for the term "'statistically independent'" we note that, with events \( A \) and \( B \) coming from physically independent systems, then for a large number of performances of the combined system, the estimated probabilities approximate the factorization presented in the definition; this is of course an expression of the generalized symmetry just mentioned.

The statistical independence of \( A \) and \( B \) implies the independence of events simply related to \( A \) and \( B \).

**PROPOSITION 2**

*If \( A \) is statistically independent of \( B \), then each event in \( \{ \emptyset, A, A^c, \mathcal{S} \} \) is statistically independent of each event in \( \{ \emptyset, B, B^c, \mathcal{S} \} \).*
Proof  The proof is straightforward. Consider the pair $A$ and $\emptyset$:

$$P(A \emptyset) = P(\emptyset) = 0,$$

$$P(A \emptyset) = P(A) \cdot 0 = 0.$$  

And consider, for example, the pair $A$ and $B^c$:

$$P(AB^c) = P(A) - P(AB) = P(A) - P(A)P(B),$$

$$P(A \emptyset) = P(A)(1 - P(B)) = P(A) - P(A)P(B).$$

Consider an example where the events clearly belong to a single random system:

EXAMPLE 1

One card from a deck: A deck of 52 playing cards is thoroughly shuffled and one card is dealt face up. Let $S$ be the event "spade" and $A$ be the event "ace." Then

$$P(S) = 13/52 = 1/4, \quad P(A) = 4/52 = 1/13,$$

$$P(SA) = 1/52 = P(S)P(A).$$

Thus the events $S$ and $A$ are statistically independent. It then follows from Proposition 2 that, say, $S^c$ and $A$ are statistically independent.

$$P(S^cA) = 3/52 = (3/4) \cdot (1/13) = P(S^c)P(A).$$

B  INDEPENDENT ALGEBRAS

With independent events $A$ and $B$. Proposition 2 shows that each event in $\{ \emptyset, A, A^c, S \}$ is independent of each event in $\{ \emptyset, B, B^c, S \}$. We now discuss independence between classes of events. Consider a sample space $\mathcal{S}$ with $\sigma$-algebra $\mathcal{A}$ of events and probability measure $P$. Let $\mathcal{A}_1$ be a class of events, say referring to certain characteristics of a system being investigated, and $\mathcal{A}_2$ be another class, say referring to other characteristics of the system. There are some advantages to thinking of the classes as being closed under the standard operations $\cup$, $\cap$, $^c$; accordingly, we will assume that they are $\sigma$-algebras.

DEFINITION 3

A $\sigma$-algebra $\mathcal{A}_1$ is statistically independent of a $\sigma$-algebra $\mathcal{A}_2$ if

$$P(A_1A_2) = P(A_1)P(A_2)$$

for each $A_1$ in $\mathcal{A}_1$ and $A_2$ in $\mathcal{A}_2$.

EXAMPLE 2

One card from a deck: One card is dealt from a well-shuffled deck (Example 1). We have seen that $S$, "spade," and $A$, "ace," are statistically independent. By Proposition 2 it follows that $\{ \emptyset, S, S^c, S \}$ and $\{ \emptyset, A, A^c, S \}$ are statistically
independent. Indeed, it should be clear that any $\sigma$-algebra describing suits is independent of any $\sigma$-algebra describing denominations.

C INDEPENDENT FUNCTIONS

Now consider independence between two functions $X$ and $Y$ defined on the sample space $\mathcal{S}$. Of course, we assume that the functions are respectable, that is, measurable, as discussed in Section 3.1.

DEFINITION 4

The functions $X$ and $Y$ are statistically independent if each event for $X$ is statistically independent of each event for $Y$.

Note that the probability statement for the independence of an event $B_1$ for $X$ and an event $B_2$ for $Y$ can be expressed directly as

\[ P_{XY}(B_1 \times B_2) = P_X(B_1)P_Y(B_2), \]

or in terms of preimage events as

\[ P(X^{-1}(B_1) \cap Y^{-1}(B_2)) = P(X^{-1}(B_1))P(Y^{-1}(B_2)). \]

The independence of $X$ and $Y$ can be related to the independence of algebras in Definition 3: let $\mathcal{B}_1$ be the events for the function $X$ and $\mathcal{B}_2$ be the events for the function $Y$; then independence says that $X^{-1}(\mathcal{B}_1)$ is statistically independent of $Y^{-1}(\mathcal{B}_2)$.

EXAMPLE 3

One card from a deck: One card is dealt from a well-shuffled deck (Examples 1 and 2). Let $X$ be the indicator function for the event $S$ and $Y$ be the indicator for the event $A$. Then $X$ and $Y$ are statistically independent; this uses Proposition 2 with the results in Example 2. This simple example has been used repeatedly; some substantial examples are given in Sections E and F.

Note that when we talk of two functions $X$ and $Y$ being statistically independent, we are saying very little about $X$ and $Y$ as functions. Rather the property centers on the probability measure on the sample space and the functions $X$ and $Y$ merely delineate events on that sample space.

We now present a simple proposition that emphasizes an important aspect of independence, an aspect that is almost trivial as based on the definitions we have used:

PROPOSITION 5

If $X$ and $Y$ are statistically independent and if $h_1$ and $h_2$ are measurable functions on the ranges of $X$ and $Y$, respectively, then $h_1(X)$ and $h_2(Y)$ are statistically independent.
Proof: From Definition 4 we have that each event for \( X \) is statistically independent of each event for \( Y \). Now consider \( h_1(X) \) and \( h_2(Y) \). An event for \( h_1(X) \) is necessarily an event for the antecedent function \( X \); see Definition 3.1.2 for a measurable function. Similarly, an event for \( h_2(Y) \) is necessarily an event for \( Y \). Thus from the initial statement we have trivially that each event for \( h_1(X) \) is independent of each event for \( h_2(Y) \). Note that the functions \( h_1 \) and \( h_2 \) just cut down on the number of events being considered.

D INDEPENDENCE ON THE LINE AND PLANE

On real spaces we have seen the importance of distribution functions, probability functions, and density functions. Now consider independence in terms of these functions.

PROPOSITION 6

Real-valued functions \( X \) and \( Y \) are statistically independent if and only if

\[
F_{XY}(x, y) = F_X(x)F_Y(y)
\]

for all \( x \) and \( y \) on the real line \( R \).

Proof If \( X \) and \( Y \) are statistically independent, then the events \((\infty, x] \times R \) and \( R \times (\infty, y] \) are statistically independent on the space \( R^2 \) for \( (X, Y) \). This gives

\[
F_{XY}(x, y) = P_{XY}(\{\infty, x] \times (\infty, y]) = P_X(\{\infty, x]}P_Y(\{\infty, y])
\]

\[= F_X(x)F_Y(y).\]

The proof of the converse is recorded in Supplement Section 3.6; it is more complicated and uses the Extension Theorem, 1.5.5 in Supplement Section 1.5.

PROPOSITION 7

For a discrete distribution, \( X \) and \( Y \) are statistically independent if and only if

\[
p_{XY}(x, y) = p_X(x)p_Y(y)
\]

for all \( x \) and \( y \) on the real line.

Proof The proof follows immediately from Proposition 6. For consider

\[
F_{XY}(x, y) = \sum_{t \leq x, u \leq y} p_{XY}(t, u),
\]

\[
F_X(x)F_Y(y) = \sum_{t \leq x} p_X(t) \sum_{u \leq y} p_Y(u) = \sum_{t \leq x, u \leq y} p_X(t)p_Y(u).
\]

If \( X \) and \( Y \) are independent, then by Proposition 6 the left sides are equal; but a distribution function determines the probability function; it follows that
\[ p_{XY} = p_X p_Y. \] Conversely, if \( p_{XY} = p_X p_Y \), then the right sides are equal and thus the left sides are equal; independence follows from Proposition 6.

**PROPOSITION 8**

*For an absolutely continuous distribution, \( X \) and \( Y \) are statistically independent if and only if*

\[ f_{XY}(x, y) = f_X(x) f_Y(y) \]

*almost everywhere on the plane (except perhaps on a set that has area measure zero).*

*Proof* The proof parallels that for the preceding proposition but uses

\[
F_{XY}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(t, u) \, dt \, du,
\]

\[
F_X(x) F_Y(y) = \int_{-\infty}^{x} f_X(t) \, dt \int_{-\infty}^{y} f_Y(u) \, du = \int_{-\infty}^{x} \int_{-\infty}^{y} f_X(t) f_Y(u) \, dt \, du.
\]

Factorization so that the variables separate is the essential part of the condition in each of these propositions. More flexible versions of these propositions are recorded as Exercises 1, 2, and 3.

**E COMBINING INDEPENDENT SYSTEMS**

Now consider the use of independence to build probability models. Suppose that we have two random systems that are physically independent of each other. For the first system with response \( s_1 \), let \( \mathbb{S}_1 \) be the sample space, \( \mathcal{A}_1 \) be the class of events, and \( P_1 \) be the probability measure. And for the second system with response \( s_2 \), let \( \mathbb{S}_2 \) be the sample space, \( \mathcal{A}_2 \) be the class of events, and \( P_2 \) be the probability measure.

Now suppose that we are interested in a performance of the combined system. The combined response is \( (s_1, s_2) \); thus the sample space \( \mathbb{S} \) is the product \( \mathbb{S}_1 \times \mathbb{S}_2 \). We would certainly be interested in events given by a product set \( A_1 \times A_2 \). Accordingly, we take \( \mathcal{A} \) to be the smallest \( \sigma \)-algebra that includes the product sets \( A_1 \times A_2 \).

And if the systems are independent, we would have \( P(A_1 \times A_2) = P_1(A_1) P_2(A_2) \) on the basis of events for the first system being statistically independent of events for the second system. The measure for the product sets then extends uniquely to a probability measure for the class \( \mathcal{A} \) of events; this uses the Extension Theorem, 1.5.5, with details as in Section 2.3B.

Thus we obtain the product space \( \mathbb{S} \), the product algebra \( \mathcal{A} \) of events, and the product probability measure \( P \). In this way we can build progressively larger models by introducing components one by one. We can thus produce quite complicated models from quite simple ingredients.

**EXAMPLE 4**

An engineer measures the lifetime \( y_1 \) of an electronic component; it is known to be exponential(\( \theta \)). He also measures the lifetime \( y_2 \) for an independent electronic
component; it is known to be exponential($2\theta$). Construct the model for the combined system with outcome $(y_1, y_2)$. The density function for $(y_1, y_2)$ is available from Proposition 8:

$$f(y_1, y_2) = \frac{1}{\theta} e^{-\frac{y_1}{\theta}} \frac{1}{2\theta} e^{-\frac{y_2}{2\theta}}$$

$$= \frac{1}{2\theta^2} e^{-\frac{(y_1+y_2)}{2\theta}} \quad \text{if } y_1, y_2 > 0$$

and zero otherwise. See Exercise 3.2.16.

---

**F SAMPLES FROM A DISTRIBUTION**

Consider a random system with response $y$. Sometimes we may be interested in a single performance of the system. Other times we may be interested in a sequence of independent performances yielding a combined response $(y_1, \ldots, y_n)$. If we have a model for a single performance, we can build a model for the combined performance by using the given model for each component and then combining by independence as in Section E.

**DEFINITION 9**

If $y_1, \ldots, y_n$ are statistically independent and if each $y_i$ has the same distribution with measure $P$, or with probability function $p$, or with density function $f$, then $y = (y_1, \ldots, y_n)'$ is a sample from the distribution given by $P$, or by $p$, or by $f$.

Alternatively, we can say that $y_1, \ldots, y_n$ are IID (independent and identically distributed) according to $P$, $p$, or $f$.

In the discrete case the joint probability function for the sample $y$ on $\mathbb{R}^n$ is

$$\rho(y) = \rho(y_1) \cdots \rho(y_n);$$

and in the absolutely continuous case, the joint probability density for the sample $y$ on $\mathbb{R}^n$ is

$$f(y) = f(y_1) \cdots f(y_n).$$

**EXAMPLE 5**

**Sample from Bernoulli($p$):** A certain drug therapy has a probability $p$ of success; let $x$ be the indicator for success. The probability function is

$$\rho(x) = p^x q^{1-x} \quad \text{if } x = 0, 1$$

and zero otherwise ($q = 1 - p$). This is the Bernoulli($p$) distribution. Now consider $n$ repetitions on different animals obtained randomly from a large population; let $(x_1, \ldots, x_n)$ be the vector of indicators. The probability function is obtained by Proposition 7:

$$\rho(x_1, \ldots, x_n) = p^{x_1} q^{1-x_1} \cdots p^{x_n} q^{1-x_n}$$

$$= \rho_{x_1, \ldots, x_n}$$

on the points $(0, 1)^n$ in $\mathbb{R}^n$ and zero elsewhere.
EXAMPLE 6

Sample from Poisson($\lambda$): Consider a count $y$ from a radioactive source having a Poisson distribution

$$p(y) = \frac{\lambda^y e^{-\lambda}}{y!} \quad \text{if } y = 0, 1, 2, \ldots$$

and zero otherwise. And suppose that we wish to consider $n$ independent performances of the system. Let $y_1, \ldots, y_n$ designate the responses for 1st, $\ldots$, $n$th performances. The independence among the various performances then prescribes statistical independence, and by Proposition 7 we obtain

$$p(y_1, \ldots, y_n) = p(y_1) \cdots p(y_n)$$

$$= \frac{\lambda^{y_1} e^{-\lambda}}{y_1!} \cdots \frac{\lambda^{y_n} e^{-\lambda}}{y_n!}$$

$$= \frac{\lambda^{y_1+y_n} e^{-n\lambda}}{\Pi y_i!} \quad y_i = 0, 1, 2, \ldots$$

and zero otherwise. We thus obtain a discrete distribution on $\mathbb{R}^n$, and it describes the combined response $y = (y_1, \ldots, y_n)'$.

In an application we may not know the true value of $\lambda$ that gives the correct Poisson distribution for the response. We may then make $n$ performances of the system and obtain an observed sample $(y_1, \ldots, y_n)$. In order to determine plausible values for $\lambda$, we will, in the statistics chapters, see how the observed sample relates to the distribution $p(\cdot, \ldots, \cdot)$ prescribed by each of the possible values for $\lambda$.

EXAMPLE 7

Sample from the normal($\mu$, $\sigma$): Consider a voltage measurement $y$ having a normal distribution($\mu$, $\sigma$):

$$f(y) = \frac{1}{\sqrt{2\pi} \sigma} \exp \left\{ -\frac{1}{2\sigma^2} (y - \mu)^2 \right\}.$$

Now suppose that we wish to consider $n$ independent performances of the system. Let $y_1, \ldots, y_n$ designate the responses for the 1st, $\ldots$, $n$th performances. Then $(y_1, \ldots, y_n)$ is a sample from the normal($\mu$, $\sigma$) distribution; the probability density function is available from Proposition 8:

$$f(y_1, \ldots, y_n) = f(y_1) \cdots f(y_n)$$

$$= (2\pi)^{-n/2} \sigma^{-n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\}.$$

Thus $y = (y_1, \ldots, y_n)'$ has a rotationally symmetric normal distribution in $\mathbb{R}^n$; it is located at some point $\mu \mathbf{1} = \mu (1, \ldots, 1)' = (\mu, \ldots, \mu)'$ on the line $\mathbf{L}(1)$.

In an application we may not know the values for $\mu$ and $\sigma$. We may then make $n$ performances of the system and obtain an observed sample $(y_1, \ldots, y_n)$. In order to determine plausible values for $(\mu, \sigma)$, we will, in the statistics chapters,
see how the observed sample relates to the distribution \(f(\cdot, \ldots, \cdot)\) prescribed by each of the possible values for \((\mu, \sigma)\).

\section*{G EXERCISES: CRITERIA FOR INDEPENDENCE}

1. Prove: \(X\) and \(Y\) are statistically independent if and only if \(F_{XY}(x, y) = g(x)h(y)\) factors so that the variables separate. Use Proposition 6.

2. Prove for the discrete case: \(X\) and \(Y\) are statistically independent if and only if \(p_{XY}(x, y) = g(x)h(y)\) factors so that the variables separate. Use Proposition 7.

3. Prove for the absolutely continuous case: \(X\) and \(Y\) are statistically independent if and only if \(f_{XY}(x, y) = g(x)h(y)\) factors so that the variables separate, except perhaps on a set that has area measure zero. Use Proposition 8.

\section*{H EXERCISES}

4. Consider the standard normal distribution for \((z_1, z_2)\) on \(\mathbb{R}^2\); see Example 3.2.9. Show that \(r\) and \(a\) are statistically independent, where \(z_1 = r \cos a\) and \(z_2 = r \sin a\).

5. \textit{Sum of chi squares}: Let \(\chi_1^2, \ldots, \chi_k^2\) be statistically independent and \(\chi_i^2\) have a chi-square \((m_i)\) distribution. Problem 3.2.27 shows that \(\chi_1^2 + \ldots + \chi_k^2\) is chi square \((m_1 + \ldots + m_k)\). Deduce by induction that \(\chi^2 = \chi_1^2 + \ldots + \chi_k^2\) is chi square \((m_1 + \ldots + m_k)\). This is called the \textit{reproductive property} of the chi-square distribution.

6. Let \((z_1, \ldots, z_k)\) be a sample from the standard normal. Show that \(\sum z_i^2\) is chi square \((k)\); recall Example 3.2.2 and Exercise 5.

\section*{I PROBLEMS}

7. Three symmetric dice are tossed and \(s = (x_1, x_2, x_3)\) is observed, where \(x_i\) is the number of points on the \(i\)th die. Determine the probability function for \(W\), where \(W(x_1, x_2, x_3) = x_1 + x_2 + x_3\); recall Examples 3.2.4 and 3.2.5.

8. Let \(\chi_1^2\) and \(\chi_2^2\) be statistically independent and \(\chi_i^2\) have a chi-square \((m_i)\) distribution. Show that \(T = \chi_1^2 + \chi_2^2\) and \(G = \chi_1^2 / \chi_2^2\) are statistically independent, and deduce that \(T\) is chi square \((m_1 + m_2)\) and \(G\) is canonical \(F(m_1, m_2)\).

9. Let \(\chi_1^2\) and \(\chi_2^2\) be statistically independent and \(\chi_i^2\) have a chi-square \((m_i)\) distribution. Show that \(T = \chi_1^2 + \chi_3^2\) and \(P = \chi_1^2 / (\chi_1^2 + \chi_3^2)\) are statistically independent, and deduce that \(T\) is chi square \((m_1 + m_3)\) and \(P\) is beta \((m_1/2, m_3/2)\).

10. Let \((z_1, z_2)\) be a sample of 2 from the standard normal. Use Problem 3.2.32 to show that \(W = z_1 + b z_2\) is normal \((0, 1 + b^2)^{1/2}\).

\textit{(continuation)} Let \(y_1\) and \(y_2\) be statistically independent and \(y_i\) normal \((\mu_i, \sigma_i^2)\). Represent \(y_1\) and \(y_2\) in terms of standard normal \(z_1\) and \(z_2\) and show that \(y_1 + y_2\) is normal \((\mu_1 + \mu_2, (\sigma_1^2 + \sigma_2^2)^{1/2})\).

\section*{J PROBLEMS: COMPUTER USAGE}

Some computer exercises are recorded in Sections 2.4H and 2.4J. Use the results there for the following theoretical and applied problems.

12. Let \(u_1\) and \(u_2\) be statistically independent and \(u_i\) be uniform on \((0, 1)\), and define \((z_1, z_2) = (-2 \ln u_1)^{1/2} \cos 2\pi u_2, (-2 \ln u_1)^{1/2} \sin 2\pi u_2\). Show that \((z_1, z_2)\) is a sample of 2 from the standard normal; recall Example 3.2.9.

\textit{(continuation)} Use a computer program for random digits to obtain a sample of 50 from the standard normal; and see where the values fall in relation to the density function.
3.4
COUNTING SAMPLE POINTS

Many qualitative scientific investigations are concerned with the frequency of occurrence of an event in a sequence of repetitions of a random system. In Section 3.5 we shall determine the marginal distributions appropriate to such frequencies. Typically, the underlying distribution for a sequence of repetitions has a variety of symmetries, symmetries that lead to the equality of some or all of the underlying probabilities. Accordingly, the derivation of various marginal probabilities can simplify to the counting of sample points. For example, with a symmetric distribution on a set \( S_0 \), we have

\[
P(A) = \frac{c(A \cap S_0)}{c(S_0)},
\]

where \( c \) is the cardinality measure. In this section we develop formulas for counting sample points in a set—for determining the cardinality of a set.

A COUNTING SEQUENCES: INDIVIDUAL POPULATIONS

Suppose that we have several finite sets, or populations, \( \mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_k \), and let \( c(\mathcal{S}_i) = N_i \) be the number of elements in the set \( \mathcal{S}_i \). Consider sequences of the form

\[
(x_1, \ldots, x_k), \quad x_i \in \mathcal{S}_i.
\]

And then consider the sample space \( S_1 \) consisting of such sequences:

\[
S_1 = \{(x_1, \ldots, x_k) : x_i \in \mathcal{S}_i\};
\]

this is the Cartesian product of the spaces \( \mathcal{S}_1, \ldots, \mathcal{S}_k \).

The cardinality of the space \( S_1 \) is

\[
(1) \quad c(S_1) = c(\mathcal{S}_1) \cdots c(\mathcal{S}_k) = N_1 \cdots N_k.
\]

For the first \( x \) can be any of the \( N_1 \) elements in \( \mathcal{S}_1 \); and given the first, the second can be any of the \( N_2 \) elements in \( \mathcal{S}_2 \); and given the first two, the third can be any of the \( N_3 \) elements in \( \mathcal{S}_3 \), and so on; thus the total number is \( N_1 \cdots N_k \).

EXAMPLE 1

\( n \) tosses of a coin: A symmetric coin is tossed \( n \) times. Let \( x_i \) be the indicator for heads on the \( i \)th toss; the space for \( x_i \) is

\[
\mathcal{S}_i = \{0, 1\}; \quad c(\mathcal{S}_i) = 2.
\]

Then for the sequence of \( n \) tosses we have

\[
(x_1, \ldots, x_n), \quad x_i \in \mathcal{S}_i,
\]

giving the sample space

\[
S_1 = \{(x_1, \ldots, x_n) : x_i \in \mathcal{S}_i\}.
\]
The cardinality of $\mathcal{S}_1$ is $c(\mathcal{S}_1) = 2^n$. We then have the probabilities
\[ p(0, \ldots, 0) = 1/2^n, \]
\[ p(1, \ldots, 1) = 1/2^n, \]
\[ P(r \text{ 1s and } n-r \text{ 0s in a specified order}) = 1/2^n. \]

**B COUNTING SEQUENCES: A SINGLE POPULATION**

Suppose that we have a finite set or population $\mathcal{S}$, and let $c(\mathcal{S}) = N$ be the number of elements in the set $\mathcal{S}$. Consider sequences of the form
\[ (x_1, \ldots, x_n), \quad x_i \in \mathcal{S}, x_i \neq x_j \quad (i \neq j), \]
where if a particular element of the population is in a position $x_i$, then it is unavailable for the remaining positions in the sequence. And then consider the sample space $\mathcal{S}_2$ consisting of such sequences:
\[ \mathcal{S}_2 = \{(x_1, \ldots, x_n): x_i \in \mathcal{S}, x_i \neq x_j \quad (i \neq j)\}. \]

*The cardinality of the space $\mathcal{S}_2$ is*
\[ c(\mathcal{S}_2) = N^{(n)} = N(N - 1) \cdots (N - n + 1) = \frac{N!}{(N-n)!}. \]

For the first $x$ can be any of the $N$ elements in $\mathcal{S}$; and given the first, the second can be any of the $N - 1$ remaining elements, and so on for the $n$ positions in the sequence; thus the total number is $N(N - 1) \cdots (N - n + 1)$. Note that $N^{(n)}$ is called $N$ descending factorial $n$; it has $n$ factors, commencing with $N$.

**EXAMPLE 2**

**Thirteen cards from a deck:** A deck of 52 playing cards is thoroughly shuffled and 13 cards are dealt in succession face up. Let $x_i$ designate the $i$th card dealt. Then for the sequence of 13 cards we have
\[ (x_1, \ldots, x_{13}), \quad x_i \in \mathcal{S}, x_i \neq x_j \quad (i \neq j), \]
giving the sample space
\[ \mathcal{S}_2 = \{(x_1, \ldots, x_{13}): x_i \in \mathcal{S}, x_i \neq x_j \quad (i \neq j)\}. \]

The cardinality of $\mathcal{S}_2$ is $52^{(13)} = 52 \cdot 51 \cdot \ldots \cdot 40$. We then have
\[ p(x_1, \ldots, x_{13}) = 1/52^{(13)} \]
for each possible sequence $(x_1, \ldots, x_{13})$ of 13 cards from 52. Now suppose that we are interested in the probability of 13 spades. The number of sequences of 13 spades is $13^{(13)} = 13!$. Thus
\[ P(13 \text{ spades}) = 13!/52^{(13)}. \]
COUNTING SETS

Suppose that we have a finite set or population $S$ and let $c(S) = N$ be the number of elements in the set $S$. Consider subsets of $S$ that contain $n$ elements:

$$\{x_1, \ldots, x_n\}, \quad x_i \in S, x_i \neq x_j \quad (i \neq j).$$

And then consider the sample space $S_3$ consisting of such subsets:

$$S_3 = \{ (x_1, \ldots, x_n) : x_i \in S, x_i \neq x_j \quad (i \neq j) \}.$$

The cardinality of the space $S_3$ is

$$c(S_3) = \binom{N}{n} = \frac{N(n)}{n!} = \frac{N}{n!(N-n)!}. \tag{3}$$

For we can form a mapping $Y$ from the preceding space $S_2$ onto the present space $S_3$ such that

$$(x_1, \ldots, x_n) \mapsto \{x_1, \ldots, x_n\};$$

that is, the sequence $(x_1, \ldots, x_n)$ is carried into the set $(x_1, \ldots, x_n)$. This is an $n!$ to one mapping; thus

$$c(S_3) = \frac{c(S_2)}{n!} = \frac{N(n)}{n!} = \binom{N}{n}.$$

Note that $\binom{N}{n}$ is called $N$ choose $n$.

EXAMPLE 3

Thirteen cards from a deck; see Example 2: Thirteen cards are dealt from a well-shuffled deck $S$ and are collected together as a hand, or set of 13 cards, without regard for the order in which they were dealt. The number of different hands is

$$c(S_3) = \binom{52}{13} = \frac{52^{(13)}}{13!} = 635,013,559,600.$$ 

We then have (recall the mapping $Y$)

$$p_Y(\{x_1, \ldots, x_{13}\}) = \frac{1}{\binom{52}{13}}$$

for possible hands $\{x_1, \ldots, x_{13}\}$ of 13 cards. The probability for 13 spades is

$$P_Y(13 \text{ spades}) = \frac{1}{\binom{52}{13}} = \frac{13!}{52^{(13)}}.$$

which agrees with the value calculated earlier on the basis of sequences of cards. The probability of a hand containing 5 spades, 3 hearts, 2 diamonds, and 3 clubs can now be calculated using, in addition, the results from Section A:

$$P(5 \text{ spades}, 3 \text{ hearts}, 2 \text{ diamonds}, 3 \text{ clubs})$$
\[
\left(\binom{13}{5} \cdot \binom{13}{3} \cdot \binom{13}{2} \cdot \binom{13}{3}\right) \cdot \binom{52}{13}.
\]

Subsets from a space \(S\) can be described in another way by placing an indicator function on the population \(S\)—to indicate whether an element is in the sample or not. Let \(w_j\) be equal to \(b\) if the \(j\)th population element is in the sample and equal to \(a\) if it is not. Then the sample space \(S_3\) can be expressed alternatively as

\[
S_3 = \{ (w_1, \ldots, w_n) : n \text{ of the } w's \text{ are equal to } b; N - n \text{ of the } w's \text{ are equal to } a \}.
\]

The cardinality of \(S_3\) is of course the number of ways in which we can choose \(n\) places for the \(b's\).

**D COUNTING SEQUENCES OF SETS**

Consider a finite set or population \(S\) and let \(c(S) = N\) be the number of elements in the set \(S\). And suppose that we form a first subset \(\{x_1, \ldots, x_{n_1}\}\) that contains \(n_1\) elements, a second subset \(\{x_{n_1+1}, \ldots, x_{n_1+n_2}\}\) that contains \(n_2\) elements, \ldots, and a last subset \(\{x_{n_1+\ldots+n_k+1}, \ldots, x_N\}\) that contains \(n_k\) elements (\(\Sigma n_j = N\)); and then consider

\[
\{(x_1, \ldots, x_{n_1}), \ldots, (x_{N-n_k+1}, \ldots, x_N)\}, \quad x_i \in S, x_i \neq x_j.
\]

The sample space \(S_4\) is the set of such elements.

*The cardinality of the space \(S_4\) is*

\[
(4) \quad c(S_4) = \binom{N}{n_1 \ldots n_k} = \frac{N!}{n_1! \ldots n_k!} = \binom{N}{n_1} \binom{N-n_1}{n_2} \ldots \binom{N-n_1-\ldots-n_{k-1}}{n_k}.
\]

For consider the space \(S_2\) with \(n = N\); let \(W\) be the mapping from the space \(S_2\) to the present space \(S_4\) such that

\[
(x_1, \ldots, x_N) \mapsto \{(x_1, \ldots, x_{n_1}), \ldots, (x_{N-n_k+1}, \ldots, x_N)\};
\]

that is, the sequence \((x_1, \ldots, x_N)\) is carried into the sequence of sets of size \(n_1, \ldots, size n_k\). This is an \(n_1! \ldots n_k!\) to one mapping; thus

\[
c(S_4) = \frac{c(S_2)}{n_1! \ldots n_k!} = \frac{N!}{n_1! \ldots n_k!}.
\]

Alternatively, we can form the first set in \(\binom{N}{n_1}\) different ways, then the second set in \(\binom{N-n_1}{n_2}\) different ways, and so on, giving the right-hand expression (4). Note that

\[
\binom{N}{n_1 \ldots n_k} \text{ is called } N \text{ choose } n_1, \ldots, n_k.
\]

**EXAMPLE 4**

A deck \(S\) of 52 playing cards is thoroughly shuffled and 13 cards are dealt to E, 13 to S, 13 to W, and 13 to N. We thus obtain a sequence of four hands each of
13 cards. The number of such bridge distributions is

\[ c(S_4) = \frac{52!}{13! \ 13! \ 13! \ 13!} \]

The probability for any particular bridge distribution is

\[ \rho_w \{(x_1, \ldots, x_{13}), \ldots, (x_{40}, \ldots, x_{52})\} = \frac{(13!)^4}{52!} \]

A sequence of subsets from a space \( S \) can be described in another way by placing an indicator function on the population \( S \) — to indicate for any element which of the \( k \) samples it goes in. Let \( w_j = b_1, \ldots, b_k \) according as the \( j \)th element of \( S \) is in the first, \( \ldots, k \)th subset, respectively. Then the sample space \( S_4' \) can be expressed alternatively as

\[ S_4' = \left\{ w_1, \ldots, w_n : \begin{array}{l}
n_1 \text{ of the } y's \text{ are equal to } b_1 \\
 \vdots \\
n_k \text{ of the } y's \text{ are equal to } b_k \end{array} \right\} \]

The cardinality of \( S_4' \) is, of course, the number of ways we can choose \( n_1 \) places for \( b_1 \)'s, then \( n_2 \) places for \( b_2 \)'s, and so on.

**E Extended Factorials**

The factorial expressions in this section can occur in various summation formulas, and it is convenient to extend their definition and thus avoid clumsy restrictions on the summations.

The **factorial function** \( n! = n(n-1) \cdots 2 \cdot 1 \) is defined initially for positive integers \( n \). This is extended to 0 by defining \( 0! = 1 \).

The **descending factorial** \( N^{(n)} = N(n-1) \cdots (N-n+1) \) is defined for all real \( N \) and initially for positive integers \( n \). This is extended to \( n = 0 \) by defining \( N^{(0)} = 1 \).

The **combinatorial function** is then defined by the formula

\[ \binom{N}{n} = \frac{N^{(n)}}{n!} \]

for all real \( N \) and for integers \( n = 0, 1, 2, \ldots ; \) and is extended by

\[ \binom{N}{0} = 1, \quad \binom{N}{n} = 0, \quad n = -1, -2, \ldots. \]

Then note that for integer \( N, n \)

\[ \binom{N}{n} = \binom{N}{N-n} \]

The **binomial theorem** can then be expressed in the form

\[ (1 + t)^N = 1 + \binom{N}{1} t + \binom{N}{2} t^2 + \cdots \]

for all real \( N \) provided that \( |t| < 1 \) and for positive integer \( N \) without restriction on \( t \). Note that for positive integers \( N \) the series terminates; otherwise, it is a power series.

The **generalized combinatorial function**
Sec. 3.4: Counting Sample Points  133

(7) \[
\binom{N}{n_1 \ldots n_k} = \frac{N!}{n_1! \ldots n_k!}
\]
is defined initially for positive integers \(n_1, \ldots, n_k\) such that \(\Sigma n_i = N\). The definition of \(0!\) gives an immediate extension to nonnegative integers \(n_1, \ldots, n_k\) such that \(\Sigma n_i = N\). We extend it to arbitrary integers \(n_1, \ldots, n_k\) and positive integer \(N\) by defining

\[
\binom{N}{n_1 \ldots n_k} = 0
\]

if any \(n_i\) is negative or if \(\Sigma n_i \neq N\).

The multinomial theorem can be expressed in the form

(8) \[(t_1 + \cdots + t_j)^N = \sum \binom{n}{n_1 \ldots n_j} t_1^{n_1} \cdots t_j^{n_j}\]

for positive integer \(N\). The summation effectively is over all sequences \((n_1, \ldots, n_j)\) with integers \(n_i \geq 0\) and \(\Sigma n_i = N\).

The values of the factorial function can be approximated for large \(n\) by Stirling's formula,

\[
\Gamma(N+1) = n! \approx \sqrt{2\pi n} n^{n+1/2} e^{-n} = \sqrt{2\pi n} \left(\frac{n}{e}\right)^n
\]

**F** **COMBINATORIAL EXERCISES**

1 Prove algebraically the following recursion relation for the combinatorial symbol (real \(n\), integer \(r\)):

\[
\binom{n}{r-1} + \binom{n}{r} = \binom{n+1}{r}
\]

Repeated use of the relation generates the Pascal triangle recording values of \(\binom{n}{r}\).

<table>
<thead>
<tr>
<th>(n)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2 Prove the preceding recursion relation in terms of the number of different sets from a population (positive integer \(n\), \(0 \leq \text{integer } r \leq n\)).

3 Prove the following relations (positive integer \(r\)):

\[
\binom{-1}{r} = (-1)^r, \quad \binom{-2}{r} = (-1)^r(r+1).
\]
For a nonnegative integer \( n \) prove that
\[
\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n} = 2^n.
\]
\[
\binom{n}{0} - \binom{n}{1} + \cdots + (-1)^n \binom{n}{n} = 0.
\]

For a positive integer \( n \) prove that
\[
\binom{n}{1} + 2 \binom{n}{2} + \cdots + n \binom{n}{n} = n 2^{n-1},
\]
\[
\binom{n}{1} - 2 \binom{n}{2} + \cdots (-1)^{n-1} n \binom{n}{n} = 0.
\]

**COMBINATORIAL PROBLEMS**

For positive integers \( n > r \), use Exercise 1 to prove that
\[
\binom{n}{r} = \binom{n-1}{r} + \binom{n-2}{r-1} + \cdots + \binom{n-r-1}{0}
\]
\[
= \binom{n-1}{r-1} + \binom{n-2}{r-1} + \cdots + \binom{r-1}{r-1}.
\]

Examine the expansion of \((1 + \theta)^n(1 + \theta)^m\) and prove that
\[
\binom{m}{0}\binom{n}{r} + \binom{m}{1}\binom{n}{r-1} + \cdots + \binom{m}{r}\binom{n}{0} = \sum_{t} \binom{m}{t}\binom{n}{r-t} = \binom{m+n}{r}
\]
for real \( m, n \) and integer \( r \).

For real \( n \) and integer \( r \), prove that
\[
\binom{-n}{r} = (-1)^r \binom{n+r-1}{r}, \quad (\frac{2^r}{r})^{2^{-2r}} = (-1)^r \binom{-1/2}{r}
\]

Use Problems 7 and 8 to prove that
\[
\binom{n}{r} = \binom{n}{r-1} + \cdots + \binom{n}{1} \neq 1.
\]
\[
\binom{n-k}{n-r} = \sum_{r} (-1)^r \binom{k}{r} \binom{n-r}{x}
\]
for integers \( n, k, \) and \( r \).

Use Problems 7 and 8 to prove that
\[
\binom{m+n+r-1}{r} = \sum_{x=0}^{r} \binom{m+r-x-1}{r-x} \binom{n+x-1}{x}
\]
for real \( m, n \) and integer \( r \).

Use Stirling's formula to show that
\[
\lim_{n \to \infty} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n/2} \Gamma\left(\frac{n}{2}\right)} = 1.
\]
H  PROBABILITY PROBLEMS

Some of the familiar probability problems that go with the combinatorial formulæ are
rephrased as distribution problems and presented in the appropriate context at the end of
Section 5.

12  A symmetrical die is rolled 5 times. Find the probability of 5 aces; of 5 even numbers; and of
5 nonaces.
13  Random digits are defined in Problem 2.1.8, the uniform distribution on $S_0 = \{0, 1, \ldots, 9\}$. An $n$-digit random number is a sample from such a uniform distribution. Find the
probability that a 5-digit number contains 5 even digits; and exactly 2 even digits.
14  Under conventional assumptions find the probability that a family with 5 children has 5 girls;
and exactly 2 girls.
15  A hand of poker contains 5 cards from a well-shuffled deck of 52 playing cards. Find the
probability that the cards are all spades.
16  Find the probability that the cards in a 5-card poker hand are all of different denominations.
17  Find the probability that a 5-card poker hand is a
   (a) Full house (3 cards of one denomination and 2 of another denomination).
   (b) Four of a kind (4 cards of one denomination and 1 other card).
10  Find the probability that a 5-card poker hand is a
   (a) Royal flush (10, J, Q, K, A in a single suit).
   (b) Straight flush (5 cards in sequence in a single suit, other than a royal flush).
   (c) Flush (5 cards in a single suit, other than a straight or royal flush).

I  PROBLEMS FROM PHYSICS

19  Suppose that $n$ identical balls are distributed among $k$ different cells. Let $y_i$ be the number of
balls in the $i$th cell. Show that the number of different arrangements $(y_1, \ldots, y_k)$ is

$$A_{n,k} = \binom{n + k - 1}{n} = \binom{n + k - 1}{k - 1}.$$

Note: A die with $k$ faces is tossed $n$ times and $(y_1, \ldots, y_k)$ is recorded where $y_i$ is the
frequency for the $i$th face; $A_{n,k}$ is the number of different vectors $(y_1, \ldots, y_k)$.

20  Consider $n$ different balls and $k$ different cells. Suppose that the balls are independently
tossed in the cells and the probabilities are equal for the different cells: the symmetrical
probability measure on $k^n$ points. Calculate the probability function $p(y_1, \ldots, y_k)$, where $y_i$
is the number of balls in the $i$th cell. Note: If each of $n$ particles can be in any of $k$ different
states, the preceding Maxwell–Boltzmann probabilities have been proposed as a description
of arrangements.

21  (continuation) Suppose that $n = k$ balls are distributed randomly among $k$ different cells.
   (a) Find the probability that all cells are occupied.
   (b) Find the probability that exactly one cell is unoccupied.

22  (continuation) Physical particles do not seem to behave as balls symmetrically into cells.
The Bose–Einstein model is a symmetrical model based on different arrangements $(y_1, \ldots, y_k)$
as sample points: thus probability $1/A_{n,k}$ for each arrangement. Consider 6 particles and
7 states. Find the probability that some state has 2 particles. some other 4 states have 1
particle each, and the remaining states have none—without specifying which states.

23  (continuation) The Fermi–Dirac model is a symmetric model based on arrangements—but
only on those arrangements with at most one particle in any state. Show that the probability
for a permissible arrangement is $1/\binom{k}{n}$. Some particles seem to follow Bose–Einstein and
others seem to follow Fermi–Dirac.
24 (continuation) Suppose that \( n = k \) particles are distributed among \( k \) states:
(a) Find the probability that all states are occupied (Bose–Einstein).
(b) Find the probability that all states are occupied (Fermi–Dirac).

J PROBLEM FROM ANALYSIS

25 Verify the following steps in a proof of Stirling's formula.

(a) \( \Gamma(n + 1) = n^{n+1/2}e^{-n} \int_{-\sqrt{n}}^{\infty} \exp \{ n \log (1 + v/\sqrt{n}) - \sqrt{n} v \} \, dv \).

(b) \( \lim_{n \to \infty} \frac{n^{n+1/2}e^{-n}}{\Gamma(n + 1)} \int_{-\sqrt{n}}^{\infty} \exp \{ n \log (1 + v/\sqrt{n}) - \sqrt{n} v \} \, dv = 1 \).

(c) \( \lim_{n \to \infty} \exp \{ n \log (1 + v/\sqrt{n}) - \sqrt{n} v \} = \exp \{-v^2/2\} \text{ uniformly on } [-\sqrt{n}, \sqrt{n}] \).

(d) \( \lim_{n \to \infty} \frac{n^{n+1/2}e^{-n}}{\Gamma(n + 1)} (2\pi)^{1/2} = 1 \).

3.5 MARGINAL DISTRIBUTIONS FOR FREQUENCIES

We are now in a position to derive various marginal distributions describing frequencies. In one direction we shall be concerned with frequencies such as the number of reactions when a new drug is administered to a sequence of animals from a certain genetic strain; or the number of sterile water samples when a sequence of samples is taken from a water source; or the number of aces with a sequence of tosses of a die; or the number of heads with a sequence of tosses of a coin. In these examples the frequency is the number of occurrences of an event in a sequence of independent repetitions of a random system and the event can occur or not occur on each repetition. The marginal distribution for such frequencies is examined in Sections A and B: binomial distribution and multinomial distribution.

In another direction we shall be concerned with frequencies such as the number of tagged fish in a sample caught from the fish in a particular lake (care needed to ensure equal probability for each possible sample), the number of people favoring a certain proposal in a sample of adult voters in a certain city (care needed to ensure equal probability for each possible sample), the number of defective tires in a sample of 50 from an incoming shipment of 1000 (thorough mixing or random numbers needed to ensure equal probability for each possible sample), or the number of spades in a hand of 13 from a well-shuffled deck of 52 playing cards. In these examples there is a finite population of people, tires, or cards, and random sampling is used to ensure that each possible sample has the same probability of being drawn, and the frequency is the number of sample elements with a particular property. The marginal distribution for such frequencies in a random sample from a finite population is examined in Sections C and D: hypergeometric distribution and multivariate hypergeometric distribution. The marginal distribution for a general frequency is examined in Section E. The derivations use the marginal concepts of Sections 3.1 and 3.2, the independence of Section 3.3, and the counting formulas of Section 3.4.
A BINOMIAL($n$, $p$) DISTRIBUTION

Consider the Bernoulli($p$) distribution for an indicator $x$:

$$p(1) = p, \quad p(0) = q,$$

or, equivalently,

$$p(x) = p^x q^{1-x}, \quad x = 0, 1,$$

and zero otherwise. Note that the letter $p$ is being used for the probability at the point $x = 1$ and quite separately is being used for the probability function; this slight notational overlap allows consistencies to be maintained elsewhere.

The indicator $x$ could register the occurrence of a reaction when a drug is administered to an animal from a certain genetic strain, or could register sterility of a water sample from a certain source, or could register the occurrence of an ace with the toss of a die, or could register the occurrence of a head with the toss of a coin.

Now consider $n$ independent performances of the system being investigated. More formally, let $(x_1, \ldots, x_n)$ be a sample from the Bernoulli($p$) distribution. Then by the independence example (Example 3.5),

$$p(x_1, \ldots, x_n) = p^{x_1} q^{1-x_1} \cdots p^{x_n} q^{1-x_n}, \quad x_j = 0, 1,$$

$$= p^y q^{n-y}, \quad x_j = 0, 1,$$

and zero otherwise, where we write $y = \sum x_j$ for the frequency of 1's in the sample $(x_1, \ldots, x_n)$.

Now consider the marginal distribution of $y = x_1 + \cdots + x_n$, the frequency of 1's in the sample $(x_1, \ldots, x_n)$. Of course, a value of $y$ occurs with any vector $(x_1, \ldots, x_n)$ that contains $y$ 1's and $n - y$ 0's. The number of such vectors is given by \( \binom{n}{y} \) in Section 3.4C; thus the marginal probability function is

$$p_Y(y) = \binom{n}{y} p^y q^{n-y}, \quad y = 0, 1, \ldots, n,$$

$$= 0, \quad \text{otherwise}.$$

This is the binomial($n$, $p$) distribution (see Example 1.3.5 and Exercise 2.1.10).

Consider some examples of the binomial distribution. The binomial(4, 1/2) and binomial(4, 3/4) are given in Table 3.6 and plotted in Figure 3.11. For comparison

<p>| TABLE 3.6 |
| Biomial distributions |
|---|---|---|</p>
<table>
<thead>
<tr>
<th>( y )</th>
<th>( p_Y(y) )</th>
<th>( F_Y(y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1/16</td>
<td>1/16</td>
</tr>
<tr>
<td>1</td>
<td>4/16</td>
<td>5/16</td>
</tr>
<tr>
<td>2</td>
<td>6/16</td>
<td>11/16</td>
</tr>
<tr>
<td>3</td>
<td>4/16</td>
<td>15/16</td>
</tr>
<tr>
<td>4</td>
<td>1/16</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( y )</th>
<th>( p_Y(y) )</th>
<th>( F_Y(y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0039</td>
<td>0.0039</td>
</tr>
<tr>
<td>1</td>
<td>0.0469</td>
<td>0.0508</td>
</tr>
<tr>
<td>2</td>
<td>0.2109</td>
<td>0.2617</td>
</tr>
<tr>
<td>3</td>
<td>0.4219</td>
<td>0.6836</td>
</tr>
<tr>
<td>4</td>
<td>0.3164</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
purposes we record a few values for the corresponding distributions with \( n = 4 \) replaced by \( n = 40 \), Table 3.7.

<table>
<thead>
<tr>
<th>( y )</th>
<th>( p_y(y) )</th>
<th>( F_y(y) )</th>
<th>( y )</th>
<th>( p_y(y) )</th>
<th>( F_y(y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>10</td>
<td>0.0008</td>
<td>0.0011</td>
<td>10</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>20</td>
<td>0.1254</td>
<td>0.5627</td>
<td>20</td>
<td>0.0004</td>
<td>0.0006</td>
</tr>
<tr>
<td>30</td>
<td>0.0008</td>
<td>0.9997</td>
<td>30</td>
<td>0.1444</td>
<td>0.5605</td>
</tr>
<tr>
<td>40</td>
<td>0.0000</td>
<td>1.0000</td>
<td>40</td>
<td>0.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

In Section 1.1B we tossed a coin 100 times and obtained the observed frequency 53 for heads. If the probability for heads is \( p \), then the distribution describing possible frequencies is binomial \((100, p)\). This can be tabulated for various values of \( p \). We could then examine these distributions to see which ones are harmonious with the observed frequency 53.

### B MULTINOMIAL \((n, p_1, \ldots, p_k)\) DISTRIBUTION

We have been considering a dichotomy for the response of the random system being investigated: for example, \{reaction, no reaction\} for the drug to an animal, \{sterile, nonsterile\} for a water sample, \{ace, nonace\} for the toss of a die, \{\( H, T \)\} for the toss of a coin. Now consider more generally some partition \( \{B_1, \ldots, B_k\} \) for the response of the random system being investigated.

Let \( p_1, \ldots, p_k \) be the probabilities corresponding to \( B_1, \ldots, B_k \) and let \( x_1, \ldots, x_k \) be the corresponding indicators, Table 3.8; note that exactly one \( x(=1) \) in each row is different from zero. This gives the multivariate Bernoulli \((p_1, \ldots, p_k)\) distribution for the vector \( x = (x_1, \ldots, x_k)' \):

\[
p(x) = p_1^{x_1} \cdots p_k^{x_k}, \quad x_i = 0, 1; \sum x_i = 1,
\]
TABLE 3.8
Indicators and probabilities

<table>
<thead>
<tr>
<th>Event</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$\cdots$</th>
<th>$x_k$</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B_1$</td>
<td>1</td>
<td>0</td>
<td>$\cdots$</td>
<td>0</td>
<td>$p_1$</td>
</tr>
<tr>
<td>$B_2$</td>
<td>0</td>
<td>1</td>
<td>$\cdots$</td>
<td>0</td>
<td>$p_2$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\cdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$B_k$</td>
<td>0</td>
<td>0</td>
<td>$\cdots$</td>
<td>1</td>
<td>$p_k$</td>
</tr>
</tbody>
</table>

and zero otherwise, where $p_i \geq 0$ and $\sum p_i = 1$.

The indicators $x_1, \ldots, x_k$ could register occurrence for "type A reaction," "type B reaction," or "no reaction" for the drug to an animal ($k = 3$), or could register occurrence for "sterility," "contaminated (nonfecal)," or "contaminated (fecal)" for the water sample, or could register occurrence for "ace," "two or three," or "four or more" for the toss of a die, or could register "H with orientation in (0, 90°)," "H with orientation in [90°, 180°]," or "T" for the toss of a coin.

Now consider $n$ independent performances of the system being investigated. More formally, let $x_1, \ldots, x_n$ be a sample from the preceding multivariate Bernoulli ($p_1, \ldots, p_k$) distribution and write

$$X = (x_1, \ldots, x_n) = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & & \vdots \\ x_{k1} & \cdots & x_{kn} \end{pmatrix},$$

where the $j$th column records the indicator values for the $j$th sample element. The joint probability function is then given by formula (3.3.7):

$$p(X) = (p_1^{x_{11}} \cdots p_k^{x_{1n}}) \cdots (p_1^{x_{k1}} \cdots p_k^{x_{kn}})$$

if $x_{ij} = 0, 1$ and $\sum_i x_{ij} = 1$

and zero otherwise, where we write $y_1 = \sum_j x_{1j}, \ldots, y_k = \sum_j x_{kj}$ for the frequencies of the events $B_1, \ldots, B_k$, respectively.

Now consider the marginal distribution for the frequencies $(y_1, \ldots, y_k)$. A particular value for $(y_1, \ldots, y_k)$ will occur for any response sequence that involves $y_1$ $B_1$'s, $y_2$ $B_2$'s for the $n$ positions in the sequence. The number of such sequences is available from the concluding remarks in Section 3.4D: the number is

$$\binom{n}{y_1} \cdots \binom{n}{y_k},$$

which is the number of ways of selecting $y_1$ places for the $B_1$'s, $y_2$ places for the $B_2$'s, and so on. Thus the marginal probability function is

(2) \[ p_Y(y_1, \ldots, y_k) = \binom{n}{y_1} \cdots \binom{n}{y_k} p_1^{y_1} \cdots p_k^{y_k} \quad \text{if } y_1 = 0, 1, \ldots; \sum y_i = n \]

and zero otherwise. This is the multinomial$(n, p_1, \ldots, p_k)$ distribution. Note directly that the probabilities add to 1 (multinomial theorem in Section 3.4E).
EXAMPLE 1

Tossing a modified die: Consider the tossing of a modified symmetrical die that has one face marked 1, two faces marked 2, and three faces marked 3. The marginal distribution for the frequencies \(y_1, y_2, y_3\) is multinomial(5, 1/6, 2/6, 3/6); the probability function is

\[
p_Y(y_1, y_2, y_3) = \binom{5}{y_1, y_2, y_3} (1/6)^{y_1}(2/6)^{y_2}(3/6)^{y_3} \quad \text{if } y_1 = 0, 1, \ldots; \Sigma y_i = 5
\]

and zero otherwise; this is shown in Table 3.9 and plotted in Figure 3.12.

<table>
<thead>
<tr>
<th>(y_1)</th>
<th>(y_2)</th>
<th>(y_3)</th>
<th>7776(p_Y)</th>
<th>(y_1)</th>
<th>(y_2)</th>
<th>(y_3)</th>
<th>7776(p_Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>5</td>
<td>243</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>120</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>4</td>
<td>405</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>360</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>4</td>
<td>810</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>480</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>3</td>
<td>270</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>240</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1080</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>3</td>
<td>1080</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2</td>
<td>90</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>540</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1080</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>2</td>
<td>720</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>32</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td>7776</td>
</tr>
</tbody>
</table>

C HYPERGEOMETRIC(N, n, p) DISTRIBUTION

Consider a collection or population of \(N\) elements, each being either a type B element or a type G element. Let \(Np\) be the number of type B elements and \(Nq\) be the number of type G elements; \(p + q = 1\). Now consider a random sample of \(n\) from the population: a sequence of \(n\) elements such that each of the \(N^\text{(a)}\) possible sequences has the same probability \(1/N^\text{(a)}\). Let \(y\) be the number of type B elements in the sample.

For example, \(y\) could be the number of tagged fish in a sample of 200 from a lake containing 1000 tagged fish and 9000 unmarked fish, or \(y\) could be the number of people favoring a proposal in a sample of 100 from the 150,000 voters in a certain city, or \(y\) could be the number of defective tires in a sample of 50 from a shipment of 1000, or \(y\) could be the number of spades in a hand of 13 from a deck of 52.

Consider briefly how this differs from the binomial or "infinite" population situation. Suppose that we think of the sample being formed one element at a time. For the first element in the sample we must have by symmetry that \(P(B) = p\) and \(P(G) = q\). After we have the first element, we can consider the possibilities for the second element. The probabilities for B and G will be slightly changed and will depend, of course, on which element was obtained for the first position. Thus successive draws are not statistically independent and to calculate probabilities we would need the concept of conditional probability which is developed in Chapter 4.

The marginal probability distribution for the frequency \(y\) can be derived from the
symmetric distribution assigning probability \( 1/N^n \) to each possible sequence or equivalently from the symmetric distribution assigning probability \( 1/\binom{N}{n} \) to each possible sample set. We use the second of these symmetric distributions and note that the number of different sets that contain \( y \) B elements and \( n - y \) G elements is available from Sections 3.4A and 3.4C:

\[
\binom{Np}{y} \cdot \binom{Nq}{n-y}.
\]

Thus we obtain the marginal probability function:

\[
\rho_y(y) = \frac{\binom{Np}{y} \cdot \binom{Nq}{n-y}}{\binom{N}{n}} \quad \text{for } y \text{ an integer}
\]

and zero otherwise. This is the \textit{hypergeometric}(N, n, p) distribution. See Exercise 2.1.12 and note the effective range of the distribution.

Consider the numerical examples listed in Table 3.10 and plotted in Figure 3.13; notice how the distribution is more concentrated than the corresponding binomial distribution in Figure 3.11. The hypergeometric probability function can be expressed in the alternative form:

\[
\rho_y(y) = \frac{n! \cdot (Np)^y \cdot (Nq)^{n-y}}{N^n}, \quad y = 0, 1, \ldots, n,
\]

\[= 0 \quad \text{otherwise},\]

which can be justified directly by considering sequences rather than sets.
Chap. 3: Marginal Probability

FIGURE 3.13
Hypergeometric(8, 4, 1/2) and hypergeometric(8, 4, 3/4) distributions. Compare with Figure 3.11.

TABLE 3.10
Example distribution

<table>
<thead>
<tr>
<th>Hypergeometric(8, 4, 1/2)</th>
<th>Hypergeometric(8, 4, 3/4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>p(y)</td>
</tr>
<tr>
<td>---</td>
<td>------</td>
</tr>
<tr>
<td>0</td>
<td>1/70</td>
</tr>
<tr>
<td>1</td>
<td>15/70</td>
</tr>
<tr>
<td>2</td>
<td>36/70</td>
</tr>
<tr>
<td>3</td>
<td>15/70</td>
</tr>
<tr>
<td>4</td>
<td>1/70</td>
</tr>
</tbody>
</table>

EXAMPLE 2

Sample of desk calculators: A random sample of 4 units is taken from a shipment of 8 used desk calculators. Let y be the number of defective calculators in the sample. The marginal distribution of y is then hypergeometric (8, 4, D/8), where D is the number of defectives in the lot. The distribution is plotted in Figure 3.13 for the cases D = 4 and D = 8.

D MULTIVARIATE HYPERGEOMETRIC
(N, n, p₁, . . . , pₖ) DISTRIBUTION

Consider the finite population further. In the preceding section we assumed that each element was either a type B element or a type G element, a dichotomy. Now suppose that an element can be a B₁ element, a B₂ element, . . . , or a Bₖ element, a partition of the possible descriptions for an element.

More specifically consider a population of N elements consisting of Np₁ elements marked B₁, . . . , Npₖ elements marked Bₖ, where ∑ pᵢ = 1. And suppose that we sample n elements in succession randomly so that each of the Nᵢ sequences of elements has the same probability 1/Nᵢ. Let y₁ be the number of B₁ elements in the sample, y₂ be the number of B₂ elements in the sample, . . . , and yₖ the number of Bₖ elements. Note of course that ∑ yᵢ = n.

The marginal probability distribution for the frequencies (y₁, . . . , yₖ) can be derived from the symmetric distribution assigning probability 1/Nᵢ to each sequence or from the symmetric distribution assigning probability 1/(N) to each possible sample
set. We use the second of these and note that the number of different sets that contain 
\( y_1, B_1 \) elements, \( \ldots, y_k B_k \) elements is available from Sections 3.4A and 3.4C:

\[
\binom{Np_1}{y_1} \cdots \binom{Np_k}{y_k}.
\]

Thus we obtain the marginal probability function:

\[
(4) \quad p_Y(y_1, \ldots, y_k) = \frac{\binom{Np_1}{y_1} \cdots \binom{Np_k}{y_k}}{\binom{N}{n}}, \quad y_i = \text{integer, } \Sigma y_i = n
\]

and zero otherwise. This is the multivariate hypergeometric\((N, n, p_1, \ldots, p_k)\) distribution.

**EXAMPLE 3**

**Sample of outboard motors:** A random sample of 3 units is taken from a shipment of 6 used outboard motors. Let \( y_1 \) be the number of unrepairable units, \( y_2 \) the number of repairable units, and \( y_3 \) the number of satisfactory units. The marginal distribution of \((y_1, y_2, y_3)\) is then multivariate hypergeometric\((6, 3, p_1, p_2, p_3)\). The distribution for the case \( 6p_1 = 1, 6p_2 = 2, 6p_3 = 3 \) is given in Table 3.11 and plotted in Figure 3.14.

**TABLE 3.11**

<table>
<thead>
<tr>
<th>( y_1 )</th>
<th>( y_2 )</th>
<th>( y_3 )</th>
<th>( p_Y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1/20</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>6/20</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>1</td>
<td>3/20</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>2</td>
<td>3/20</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>6/20</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1/20</td>
</tr>
</tbody>
</table>

**FIGURE 3.14**

Multivariate hypergeometric\((6, 3, 1/6, 2/6, 3/6)\) distribution plotted against \((y_1, y_2)\); recall that \( y_3 = 3 - y_1 - y_2 \) or on points on \( \Sigma y_i = 3 \) in \( R^3 \).
E GENERAL FREQUENCY DISTRIBUTION

Consider a sample space $S$ with class $\mathcal{G}$ of events and probability measure $P$. And suppose that an investigator is interested in certain events $A_1, A_2, \ldots, A_n$, specifically, in the number of these events that occur in any performance of the system. Let

$Y = Y(s)$ be the number of events $A_1, \ldots, A_n$ that occur with the sample point $s$.

$Y(s) = \text{number of } A_j \text{ such that } s \in A_j.$

The probability function for $Y$ is available from Problem 1.5.3:

$$\rho_Y(y) = S_y - \binom{y + 1}{y} S_{y+1} + \cdots + (-1)^{n-y} \binom{n}{y} S_n,$$

where $S_0 = 1$ and

$$S_r = \sum_{i_1 < \cdots < i_r} P(A_{i_1} \cdots A_{i_r})$$

is the sum of the probabilities for the $r$th-order intersection sets. The probability for at least $r$ of the events is recorded in Problem 1.5.5.

Consider the classical example for this general distribution:

EXAMPLE 4

A well-shuffled deck of $N = 52$ playing cards is dealt successively as the dealer calls AS, KS, QS, $\ldots$, 3C, 2C. Let $A_i$ be the event the $i$th call correctly identifies the $i$th card dealt. And let $Y$ be the number of cards called correctly.

$$P(A_i) = \frac{(N-1)!}{N!} = \frac{1}{N} = P(A_i)$$

$$P(A_1A_2) = \frac{(N-2)!}{N!} = \frac{1}{N(N-1)} = P(A_1A_2), \quad i \neq j.$$  

$$P(A_1A_2A_3) = \frac{(N-3)!}{N!} = \frac{1}{N(N-1)(N-2)} = P(A_1A_2A_3), \quad i \neq j \neq k \neq i.$$  

Thus

$S_0 = 1,$

$S_1 = \sum P(A_i) = 1.$

$S_2 = \sum \frac{1}{N(N-1)} = \binom{N}{2} \frac{1}{N(N-1)} = \frac{1}{2}.$

$S_3 = \binom{N}{3} \frac{1}{N(N-1)(N-2)} = \frac{1}{3}.$

The distribution can now be given, Table 3.12. The Poisson(1) probabilities provide a convenient approximation:

$$\rho_Y(y) \approx \frac{1}{y!} e^{-1}.$$
TABLE 3.12
Distribution of cards

<table>
<thead>
<tr>
<th>y</th>
<th>$p_y(y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$1 \left(1 - 1 + \frac{1}{2!} - \cdots + \frac{1}{52!}\right)$</td>
</tr>
<tr>
<td>1</td>
<td>$1 \left(1 - 1 + \frac{1}{2!} - \cdots - \frac{1}{51!}\right)$</td>
</tr>
<tr>
<td>2</td>
<td>$\frac{1}{2!} \left(1 - 1 + \frac{1}{2!} - \cdots + \frac{1}{50!}\right)$</td>
</tr>
<tr>
<td>3</td>
<td>$\frac{1}{3!} \left(1 - 1 + \frac{1}{2!} - \cdots - \frac{1}{49!}\right)$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>51</td>
<td>$\frac{1}{51!} (1 - 1)$</td>
</tr>
<tr>
<td>52</td>
<td>$\frac{1}{52!} (1)$</td>
</tr>
</tbody>
</table>

**F EXERCISES: SAMPLE FROM A RANDOM SYSTEM**

1. Tabulate the binomial(3, 1/2) probability function.
2. Tabulate the binomial(5, 1/2) probability function.
3. Let $y$ be the number of even digits in a three-digit random number; tabulate the probability function for $y$.
4. Let $y$ be the number of zeros in a three-digit random number; tabulate the probability function for $y$.
5. Tabulate the multinomial(3, 1/3, 1/3, 1/3) probability function.
6. Tabulate the multinomial(3, 1/10, 3/10, 6/10) probability function.
7. Consider a symmetric die and suppose that two faces are labeled 1, two are labeled 2, and two are labeled 3. Consider three tosses and tabulate the probability function for $(y_1, y_2, y_3)$, where $y_i$ is the number of faces that show $i$.
8. Consider a four-digit random number and suppose that $(y_1, y_2, y_3)$ is reported, where $y_1$ is the number of 1s, $y_2$ is the number of 2s, and $y_3$ is the number of other digits. Record the probability function for $(y_1, y_2, y_3)$.
9. An article from production can be acceptable, repairable, or scrap. Suppose that production is stable and $p_1$, $p_2$, and $p_3$ are the respective probabilities for the condition of an article. And suppose that articles are collected in lots of 100. Give the probability function for $(y_1, y_2, y_3)$, where $y_1$, $y_2$, and $y_3$ are the number of items of the respective conditions in a lot.
10. A certain pair of black mice produce an offspring that is brown with probability 1/4 and black with probability 3/4. Let $y$ be the number of black mice in a litter of 5. Record the probability function for $y$; ignore the possibility of twins.
11. A bioassay procedure exposes 5 animals to a first dosage and 5 animals to a second dosage of a drug. Let $y_1$ be the number with a reaction in the first group and $y_2$ the number in the second group. Let $p_1$ and $p_2$ be the probability of reaction at the two dosages. Record the probability function for $(y_1, y_2)$. Assume independence between animals.
12. A coin may be biased. You decide to test the coin by making $n$ tosses ($= 5$ to simplify arithmetic; too small realistically). As a procedure for making a tentative judgment, you decide to accept unbiasedness if the number of heads is not equal to the extremes 0 and 5 and reject unbiasedness if the extremes occur. Let $p = P(H)$ be the probability for the
coin. If \( p = 1/2 \), what is the probability of rejection? If \( p = 3/4 \), what is the probability of rejection? Let \( \Omega(p) \) be the probability of rejection as a function of \( p \). Calculate the power function \( \Omega \) of this mechanized judgment procedure. Plot it. Is the procedure very reliable at this sample size?

**G  EXERCISES: SAMPLE FROM A FINITE POPULATION**

13 Tabulate the hypergeometric(6, 3, 1/2) probability function.
14 Tabulate the hypergeometric(6, 5, 1/2) probability function.
15 Three coils of cable are randomly sampled from a store of 6 coils. Let \( y \) be the number of sampled coils with defects. Tabulate the probability function for \( y \) if half the coils in the store contain defects.
16 Record the probability function for \( y \), the number of aces in a hand of 13 from a well-shuffled deck.
17 Record the probability function for \( z \), the number of aces in the combined hands (13 each) for \( N \) and \( S \) in a bridge distribution of 13 cards to each of 4 players.
18 Consider an urn containing 5 balls: 1 ball numbered 1, 1 ball numbered 2, and 3 balls numbered 3. The urn is thoroughly stirred and 3 balls are drawn at random. And \( (y_1, y_2, y_3) \) is reported; \( y_i \) is the number of \( i \) balls in the sample. Record the probability function and tabulate.
19 A hand of 13 cards is dealt from a well-shuffled standard deck of 52 cards. Calculate the probability function
   (a) \( p_1(x) \), where \( x \) is the number of spades in the hand.
   (b) \( p_2(y) \), where \( y \) is the number of hearts in the hand.
   (c) \( p_3(x, y) \), where \( x \) and \( y \) are as given in parts (a) and (b).
20 A hand of 13 cards is dealt from a well-shuffled deck; let \( x, y, z, \) and \( w \) be the number of spades, hearts, diamonds, clubs, respectively. Record the probability function \( p(x, y, z, w) \).
21 An urn contains \( R \) red balls, \( W \) white balls, \( B \) blue balls, where \( N = R + W + B \). The urn is thoroughly mixed and a random sample of \( n \) balls is drawn; let \( r, w, \) and \( b \) be the number of red, white, and blue balls, respectively. Record the probability function for \( (r, w, b) \).
22 A production lot of 100 items contains \( D \) defectives. A sample of 5 is taken at random from the lot and the number of defectives \( y \) is reported.
   (a) Record the probability function for \( y \).
   (b) Tabulate the probability function for the case \( D = 0 \); the case \( D = 1 \); and the case \( D = 2 \).
   (c) A quality-control procedure is to reject the lot if any defectives are found; the operating characteristic function \( Q \) of the procedure records \( Q(D) \), the probability of acceptance when the lot contains \( D \) defectives. Partially tabulate and sketch the operating characteristic \( Q \).

**H  COMPUTER EXERCISES**

23 Obtain a sample of 1000 random digits and tabulate the proportion of 0's, 1's, ..., 9's. Does the empirical distribution seem reasonably symmetrical?
24 Let \( y \) be the number of even digits in a block of five random digits. For a sample of 1000 blocks, tabulate the proportion with \( y = 0, y = 1, y = 2, y = 3, y = 4, \) and \( y = 5 \). Compare with the theoretical distribution in Exercise 2.
25 Let \( y_1, y_2, \) and \( y_3 \) be, respectively, the number of 0's, the number that are multiples of 3, and the number of other digits in a block of three random digits. For a sample of 1000 blocks,
tabulate the proportion for each value for \( (y_1, y_2, y_3) \). Compare with the theoretical distribution in Exercise 6.

### PROBLEMS

26 Binomial\((n, p)\) probabilities can be calculated in sequence. Show that \( r(y) = \rho(y) / \rho(y - 1) = (n - y + 1)/p / \rho(y) \), where \( \rho(y) \) is the binomial probability function.

27 (continuation) The mode of a discrete distribution is the sample point having maximum probability. Show that the mode of the binomial distribution is \( \lceil (n + 1)p \rceil \), where \( \lceil x \rceil \) is the largest integer \( \leq x \). However, if \( (n + 1)p \) is an integer, there is a double mode, at \( (n + 1)p - 1 \) and at \( (n + 1)p \).

28 Let \( F \) be the distribution function for the binomial\((n, p)\) distribution. Show that at the integers \( y = 0, 1, \ldots, n - 1 \),

\[
F(y) = \sum_{x=0}^{y} \binom{n}{x} p^x q^{n-x} = \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y)} \int_0^1 t^y(1-t)^{n-y-1} dt;
\]

the integral is an incomplete beta integral.

29 Sum of binomials: Let \( y_1 \) and \( y_2 \) be independent and \( y_1 \) be binomial\((n_1, p)\). Use the convolution formula (3.2.9) to show that \( y_1 + y_2 \) is binomial\((n_1 + n_2, p)\).

30 (continuation) Represent each \( y_i \) in terms of a sample from the Bernoulli\((p)\) and thus show that \( y_1 + y_2 \) is binomial\((n_1 + n_2, p)\).

31 Poisson\((\lambda)\) probabilities can be calculated in sequence: calculate \( r(y) = \rho(y) / \rho(y - 1) \).

32 (continuation) Show that the mode of the Poisson\((\lambda)\) is \( \lceil \lambda \rceil \); compare with Problem 27. However, if \( \lambda \) is an integer, there is a double mode, at \( \lambda - 1 \) and at \( \lambda \).

33 By appropriate differentiation and integration, show that the Poisson distribution function can be reexpressed as an incomplete gamma integral:

\[
F(y) = \sum_{0 \leq x \leq y} e^{-\lambda x} \frac{x^y}{x!} = \frac{1}{y!} \int_{\lambda}^{\infty} e^{-t y} dt
\]

for \( y = 0, 1, 2, \ldots \).

34 Sum of Poisson: Let \( y_1 \) and \( y_2 \) be independent and \( y_1 \) be Poisson\((\lambda_1)\). Use the convolution formula (3.2.9) to show that \( y_1 + y_2 \) is Poisson\((\lambda_1 + \lambda_2)\).

35 Let \( \rho(y; p) \) be the binomial\((n, p)\) probability function. Let \( \lambda = np \) be fixed; then \( p = \lambda/n \); show that \( \lim \rho(y; p) \) as \( n \to \infty \) is given by the Poisson\((\lambda)\) probability function.

36 Negative binomial\((r, p)\) distribution: Consider a sequence of independent indicators, each Bernoulli\((p)\). Let \( y \) be the number of \( 0 \)'s until \( r 1 \)'s have occurred. Show that the marginal distribution of \( y \) is negative binomial\((r, p)\); recall Exercise 2.1.11.

37 Combining multinomial coordinates: Let \( (y_1, y_2, y_3) \) be multinomial\((n, p_1, p_2, p_3)\) and consider \( (y_1, y_2, y_3) \). Show that \( y_1 \) is binomial\((n, p_1)\).

38 (continuation) Let \( (y_1, \ldots, y_k) \) be multinomial\((n, p_1, \ldots, p_k)\). Show that the marginal distribution of \( (y_1, \ldots, y_{k-2}, y_{k-1} + y_k) \) is multinomial\((n, p_1, \ldots, p_{k-2}, p_{k-1} + p_k)\).

39 (continuation) Let \( (y_1, \ldots, y_k) \) be multinomial\((n, p_1, \ldots, p_k)\). And let \( w_i \) be the sum of the first \( r_i \)'s, \( w_s \) the sum of the next \( r_s \)'s, \ldots, \( w_k \) the sum of the last \( r_k \)'s. And let \( P_i \) be the sum of the first \( p_i \)'s, and so on for \( P_2, \ldots, P_s \). Use Problem 38 to justify that the marginal of \( (w_1, \ldots, w_s) \) is multinomial\((n, P_1, \ldots, P_s)\).

40 Let \( \rho(y_1, \ldots, y_k) \) be the multinomial\((n, p_1, \ldots, p_k)\) probability function. Let \( \lambda_1 = np_1, \ldots, \lambda_{k-1} = np_{k-1} \) be fixed; then \( p_i = \lambda_i/n \) for \( i = 1, \ldots, k - 1 \). Show that as \( n \to \infty \), the probability function for \( y_1, \ldots, y_{k-1} \) converges to the probability function for \( k - 1 \) independent Poissons with parameters \( \lambda_1, \ldots, \lambda_{k-1} \).

41 Let \( (y_1, \ldots, y_k) \) be multivariate hypergeometric\((N, n, p_1, \ldots, p_k)\). Show that the marginal distribution of \( (y_1, \ldots, y_{k-2}, y_{k-1} + y_k) \) is multivariate hypergeometric\((N, n, p_1, \ldots, p_{k-2}, p_{k-1} + p_k)\). Use Problem 3.4.7.
J PROBLEMS: GENERAL FREQUENCY DISTRIBUTION

42 Use the general frequency distribution (5) to derive the binomial(n, p) probability function for the number of 1's in a sample of n from the Bernoulli(p).

43 Derive the probability function for the number y of ace-king pairs in a bridge hand.

44 Derive the probability that a bridge hand contains at least one club from each suit.

45 Derive the probability that at least one of the four hands at bridge contains a complete suit.

46 General bivariate frequency distribution: Consider events $A_1, \ldots, A_n$ and events $B_1, \ldots, B_m$ for a random system. Let x be the number of events $A$ that occur and y be the number of events $B$ that occur. Show that

$$P(x, y) = S_{x,y} = \binom{x}{x} S_{x+1,y} = \binom{y+1}{y} S_{x,y+1} \cdots$$

$$= \sum_{\alpha, \beta = 0}^{\alpha} (-1)^{\alpha + \beta} \binom{x + \alpha}{x} \binom{y + \beta}{y} S_{x+\alpha, y+\beta},$$

where

$$S_{x,y} = \sum_{i_1 < \cdots < i_x} P(A_{i_1} \cdots A_{i_x} B_{h} \cdots B_{y}), \quad x \leq n, \quad y \leq m,$$

$$= 0, \quad \text{otherwise.}$$

Use Problem 2 of Supplément Section 1.5.

K PROBLEMS: MULTINOMIALS OVER OTHER DISTRIBUTIONS

47 Two symmetric coins are tossed 4 times and $(y_1, y_2, y_3)$ is observed, where $y_1$ is the number of double heads, $y_2$ is the number of double tails, and $y_3$ is the number of mixed tosses. Derive the probability function for $(y_1, y_2, y_3)$.

48 (continuation of Exercise 18) The urn is stirred, a ball is drawn and replaced; the urn is stirred, a ball is drawn and replaced; the urn is stirred, a ball is drawn and replaced; $(y_1, y_2, y_3)$ is reported, where $y_i$ is the number of $i$ balls in the sample. Derive the probability function.

49 (continuation of Exercise 16) For 10 successive hands of bridge with thorough shuffling at each step, find the probability function for $(y_0, y_1, y_2, y_3, y_4)$, where $y_i$ is the number of hands with $i$ aces.

SUPPLEMENTARY MATERIAL

Supplement Section 3.6 is on pages 526-537.

The proof is given for the following result: $F_{XY} = F_X F_Y$ is a necessary and sufficient condition for the independence of $X$ and $Y$.

The normal theory basis for chi square, Student t, and $F$ is an application of the theory in this chapter and is presented in detail. The noncentral theory is also presented, in text and problems.

Some order statistic distributions illustrate the theory in this chapter and are developed in detail.
Product spaces are the means by which bigger models are built from smaller initial models; the formation of product spaces, as touched on in Section 3.3, is examined in detail.

Various results for measurable functions are recorded; for example, the composition of measurable functions is measurable; and the limit of a sequence of measurable functions is measurable.

NOTES AND REFERENCES

For some references to other texts, see the Notes and References in Chapter 1. In this chapter we have organized the material in terms of two basic probability concepts, marginal probability and independence, and in terms of a major application, the derivation of the standard frequency distributions, the binomial and multinomial. These distributions need the two probability concepts, and we have chosen to examine the concepts first, in special detail. We feel that the two probability concepts are very important, and have separated the material accordingly. The common separation by type of distribution, discrete or continuous, has been deemphasized, and appears just as a minor difference in formulas, $\Sigma$ or $\int$. 
In Chapter 3 we examined how probability is mapped from one space to another by means of functions (or random variables). And we discussed various ways of describing the resulting marginal distributions.

With marginal probability as the basis we then examined the important concept of statistical independence. On the surface this is concerned with how a large model can sometimes be split into essentially independent components. In fact, however, the concept is more concerned with combining simple models to form larger more complex models—with model building.

We then used independence and marginal probability to derive the important binomial, multinomial, and hypergeometric distributions.

All our examples so far have involved a random system about to be performed or a random system that has been performed with the outcome concealed from the investigator. These two situations represent, in fact, two extremes, and we are often somewhere in between: a random system has been performed and the outcome is partially concealed from the investigator. Conditional probability is concerned with this intermediate situation involving partial information concerning an outcome.

In Chapter 4 we examine this very important and fundamental concept, conditional probability: Section 4.1 is concerned with the general definition and interpretation; and Section 4.2 is concerned with formulas and methods on real spaces.
4

CONDITIONAL PROBABILITY

4.1

DEFINITION AND INTERPRETATION

Conditional probability is concerned with describing a real-world system when there is partial information concerning the outcome of the system. In this section we examine the definition and interpretation of conditional probability. Then in Section 4.2 we examine various formulas and methods for calculating conditional probability on real spaces.

Conditional probability is the most important concept of probability theory—after probability itself. We have delayed its introduction to this point in order to have a broad contact with probability methods and be able to examine the concept in depth.

Consider an engineer investigating reliability. A control system has two primary components in series. The first component is a new unit and the second component has been in service and survived 900 hours. The distribution for the life of new components is known to be approximately normal(1000, 100). For the first component we can use this normal distribution directly. However, for the second component we are interested in residual lifetime given survival for 900 hours: we know that the realized lifetime for the second component will be at least 900 hours.

Consider an actuary investigating 1-year term insurance policies for 40-year-olds. Actuarial investigations lead to life tables that contain (implicitly) the distribution of lifetime for a person just born in a certain racial and social context. However, for 40-year-olds we know that the realized lifetime is at least 40; in fact, the outcomes of interest have been selected to be those with lifetime at least 40 and not yet 41.

Consider a geneticist examining a certain strain of brown and black mice. A gene for color can be dominant $B$ for black or recessive $b$ for brown. A mouse has two genes for color and can be of the following three genotypes: $BB, Bb, bb$. A mouse $bb$ without a dominant $B$ is brown; a mouse with a dominant $B$ is black. Now consider the mating of a $Bb$ mouse with a $Bb$ mouse. Genetic theory prescribes the following probability for each kind of offspring:

<table>
<thead>
<tr>
<th></th>
<th>$BB$</th>
<th>$Bb$</th>
<th>$bb$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>1/4</td>
<td>1/2</td>
<td>1/4</td>
</tr>
</tbody>
</table>

But suppose that the geneticist is interested in a particular offspring that is black. Clearly, the case $bb$ is excluded. He is interested in outcomes of the random system that have been selected on the basis of having a dominant black gene.
A SELECTION OF PERFORMANCES: EXAMPLE

Consider a random system involving the determination of the sex of an offspring. Let $M$ denote male and $F$ denote female; then in the standard context we have the distribution

\[
\begin{array}{cc}
M & F \\
1/2 & 1/2 \\
\end{array}
\]

Or consider a system in which repetitions are more simply visualized: the tossing of a coin and observing the upward face. Let $H$ denote heads and $T$ denote tails; then in the standard context we have the distribution

\[
\begin{array}{cc}
H & T \\
1/2 & 1/2 \\
\end{array}
\]

For this case suppose that we have an operator who tosses the coin twice and records $(x_1, x_2)$, where $x_i$ is the indication for heads on the $i$th toss; then we have

\[
\begin{array}{cccc}
(0, 0) & (0, 1) & (1, 0) & (1, 1) \\
1/4 & 1/4 & 1/4 & 1/4 \\
\end{array}
\]

Now suppose that the operator selects certain performances and ignores the remaining performances. Specifically, suppose that the operator selects those performances on which $C$, "at least one head," occurs and ignores those performances on which $C^c$ occurs: see Figure 4.1.

We now discuss the probability model appropriate to such selected performances. Consider a large number, say 1000, of repetitions of the initial random system involving two tosses. The performances yielding $(0, 1), (1, 0), (1, 1)$ will pass selec-
tion; there will be approximately 250 for each of these outcomes. The performances yielding (0, 0) will not pass selection; there will be approximately 250 of these rejected outcomes. Among selected performances we have the proportions

\[
\frac{250}{750} \quad \frac{250}{750} \quad \frac{250}{750}
\]

for the three possible outcomes

(0, 1) \quad (1, 0) \quad (1, 1).

Accordingly, we write

\[
P((0, 1) : C) = \frac{1}{3}, \quad P((1, 0) : C) = \frac{1}{3}, \quad P((1, 1) : C) = \frac{1}{3}
\]

for the estimated probabilities given the event C. These are called "conditional probabilities given the event C" and are interpreted in terms of performances of the system selected on the basis of the event C.

**B DEFINITION OF CONDITIONAL PROBABILITY**

Now consider in general a random system with sample space $S$, with class $\mathcal{A}$ of events, and with probability measure $P$. And suppose that performances of the system are selected in accord with an event $C$; thus any performance yielding an outcome in $C$ is reported to an investigator and the remaining outcomes in $C^c$ are ignored. We investigate probabilities that describe selected performances; see Figure 4.2.

Consider a large number $N$ of performances of the basic system. The number of outcomes that pass selection is approximately $NP(C)$. Suppose now that we are interested in some event $A$. The number of outcomes that pass selection and yield the event $A$ is approximately $NP(AC)$. Thus for the investigator examining selected performances, the proportion or estimated probability for the event $A$ is

**FIGURE 4.2**

Selection in accord with event C. The proportion for A given selection is $P(A : C) = P(AC) / P(C)$.
\[
\frac{NP(AC)}{NP(C)} = \frac{P(AC)}{P(C)}.
\]

This leads us to the following formal definition:

**DEFINITION 1**

The **conditional probability** of \(A\) given \(C\) is

\[
(1) \quad P_C(A) = P(A : C) = \frac{P(AC)}{P(C)}
\]

provided that \(P(C) \neq 0\).

The interpretation of \(P_C(A)\) is in terms of the long-run proportions for the event \(A\) among outcomes that have passed selection based on \(C\).

Consider the construction of a model for a random system with selection. Let \((\mathbb{S}, \mathcal{A}, P)\) be the model for the unselected system, and let \(C\) be the event for selection. For the selected system it is convenient to use the same sample space \(\mathbb{S}\) even though certain sample points cannot occur. And it is convenient to use the same class \(\mathcal{A}\) of events. The function \(P_C\) given by Definition 1 gives the probability for any event \(A\) given the condition \(C\); it is straightforward to show that \(P_C\) is a probability measure (Problem 7). Thus we obtain the model

\[(\mathbb{S}, \mathcal{A}, P_C)\]

describing the selected system. Informally, we shall refer to the conditional distribution \(P_C\) for the \(C\) selected performances of the system.

A conditional distribution \(P_C\) is used to describe a partially concealed realization from a random system, the available information being represented by \(C\). Some examples have been mentioned earlier in this section; further examples are given in Section C. A conditional distribution is the appropriate distribution for a great many applications.

The definition of conditional probability gives us an alternative condition for independence:

**PROPOSITION 2**

If \(P(C) \neq 0\), then the events \(A\) and \(C\) are statistically independent if and only if the conditional probability of \(A\) given \(C\) is equal to the marginal probability of \(A\), that is, \(P(A : C) = P(A)\).

The proof follows immediately from Definition 3.3.1.

**C EXAMPLES**

Consider some examples with coins, urns, and cards, simple examples that focus clearly on certain important aspects of conditional probability.
A coin is tossed. If \( H \), then go to \( H \) urn, which has 2 \( W \) and 1 \( B \) balls; draw one ball. If \( T \), then go to \( T \) urn, which has 1 \( W \) and 1 \( B \) balls; draw one ball.

**Example 1**

**A priori probability:** Suppose that a technician tosses a symmetric coin. If the coin shows \( H \), he mixes the balls in an \( H \) urn containing two white balls and one black ball, and then draws one ball and notes its color. If the coin shows \( T \), he mixes the balls in a \( T \) urn containing one white ball and one black ball, and then draws one ball and notes its color. An outcome from these operations gives the face for the coin and the color for the ball. This is presented schematically in Figure 4.3 and as a probability tree in Figure 4.4.

As a first situation, suppose that the technician tosses the coin and reports the result to the investigator before drawing from the appropriate urn. The investigator receives the report \( H \) and is concerned with the probability for \( B \), "black ball." The probability refers to a future draw from the \( H \) urn and is

\[
P(B : H) = \frac{1/6}{3/6} = \frac{1}{3}.
\]

This is an a priori probability describing an outcome in the future. Note that the selected performances all involve drawing a single ball from the \( H \) urn, and thus refer to an ordinary performance of a component system; the probability can be calculated directly from the component system.

**Example 2**

**A posteriori probability:** Consider a second situation for the preceding coin–urn example. Suppose that the technician tosses the coin, draws from the urn, and then reports the face showing on the coin but nothing concerning the color of the ball. The investigator receives the report \( H \) and is concerned with the probability for \( B \), "black ball." The probability refers to a realization of the full system. For the selected performances, however, we are concerned essentially with the same \( H \) urn as in Example 1; we have

\[
P(B : H) = \frac{1/6}{3/6} = \frac{1}{3}.
\]

This is an a posteriori probability describing a realized outcome that is partially concealed from the investigator. The interpretation is in terms of a large number of
A coin is tossed and a ball is drawn accordingly from the \( H \) urn or the \( T \) urn. Branch probabilities and terminal probabilities are recorded.

observations on the selected system, selected to agree with the information available concerning the realization.

**EXAMPLE 3**

**A posteriori probability for an antecedent event:** Consider a third situation for the coin-urn example. Suppose that the technician tosses the coin, draws from the urn, and then reports the color of the ball but nothing concerning the face on the coin. The investigator receives the report \( B \), "black ball," and is concerned with the probability for \( H \). The probability refers to a realized toss of the coin, indeed to a coin toss that was antecedent in time to the drawing of the black ball:

\[
P(H : B) = \frac{1/6}{1/6 + 1/4} = \frac{2}{5}.
\]

This is also called an a posteriori probability. Indeed, it is a special a posteriori probability, a probability for an event \( H \) that would be antecedent in time to the event \( B \) representing the available information.

**EXAMPLE 4**

**Card distributions at bridge:** A deck of 52 playing cards is dealt so that 13 cards go to each of E, S, W, and N. Suppose that N can see the cards in his own hand and also those in the hand of S. And suppose that he observes 10 spades and is interested in the probability that the remaining 3 spades are split 2 to E and 1 to W. Let \( C \) be the event 10 spades in the combined hands of N and S, or equivalently 3 spades in the hands of E and W. And let \( A \) be the event 2 spades in the hand of E and 1 spade in the hand of W. Then

\[
P(A : C) = \frac{P(AC)}{P(C)} = \frac{\binom{13}{2} \binom{39}{11} \cdot \binom{11}{2} \binom{28}{12} / \binom{52}{13} \binom{39}{13}}{\binom{13}{3} \binom{39}{23} / \binom{52}{26}};
\]

this follows the pattern of calculation in Section 3.4: for the numerator, a set of 13 for E and then a set of 13 for W; for the denominator, a set of 26 for E and W.
For applications we need to be very careful when considering such a probability. Perhaps certain bidding has taken place, bidding that to some degree partially eliminates possibilities in $C$; for example, certain card distributions might require a different pattern of play and we would not be faced with a concern for the spade split. The probability as just calculated could then be inappropriate.

**D MULTIPLICATION FORMULAS**

So far we have been thinking of conditional probability as giving a description of a system that has been selected, or giving a description of a realization from such a system where there is partial information concerning the realization. Conditional probability, however, can also be used for computational purposes.

The definition of conditional probability can be rewritten as a *multiplication formula*,

$$P(AC) = P(C)P(A : C) \quad \text{if } P(C) \neq 0.$$  

This can be used for calculating probabilities if we have available some direct method for calculating the conditional probability; symmetry can often be used for this.

The multiplication formula can be applied repeatedly. For example, with $P(BC) \neq 0$ we have

$$P(ABC) = P(C) \cdot \frac{P(BC)}{P(C)} \cdot \frac{P(ABC)}{P(BC)} = P(C)P(B : C)P(A : BC).$$

Thus if appropriate conditional probabilities are available, say from symmetry, then the probability for a compound event can be calculated by multiplying together the appropriate conditional probabilities.

**EXAMPLE 5**

**Randomization:** Treatments $T_1$, $T_2$, $T_3$, and $T_4$ are to be randomly assigned to guinea pigs $P_1$, $P_2$, $P_3$, and $P_4$. Four cards with designations $T_1$, $T_2$, $T_3$, and $T_4$ are thoroughly shuffled and dealt one by one to positions $P_1$, $P_2$, $P_3$, and $P_4$; the treatments are given accordingly to the animals. Find the probability that the treatments are assigned $T_1$, $T_2$, and $T_3$, respectively, to $P_1$, $P_2$, and $P_3$. The probability that $T_1$ goes to $P_1$ is $1/4$, as $T_1$ is one of 4 symmetrical possibilities for $P_1$; the probability that $T_2$ goes to $P_2$ given $T_1$ to $P_1$ is $1/3$, as $T_2$ is one of the 3 remaining symmetrical possibilities for $P_2$. The probability that $T_3$ goes to $P_3$ given $T_1$, $T_2$, respectively, to $P_1$, $P_2$ is $1/2$, as $T_3$ is one of 2 remaining symmetrical possibilities for $P_3$. Thus the required probability is $$(1/4) \cdot (1/3) \cdot (1/2) = 1/24.$$ This particular probability, however, is, of course, easily calculated directly.

**EXAMPLE 6**

**Industrial acceptance sampling:** Consider a lot containing $N = 10$ articles of which $D = 3$ are defective and $N - D = 7$ are good. Suppose that the articles are thoroughly mixed and then tested one by one. Find the probability that the last defective is found on the fifth test. Let $C$ be the event ''2 defectives in the first 4 tested,'' and $A$ be ''a defective on the fifth test.''

The required probability is then
\[ P(AC) = P(C) \cdot P(A \mid C) = \frac{\binom{3}{2} \binom{7}{2}}{\binom{10}{4}} \cdot \frac{1}{6}. \]

The first factor is calculated from symmetry for a set of 4; the second factor from symmetry for a set of 1 but from the lot as modified by the first 4 removed.

**E MARGINAL PROBABILITIES AND INDEPENDENCE**

Sometimes conditional probabilities can be used in a somewhat different way to calculate a marginal probability, say \( P(A) \). Suppose that we have available the conditional probabilities \( P(A \mid C) \) and \( P(A \mid C^c) \) and we find that

\[ (4) \quad P(A \mid C) = P(A \mid C^c). \]

Then

\[ (5) \quad P(A) = P(AC) + P(AC^c) = P(C)P(A \mid C) + P(C^c)P(A \mid C^c) = P(A \mid C)P(C) + P(C^c) = P(A \mid C); \]

thus we have calculated the probability \( P(A) \).

*Note:* If equation (4) holds, then we can deduce (5) and obtain by Proposition 2 the independence of \( A \) and \( C \). Conversely, if \( A \) and \( C \) are independent [with \( P(C) \neq 0 \neq P(C^c) \)], then Proposition 2 gives equation (4). This proves the essentials of the following proposition, presenting independence in terms of conditional probability.

**PROPOSITION 3**

\( A \) and \( C \) are statistically independent if and only if

\[ P(A \mid C) = P(A \mid C^c) \]

wherever defined; that is, the conditional probability does not depend on the condition.

**EXAMPLE 7**

*Coin and two urns:* A symmetric coin is tossed. If the coin shows \( H \), "heads," then a ball is drawn from an \( H \) urn containing 1 black ball and 2 white balls; if the coin shows \( T \), "tails," then a ball is drawn from a \( T \) urn containing 2 black balls and 4 white balls. Clearly from symmetry we have

\[ P(B \mid H) = 1/3, \quad P(B \mid T) = 2/6 = 1/3. \]

Then by the preceding proposition we have that \( B \) and \( H \) are statistically independent and that

\[ P(B) = P(B \mid H) = 1/3. \]

More generally suppose that we have a partition \( \{C_1, \ldots, C_n\} \) of the sample space \( \mathcal{S} \) and find that
(6) \( P(A : C_i) \) is functionally independent of \( i = 1, \ldots, r \).

Then

\[
(7) \quad P(A) = \sum_{i=1}^{r} P(AC_i) = \sum_{i=1}^{r} P(C_i)P(A : C_i) = P(A : C_i) \sum_{i=1}^{r} P(C_i) = P(A : C_i);
\]

thus we have calculated the probability \( P(A) \).

**Note:** If equation (6) holds, then we can deduce (7) and then obtain by Proposition 2 the independence of \( A \) and each \( C_i \) in the partition; it follows that \( A \) is independent of the \( \sigma \)-algebra generated by the \( C \)'s. Conversely, if \( A \) is independent of the \( C_i \)'s, then the conditional probabilities \( P(A : C_i) \) do not depend on \( C_i \). This proves the following:

**PROPOSITION 4**

\( A \) is statistically independent of the partition \( \{C_1, \ldots, C_r\} \) if and only if

\[
P(A : C_i)
\]

wherever defined is functionally independent of \( i \); that is, the conditional probability is functionally independent of the condition.

**F PROBABILITIES FOR ANTECEDENT EVENTS**

Example 3 records a simple example of an a posteriori probability for an antecedent event. Consider this more generally. Suppose that we have a system with two stages in sequence. For the first stage, let \( H_1, \ldots, H_r \) be disjoint events with probabilities \( P(H_1), \ldots, P(H_r) \). And for the second stage, let \( A \) be the observable event and suppose that its probability \( P(A : H_i) \) depends on the outcome \( H_i \) from the first stage; see Figure 4.5.

Now suppose that we have observed the outcome from the second stage, say the occurrence of \( A \); and suppose that we are interested in probabilities concerning antecedent events from the first stage. From Definition 1, we have

**FIGURE 4.5**

First stage, giving \( H_1, \ldots, H_r \); and second stage, giving \( A \) or \( A^c \).
Sec. 4.1: Definition and Interpretation

(8) \[ P(H_i:A) = \frac{P(H_i)P(A:H_i)}{P(A)} = \frac{P(H_i)P(A:H_i)}{\sum_1^r P(H_i)P(A:H_i)} \]

This is the ordinary expression for conditional probability, but note that we are expressing the probability for an antecedent event, stage 2 to stage 1, in terms of the forward probabilities, stage 1 to stage 2. Formula (8) is sometimes called Bayes formula. For a simple example, see the earlier Example 3.

Now consider a system that can be in any one of several conditions \( H_1, \ldots, H_r \). And suppose that the system is performing an outcome characterized by the event \( A \). The investigator, of course, is interested in which condition \( H_i \) applies to the system. Certainly he can examine the probability for what has occurred; the probability \( P(A:H_i) \) as a function of the condition \( H_i \); call it the likelihood of \( H_i \). Some statistical analysis from this point of view is developed in Chapter 8.

Some statisticians feel that it is appropriate to attach numbers \( P(H_1), \ldots, P(H_r) \) to the possible conditions \( H_1, \ldots, H_r \); the numbers are chosen to represent in some way feelings concerning the plausibility of the various \( H_1, \ldots, H_r \). Then they treat these numbers as probabilities for a hypothetical initial system and call them prior probabilities. The use of formula (8) then gives the posterior probabilities

\[ P(H_i:A) = \frac{P(H_i)P(A:H_i)}{\sum_1^r P(H_i)P(A:H_i)}, \quad i = 1, \ldots, r. \]

These “probabilities” have meaning at most to the degree that it is reasonable or proper to describe feelings by means of numbers \( P(H_1), \ldots, P(H_r) \). This Bayesian analysis will be discussed briefly in Supplement Section 9.6.

G CAUTIOUS WITH CONDITIONAL PROBABILITY

In Example 4 we indicated briefly certain hazards that can arise in the application of conditional probabilities. We chose a card-dealing example for this as the basic calculations are easy and yet the potential for complications in applications is very large. The example involved a distribution of 52 cards into bridge hands, the observation of hands N and S, and the question of how cards are split between E and W.

The picture becomes more complicated and indeed realistic if we think of initial bidding and then play until a small number of cards remains in each hand. A probability calculation can be made as in Example 4; it would typically be based on symmetry and on the cards actually observed. There are a variety of other factors, however: the pattern of bidding can partially or completely eliminate some of the possibilities in the formal condition; the pattern of play can also partially or completely eliminate some of the possibilities; and indeed the players can use randomness in their choice of play. In short, there can be many complications that can invalidate the simplistic calculation based on symmetry among the various possibilities. However, if the conventions of bidding and play and any randomness used therein are clearly specified, then a conditional probability can be calculated from the corresponding, more complex model.

The interpretation of conditional probability is in terms of a selected system. This interpretation, in turn, provides the key to the proper application of the concept. Three conditions are sufficient for its application to a partially concealed realization of a random system.
The investigator receives information concerning the performance of the system; let $C$ be the set of outcomes for which the information is true. The investigator must know that:

1. **One element of $C$ has occurred.** In effect, this says that he knows that the information concerning the performance is true. The investigator must also know something about the source of the information concerning the realization, that is, how it was generated and how it was selected for transmission to the investigator.

2. **Each element in $C$ could have occurred.** In effect, this says that there must be no explicit selection; all the apparent possibilities in $C$ must be real possibilities and not excluded by the way the information was generated or transmitted to the investigator.

And the generation and transmission of the information must be such that an identical situation would be obtained for each possibility in $C$.

3. **Each element in $C$ would lead to the same information.** In effect, this says that there must be no implicit selection; the full information to the investigator must be identical to the information that the outcome is in $C$.

These conditions require that the system and the generation of the information be equivalent to a selected system as described in Section B.

### EXERCISES

A technician tosses a symmetrical copper coin and a symmetrical steel coin and then reports to an investigator, "the number of heads is at least 1." Determine the probability that both are heads if the technician's reporting pattern is as follows:

1. The technician observes both coins and reports "all tails" if the number of heads is zero; reports "the number of heads is at least 1" if the number of heads is 1 or 2.

2. The technician observes both coins and reports "all tails" if the number of heads is zero; reports "the number of heads is at least 1" if the number of heads is 1; reports "all heads" if the number of heads is 2.

3. The technician observes the copper coin and reports "the number of heads is at least 1" if the copper coin shows heads; reports "the number of tails is at least 1" if the copper coin shows tails.

### PROBLEMS

4. A bridge player finds exactly 9 spades in his own and partner's hands. Give a combinatorial expression for the probability that the remaining 4 spades split 4–0 in the hands left–right; split $x–y$ in the hands left–right where $x + y = 4$.

5. A bridge player finds exactly 8 diamonds in his hand. Give an expression for the probability that his partner has 0 diamonds; at least 1 diamond; and at least 2 diamonds.

6. From family background it is known that with probability $1/2$ a woman is a carrier of hemophilia. If she is a carrier, then with probability $1/2$ a male offspring has the disease; transmission is independent for different offspring. If she is not a carrier, then a male offspring does not inherit the disease. Determine the probability that she is a carrier if her first son is normal; and if her first two sons are normal.

7. Show that $P_C$ in Definition 1 is a probability measure.

8. A family has two children. It is reported that there is at least one boy. Under three reporting patterns as indicated by Exercises 1, 2, and 3, with "heads" replaced by "boys," determine the probability that both children are boys.
9 As a first approximation, assume that birthdays are uniformly distributed through a year of 365 days. Give an expression for the probability that \( n \) people sampled at random have different birthdays. Obtain a rough approximation for the probability. Find the smallest \( n \) for which the probability is less than 1/2.

10 The student organizer wants to choose a class representative for a class of 25. He thoroughly shuffles a deck of 52 cards and then proceeds through the class dealing one card to each student. The student receiving the first ace becomes the class representative. Give an expression for the probability that

(a) The first student becomes class representative on the first run through the class.

(b) The \( i \)th student in the class becomes class representative on the first run through the class.

(c) No class representative is chosen on the first run through the class.

(d) The first student becomes class representative.

11 In the game of craps, one of the players throws a pair of symmetric dice. On the first throw he wins if he gets 7 or 11 points and he loses if he gets 2, 3, or 12 points. Otherwise, he continues to toss until he obtains either the original number of points or 7 points. If he duplicates his original points, he wins; if he gets a 7, he loses. Find the probability that the player with the dice wins.

### 4.2

**CONDITIONAL PROBABILITY ON THE PLANE**

We have noted before the importance of real spaces for probability theory. In this section we examine conditional probability on the plane and more generally on real spaces.

#### A CONDITIONAL PROBABILITY FUNCTION

Consider a distribution for \((x_1, \ldots, x_k)\) on \(\mathbb{R}^k\) with probability function \(p\). And suppose that we want the conditional distribution given that a certain function \(Y\) takes the value \(y\). Then directly from Definition 4.1.1 we obtain

\[
(1) \quad p(x_1, \ldots, x_k : Y = y) = \frac{p(x_1, \ldots, x_k)}{p_Y(y)} \quad \text{if } Y(x_1, \ldots, x_k) = y
\]

and zero otherwise. Note that the denominator is purely a norming constant that scales up in value the original probabilities along the contour \(Y(x_1, \ldots, x_k) = y\); thus the relative conditional probabilities for points on the contour are just given by the original probability function. For example, in a safety investigation where \(x_i\) is the annual total number of accidents in plant \(i\), we could be interested in the conditional distribution of \((x_1, \ldots, x_4)\) given the annual total accidents \(y = x_1 + \cdots + x_4\) for a company.

#### EXAMPLE 1

**Conditioning two Poissons:** Consider a Poisson(\(\lambda_1\)) distribution for \(y_1\) and an independent Poisson(\(\lambda_2\)) distribution for \(y_2\). Then by Section 3.3E the joint probability function on the plane is given by
\[ \rho(y_1, y_2) = \frac{\lambda_1^{y_1}e^{-\lambda_1}}{y_1!} \frac{\lambda_2^{y_2}e^{-\lambda_2}}{y_2!} \quad \text{if } y_1, y_2 = 0, 1, 2, \ldots \]

and zero otherwise. And suppose that we want the conditional distribution given the value of the total \( T(y_1, y_2) = y_1 + y_2 = t \). The marginal probability function for \( T \) is available from formula (3.2.9),

\[
P_T(t) = \sum_{y_2} \frac{\lambda_1^{y_2}e^{-\lambda_1}}{y_2!} \frac{\lambda_2^{y_2}e^{-\lambda_2}}{y_2!} \frac{t!}{y_2!(t - y_2)!} \left( \frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^{y_2} \left( \frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^{t-y_2} = \frac{(\lambda_1 + \lambda_2)^t e^{-\lambda_1-\lambda_2}}{t!},
\]

where a sum of binomial probabilities [formula (3.5.1)] is 1; note that \( T \) is Poisson(\( \lambda_1 + \lambda_2 \)). Then by division using formula (1) we obtain the conditional distribution given the total \( y_1 + y_2 = t \):

\[
(2) \quad \rho(y_1, y_2; t) = \frac{\lambda_1^{y_1}e^{-\lambda_1} \lambda_2^{y_2}e^{-\lambda_2}/y_1!y_2!}{(\lambda_1 + \lambda_2)^t e^{-\lambda_1-\lambda_2}/t!} \quad \text{if } y_1, y_2 = 0, 1, 2, \ldots \quad \text{if } y_1 + y_2 = t,
\]

\[
= \left( \frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^{y_2} \left( \frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^{t-y_2} \quad y_2 = 0, 1, \ldots, t,
\]

and zero otherwise. Note that this distribution along a possible contour \( y_1 + y_2 = t \) is binomial(\( t, \lambda_2/(\lambda_1 + \lambda_2) \)) in terms of \( y_2 \), and similarly is binomial (\( t, \lambda_1/(\lambda_1 + \lambda_2) \)) in terms of \( y_1 \); see Figure 4.6. Thus if we are counting radioactive particles (with a Poisson distribution) over two time intervals and obtain a particular total count, then the events segregate between the two time intervals in accord

![Figure 4.6](image)

**FIGURE 4.6**

Poisson(\( \lambda_1 \)) for \( y_1 \); and independent Poisson(\( \lambda_2 \)) for \( y_2 \). The \( y_1 + y_2 = 3 \) section has the binomial(\( t, \lambda_1/(\lambda_1 + \lambda_2) \)) distribution for \( y_1 \). Illustration has \( \lambda_1 = \lambda_2 = 1 \).
with the binomial distribution; this conditional distribution can be useful for making statistical tests for the equality of $\lambda_1$ and $\lambda_2$ in an application.

Sometimes we are interested in very simple functions—coordinate projection functions.

Consider a discrete distribution on the plane $R^2$ with probability function $p$. And suppose that we want the conditional distribution for the second coordinate $y$ given the value of the first coordinate $x$. A condition that specifies the value of the first coordinate $x$ is a condition that specifies the value of the projection function onto the first axis; see Figure 4.7. Note that the second coordinate conveniently labels the points on the resulting $x$ section.

Consider a first-coordinate value $x$ having $p_1(x) \neq 0$. Then the conditional probability function for $y$ given $x$ is obtained directly from Definition 4.1.1:

$$p_2(y : x) = \frac{p(x, y)}{p_1(x)}$$

Note the form of $p_2(y : x)$: as a function of $y$ it comes entirely from the numerator $p(x, y)$; the denominator is present only as a norming constant, so summation over $y$ gives unity. For example, we might be interested in the distribution of the number of white blood cells on a slide given the number of red blood cells on the slide.

The definition (3) for $p_2(y : x)$ can be rearranged to give the multiplication formula

$$p(x, y) = p_1(x)p_2(y : x),$$

where the second factor is undefined for $x$ values having $p_1(x) = 0$.

The preceding multiplication formula, together with Exercise 3.3.2, proves the following proposition:

**PROPOSITION 1**

For the discrete case, $x$ and $y$ are statistically independent if and only if $p_2(y : x)$, where defined, is functionally independent of $x$.

Note the simple consequence: if $p_2(y : x)$ does not depend on $x$, then $p_2(y : x)$ is also the marginal probability function for $y$.

In a similar manner we can define the conditional probability function for $x$ given $y$:

$$p_1(x : y) = \frac{p(x, y)}{p_2(y)}$$

provided that $p_2(y) \neq 0$. And if the probability on the plane $R^2$ has come from some antecedent space, then the conditional probability function for $Y$ given $X$ can be presented as

$$p_{Y \mid X}(y : x) = \frac{p_{XY}(x, y)}{p_X(x)}$$

provided that $p_X(x) \neq 0$.

Analogous formulas exist for a distribution on $R^k$ with probability function $p$. For example, the conditional probability function for $(y_4, \ldots, y_k)$ given $(y_1, y_2, y_3)$ is
\[ p_{y_1 \ldots y_k | y_4, \ldots, y_k : y_1, y_2, y_3} = \frac{p(y_1, \ldots, y_k)}{p_{123}(y_1, y_2, y_3)} \]

provided that the denominator is nonzero. Note that the numerator gives the functional dependence and the denominator provides just the norming constant.

**EXAMPLE 2**

**Conditioning the multinomial:** First consider the multinomial Example 3.5.1: tossing a symmetric die with one face marked 1, two faces marked 2, and three faces marked 3; record \((y_1, y_2, y_3)\), where \(y_i\) is the number of \(i\) faces in 5 tosses. The probability function can be expressed in terms of \((y_1, y_2)\) since \(y_3 = 5 - y_1 - y_2\); the values of \(7776 p_{y_1 y_2}\) as retabulated from Example 3.5.1 are recorded in Table 4.1.

**TABLE 4.1**

<table>
<thead>
<tr>
<th>(y_2)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>243</td>
<td>405</td>
<td>270</td>
<td>90</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>810</td>
<td>1080</td>
<td>540</td>
<td>120</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1080</td>
<td>1080</td>
<td>360</td>
<td>40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>720</td>
<td>480</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>240</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ 1250 \]

**FIGURE 4.7**

Multinomial(5, 1/6, 2/6, 3/6). The marginal for \(y_1\) is binomial(5, 1/6). The conditional for \(y_2\) given \(y_1 = 2\) is binomial(3, 2/5).
Now suppose that we want the conditional distribution given that \( y_1 = 2 \). The marginal probability for \( y_1 = 2 \) is obtained by summing over \( y_2 \); we get 1250/7776. The conditional distribution for \( y_2 \), given \( y_1 = 2 \), is then obtained by division from values in the \( y_1 = 2 \) column; we obtain

<table>
<thead>
<tr>
<th>( y_2 )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_2(y_2:2) )</td>
<td>27</td>
<td>54</td>
<td>36</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>125</td>
<td>125</td>
<td>125</td>
<td>125</td>
</tr>
</tbody>
</table>

which is plotted in Figure 4.7; this is binomial(3, 2/5).

Now in general consider the multinomial(\( n, p_1, p_2, p_3 \)) distribution for \( (y_1, y_2, y_3) \); this is effectively a distribution for \( (y_1, y_2) \) in the plane, since \( y_3 = n - y_1 - y_2 \). The probability function is

\[
p(y_1, y_2) = \frac{n!}{y_1! \, y_2! \, (n - y_1 - y_2)!} \, p_1^{y_1} p_2^{y_2} p_3^{n-y_1-y_2}
\]

on the points \( y_i = 0, 1, \ldots \) with \( y_1 + y_2 \leq n \). The marginal probability function for \( y_1 \) can be obtained by summing out \( y_2 \):

\[
(7) \quad p_1(y_1) = \sum_{y_2} p(y_1, y_2)
\]

\[
= \frac{n!}{y_1! \, (n - y_1)!} \, p_1^{y_1} (p_2 + p_3)^{n-y_1}
\]

\[
\times \sum_{y_2} \frac{(n - y_1)!}{y_2! \, (n - y_1 - y_2)!} \, \left( \frac{p_2}{p_2 + p_3} \right)^{y_2} \left( \frac{p_3}{p_2 + p_3} \right)^{n-y_1-y_2}
\]

\[
= \binom{n}{y_1} p_1^{y_1} (p_2 + p_3)^{n-y_1}
\]

where a sum of binomial probabilities [formula (3.5.1)] is 1; note that the marginal distribution of \( y_1 \) is binomial(\( n, p_1 \)). The conditional distribution for \( y_2 \) given \( y_1 \) can then be calculated from formula (3):

\[
(8) \quad p_2(y_2 : y_1) = \frac{\binom{n}{y_1} p_1^{y_1} p_2^{y_2} p_3^{n-y_1-y_2}}{\binom{n}{y_1} p_1^{y_1} (p_2 + p_3)^{n-y_1}}
\]

\[
= \binom{n-y_1}{y_2} \left( \frac{p_2}{p_2 + p_3} \right)^{y_2} \left( \frac{p_3}{p_2 + p_3} \right)^{n-y_1-y_2}
\]

for \( y_2 = 0, 1, \ldots, n - y_1 \) and zero elsewhere; note that this conditional distribution for \( y_2 \) is binomial\( (n - y_1, p_2/(p_2 + p_3)) \). For \( n \) repetitions where \( y_1, y_2, \) and \( y_3 \) count the number of events \( A_1, A_2, \) and \( A_3 \), we see that with \( y_1 \) repetitions giving the first kind of event, there are \( n - y_1 \) repetitions that can be either the second or third kind of event; the formula above shows that the two kinds of events then
segregate according to the binomial distribution with probabilities $p_2/(p_2 + p_3)$ and $p_3/(p_2 + p_3)$ for the second and third kinds of events.

B CONDITIONAL PROBABILITY DENSITY GIVEN A PROJECTION

Consider an absolutely continuous distribution on the plane $R^2$ with probability density function $f$. And suppose that we are interested in the conditional distribution of the second coordinate $y$ given the value of the first coordinate $x$, for example, say, the distribution for the height of eldest son given the height of the father.

The probability for a particular value $x$ of the first coordinate is of course 0. Thus we cannot directly use the Definition 4.1.1 for conditional probability and must reconsider the selection interpretation discussed in Section 4.1B. If we consider the practical meaning of the first coordinate being equal to $x$ we are led to consider an interval $(x - h, x + h)$ and to contemplate the distribution for the second coordinate $y$ given that the first coordinate lies in the interval $(x - h, x + h)$; see Figure 4.8.

Let $G_h(y)$ be the distribution function for the conditional distribution of $y$ given that $x$ lies in the interval $(x - h, x + h)$. Then from Definition 4.1.1 we obtain

$$G_h(y) = \frac{\int_{x-h}^{x+h} \int_{-\infty}^{y} f(t, u) \, dt \, du}{\int_{x-h}^{x+h} \int_{-\infty}^{\infty} f(t, u) \, dt \, du} = \frac{\int_{x-h}^{x+h} \int_{-\infty}^{y} f(t, u) \, dt \, du/2h}{\int_{x-h}^{x+h} \int_{-\infty}^{\infty} f(t, u) \, dt \, du/2h}.$$

Now consider the limit as $h \to 0$. Both the numerator and denominator produce a derivative with respect to a limit of integration; we obtain

FIGURE 4.8
Joint density $f$; marginal density $f_1$ for $x$; and conditional density $f_2(t: x)$ for $y$ given $x$. 
\[ G(y) = \lim_{h \to 0} G_h(y) = \frac{\int_{-\infty}^{y} f(x, u) \, du}{\int_{-\infty}^{\infty} f(x, u) \, du} = \frac{\int_{-\infty}^{y} f(x, u) \, du}{f_1(x)} \]

This is undefined if \( f_1(x) = 0 \); and it is undefined more generally if the marginal density is not given by a limit (but such points form a set with length measure zero). Now by differentiation we obtain the conditional probability density function for \( y \) given \( x \):

\[ f_2(y: x) = \frac{f(x, y)}{f_1(x)}. \]

Note the form of \( f_2(y: x) \): as a function of \( y \) it comes entirely from the numerator \( f(x, y) \); the denominator is present only as a norming constant. Thus the form of the conditional distribution is given by the \( x \) section through the joint density function.

Formula (10) for \( f_2(y: x) \) can be rearranged to give the multiplication formula

\[ f(x, y) = f_1(x) f_2(y: x), \]

where the second factor may be undefined for points \( x \) in a set having probability zero.

The preceding multiplication formula, together with Exercise 3.3.3, proves the following proposition:

**PROPOSITION 2**

\( x \) and \( y \) are statistically independent if and only if \( f_2(y: x) \) is functionally independent of \( x \) (except possibly on a set having probability zero).

Note the simple consequence: if \( f_2(y: x) \) does not depend on \( x \), then \( f_2(y: x) \) is also the marginal density function for \( y \).

In a similar manner we can define the conditional probability density function for \( x \) given \( y \):

\[ f_1(x: y) = \frac{f(x, y)}{f_2(y)}. \]

And if the probability on \( \mathbb{R}^2 \) has come from some antecedent space, then the conditional probability density function for \( Y \) given \( X \) can be presented as

\[ f_{Y,X}(y: x) = \frac{f_{XY}(x, y)}{f_X(x)} \]

provided that \( f_X(x) \neq 0 \).

Analogous formulas exist for a distribution on \( \mathbb{R}^k \) with probability density function \( f \). For example, the conditional probability density function for \( (y_4, \ldots, y_k) \) given: \( (y_1, y_2, y_3) \) is

\[ f_{4, \ldots, k}(y_4, \ldots, y_k: y_1, y_2, y_3) = \frac{f(y_1, \ldots, y_k)}{f_{123}(y_1, y_2, y_3)}. \]

A conditional density function can be used for calculations. For example, to find the probability in a set \( A \) (Figure 4.9) we can write
(15) \[ P(A) = \int_{x_1}^{x_2} \left( \int_{y_1(x)}^{y_2(x)} f_2(y|x) \, dy \right) f_1(x) \, dx, \]

where the conditional probability
\[ \int_{y_1(x)}^{y_2(x)} f_2(y|x) \, dy \]
is multiplied by the marginal probability \( f_1(x) \, dx \) and integrated. Note that the expression (15) is just a rearrangement of the standard expression
\[ P(A) = \int_{x_1}^{x_2} \int_{y_1(x)}^{y_2(x)} f(x, y) \, dx \, dy. \]

Perhaps the most convenient way to remember the conditional density is as part of a probability differential
(16) \[ f(x, y) \, dx \, dy = f_1(x) \, dx \cdot f_2(y|x) \, dy, \]
in which the probability \( f_1(x) \, dx \) for an increment for the first coordinate is multiplied by the conditional probability \( f_2(y|x) \, dy \) for an increment for the second coordinate.

**EXAMPLE 3**

**Bivariate distribution:** Consider further the bivariate Example 3.2.6 with density function
\[ f(x, y) = 6x \quad \text{if} \quad 0 < x < 1, \quad 0 < y < 1 - x, \]
and zero otherwise. The marginal density for \( x \) was calculated in Example 3.2.6:
\[ f_1(x) = 6x(1 - x) \quad \text{if} \quad 0 < x < 1 \]
and zero otherwise. The conditional density for \( y \) can then be obtained from formula (10):
(17) \[ f_2(y|x) = (1 - x)^{-1} \quad \text{if} \quad 0 < y < 1 - x \]
and zero otherwise. These are sketched in Figure 4.10.
FIGURE 4.10
Conditional density \( f_2(y : x) \) for \( y \) given \( x \).

Note that the initial density \( 6x \) apparently factors into a function \( 6x \) of the first variable and a function \( 1 \) of the second variable. It is not, however, a factorization of the function \( f \) which is defined on \( \mathbb{R}^2 \); in particular, the region of positive density is not a product set. Thus \( x \) and \( y \) are not statistically independent; certainly this is evident from the conditional density \( f_2(y : x) \), which is uniform \((0, 1 - x)\) and does depend on \( x \).

The marginal density for \( y \) was calculated in Example 3.2.6:

\[
f_2(y) = 3(1 - y)^2 \quad \text{if} \quad 0 < y < 1
\]

and zero otherwise. The conditional density for \( x \) can then be obtained from formula (10):

\[
f_1(x : y) = \frac{2x}{(1 - y)^2} \quad \text{if} \quad 0 < x < 1 - y
\]

and zero otherwise; this is a triangular distribution of \((0, 1 - y)\).

---

**C CONDITIONAL PROBABILITY DENSITY GIVEN A GENERAL FUNCTION**

Consider a distribution for \((x_1, \ldots, x_k)\) on \( \mathbb{R}^k \) with density function \( g \). And suppose that we are interested in the conditional distribution given that a certain function \( Y(x_1, \ldots, x_k) \) has the value \( y \); for example, the conditional distribution of the yields \( x_1, \ldots, x_7 \) on plots \( 1, \ldots, 7 \) given the total yield \( x_1 + \cdots + x_7 \) for the seven plots. The basic procedure is to make a change of variable on \( \mathbb{R}^k \) so as to separate
Y(x1, . . . , xk) as one of the new variables and then to use the results appropriate to projections from Section B. We examine the details for the case k = 2.

Consider a distribution for (x1, x2) on R^2 with density function g. And suppose that we want the conditional distribution given a value for some function Y1. For most reasonable functions Y1, we can find a complementing function Y2 such that the mapping

(\(y_1, y_2\) = (Y1(x1, x2), Y2(x1, x2))

is one to one and continuously differentiable each way; let

\((x_1, x_2) = (X_1(y_1, y_2), X_2(y_1, y_2))\)

designate the inverse mapping; see Figure 3.10.

In terms of differentials we can present the distribution for the original variables \((x_1, x_2)\) and for the new variables \((y_1, y_2)\):

\[ g(x_1, x_2) \, dx_1 \, dx_2 = g(X_1(y_1, y_2), X_2(y_1, y_2)) \left| \frac{\partial X_1, X_2}{\partial y_1, y_2} \right| \, dy_1 \, dy_2. \]

The conditional density for \(y_2\) given \(y_1\) is then available from formula (10) for density given a projection:

\[
\begin{equation}
\int_{-\infty}^{\infty} g(X_1(y_1, u), X_2(y_1, u)) \left| \frac{\partial X_1, X_2}{\partial y_1, u} \right| \, du.
\end{equation}
\]

Note that \(y_2\) labels the points along the \(y_1\) contour on the original space; and note that

\[
\left| \frac{\partial X_1, X_2}{\partial y_1, y_2} \right|
\]

allows for variation in the width between the contours \(y_1\) and \(y_1 + dy_1\).

**EXAMPLE 4**

**Symmetric normal distribution on R^2:** Consider a sample \((x_1, x_2)\) from the central normal(0, \(\sigma\)) distribution, and suppose that we want the conditional distribution given the sample sum \(t = T(x_1, x_2) = x_1 + x_2\). Some preimage contours for the sum function are indicated in Figure 4.11; note that they are lines orthogonal to the 1-vector \(\mathbf{1} = (1, 1)'\). Thus we are interested in the conditional distribution along a strip such as that indicated in the figure.

For our basic procedure we need a second function whose preimage contours are cross sectional to those just described. A convenient set of contours is given by the lines parallel to the 1-vector and thus orthogonal to the first set of lines; a function that describes these contours is the difference function \(x_2 - x_1\).

We now choose simple coordinates that measure Euclidean distance along the two sets of contours:

\[
y_1 = \frac{x_1}{\sqrt{2}} + \frac{x_2}{\sqrt{2}},
\]

\[
y_2 = -\frac{x_1}{\sqrt{2}} + \frac{x_2}{\sqrt{2}}.
\]
FIGURE 4.11
Preimage contours for the sum function; and preimage contours for the difference function.

or, in matrix form,

\[
\begin{pmatrix}
  y_1 \\
  y_2
\end{pmatrix}
= \begin{pmatrix}
  \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\
  -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}
\end{pmatrix}
\begin{pmatrix}
  x_1 \\
  x_2
\end{pmatrix}
\quad \text{or} \quad y = Ax.
\]

This transformation is an orthogonal transformation. Note that the row vectors in
\( A \) have unit length and are mutually orthogonal; that is, the inner product matrix
\( AA' = I \) is the identity matrix; for details see Supplement Section 2.5. (In general,
a square matrix \( A \) is called orthogonal if it has the property \( AA' = I \) or \( A'A = I \),
and correspondingly the transformation \( y = Ax \) is called "orthogonal"; note that
such a transformation does not change the sum of squares

\[
\sum y_i^2 = y'y = x'A'Ax = x'x = \sum x_i^2;
\]
thus Euclidean distance is preserved.) The Jacobian determinant is

\[
\left| \frac{\partial y}{\partial x} \right| = \begin{vmatrix}
  \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\
  -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}
\end{vmatrix} = 1.
\]

The initial distribution can now be expressed in terms of the new variables

\[
\frac{1}{2\pi\sigma^2} \exp \left\{ -\frac{x_1^2 + x_2^2}{2\sigma^2} \right\} dx_1 \, dx_2 = \frac{1}{2\pi\sigma^2} \exp \left\{ -\frac{y_1^2 + y_2^2}{2\sigma^2} \right\} \cdot 1 \cdot dy_1 \, dy_2;
\]

hence the distribution of \((y_1, y_2)\) is that of a sample of 2 from the central normal(0, \(\sigma\)) distribution. Thus the marginal distribution of \(y_1\) is normal(0, \(\sigma\)), and the conditional of \(y_2\) given \(y_1\) is normal(0, \(\sigma\)). In conclusion, note that the conditional distribution of \((x_1, x_2) = (y_1/\sqrt{2} - y_2/\sqrt{2}, y_1/\sqrt{2} + y_2/\sqrt{2})\) is normal on the
contour with the given value for \(x_1 + x_2 = \sqrt{2}y_1\); it is located at \((y_1/\sqrt{2}, y_1/\sqrt{2})\)
and has scaling \(\sigma\).
**D EXERCISES: PROBABILITY FUNCTIONS**

1. Consider the multinomial \( (5, 1/6, 2/6, 3/6) \) for \( (y_1, y_2) \) as tabulated in Example 2.
   (a) Tabulate the conditional distribution of \( y_2 \) given \( y_1 = 1 \); verify that it is binomial \((4, 2/5)\).
   (b) Tabulate the conditional distribution of \( y_2 \) given \( y_1 = 0 \); verify that it is binomial \((5, 2/5)\).

2. Automatic transmissions received for repair are unrepairable with probability \(1/6\), repairable with probability \(2/6\), and actually in satisfactory condition with probability \(3/6\); see Example 2. Unrepairable units are quickly identified. A consignment of 5 units contains 1 unrepairable unit; determine the distribution of \( y_2 \), the number of units that are repairable.

3. Consider the multivariate hypergeometric \((6, 3, 1/6, 2/6, 3/6)\) in Example 3.5.3; tabulate the probability function in terms of \((y_1, y_2)\), where \( y_3 = 3 - y_1 - y_2 \).
   (a) Tabulate the conditional distribution of \( y_2 \) given \( y_1 = 1 \); verify that it is hypergeometric \((5, 2, 2/5)\).
   (b) Tabulate the conditional distribution of \( y_2 \) given \( y_1 = 0 \); verify that it is hypergeometric \((5, 3, 2/5)\).

4. **Sample of outboard motors** (Example 3.5.3). The dealer reports that the sample of 3 contains 1 unrepairable unit; determine the distribution of \( y_2 \), the number of repairable units.

5. **Industrial acceptance sampling** (Example 4.1.6). A lot contains \( N = 10 \) articles of which \( D = 3 \) are defective and \( N - D = 7 \) are good. The articles are tested successively. Verify the expression
   \[
   \binom{4}{2} \frac{3^{12} 7^{10}}{10^{15}}
   \]
   for the probability that the last defective is found on the fifth test by considering possible sequences of 5 articles for testing.

6. (continuation) Verify the expression
   \[
   \begin{pmatrix} 3 & 0 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 7 \\ 2 & 0 & 5 \end{pmatrix} \begin{pmatrix} 10 \\ 4 & 1 & 5 \end{pmatrix}
   \]
   for the probability just described by considering a sequence of 3 sets of size 4, 1, 5.

**E PROBLEMS: PROBABILITY FUNCTIONS**

7. Consider independent \( y_1, \ldots, y_k \), where \( y_i \) is Poisson(\( \lambda_i \)). Show that the conditional distribution of \((y_1, \ldots, y_k)\) given the total \( T(y_1, \ldots, y_k) = y_1 + \ldots + y_k = \gamma \) is multinomial(\( \gamma, \lambda_1/\gamma, \ldots, \lambda_k/\gamma \)), where \( \lambda = \lambda_1 + \ldots + \lambda_k \). Use Problem 3.5.34.

8. Let \((y_1, \ldots, y_k)\) be multinomial \((n, p_1, \ldots, p_k)\). Show that the conditional distribution of \((y_{r+1}, \ldots, y_k)\) given \((y_1, \ldots, y_r)\) is multinomial\((n - y_1 - \ldots - y_r, p_{r+1}/p, \ldots, p_k/p)\), where \( p = p_{r+1} + \ldots + p_k \). Use Problem 3.5.38.

9. Let \((y_1, \ldots, y_k)\) be multivariate hypergeometric\((N, n, p_1, \ldots, p_k)\). Show that the conditional distribution of \((y_{r+1}, \ldots, y_k)\) given \((y_1, \ldots, y_r)\) is multivariate hypergeometric\((Np, n - y_1 - \ldots - y_r, p_{r+1}/p, \ldots, p_k/p)\), where \( p = p_{r+1} + \ldots + p_k \). Use Problem 3.5.41.

**F EXERCISES: PROBABILITY DENSITY FUNCTIONS**

10. Consider a distribution on \( \mathbb{R}^2 \) with density
   \[
   f(y_1, y_2) = 18e^{-6y_1-3y_2}, \quad y_1 > 0, y_2 > 0
   \]
and zero otherwise. Determine the conditional density function for $y_2$ given $y_1$; see Exercise 3.2.16.

11 Consider a distribution on $R^2$ with density
\[ f(y_1, y_2) = 2e^{-y_1-y_2}, \quad 0 < y_1 < y_2 < \infty, \]
and zero otherwise. Determine the conditional density function for $y_2$ given $y_1$; see Exercise 3.2.17.

12 Let $f$ be the density function for the uniform distribution over the disk $y_1^2 + y_2^2 < 1$. Calculate the conditional density $f_2(\cdot | y_1)$ for $y_2$ given $y_1$; see Exercise 3.2.18.

13 Consider the uniform distribution over the unit ball $y_1^2 + y_2^2 + y_3^2 < 1$. Determine the conditional distribution of $y_3$ given $(y_1, y_2)$; and the conditional distribution of $(y_2, y_3)$ given $y_1$; see Exercise 3.2.19.

14 Consider the density on $R^k$ given by $f(y_1, \ldots , y_k) = h_1(y_1) \cdots h_k(y_k)$, where each $h_i$ is a density on $R$. Determine the conditional density of $(y_3, \ldots , y_k)$ given $(y_1, y_2, y_3)$; see Exercise 3.2.20.

15 Consider the Pareto($\alpha$) distribution on $R^2$ with density $f$ given by ($\alpha > 0$)
\[ f(y_1, y_2) = \alpha(\alpha + 1)(1 + y_1 + y_2)^{-\alpha-2}, \quad 0 < y_1, 0 < y_2, \]
and zero otherwise. Determine the conditional distribution of $y_2$ given $y_1$; see Exercise 3.2.21.

G PROBLEMS: PROBABILITY DENSITY FUNCTIONS

16 Consider the canonical Student($n$) distribution for $(y_1, y_2)$ on $R^2$. Determine the conditional density for $y_2$ given $y_1$; see Exercise 3.2.22. Verify that the conditional distribution can be described as $y_2/(1 + y_1^2)^{1/2}$ is canonical Student($n + 1$). Note that the conditional distribution of $y_2$ is more spread out with larger values of $|y_1|$.

17 Consider the uniform distribution for $(x_1, x_2)$ over the unit square $(0, 1)^2$. Determine the distribution of $(y_1, y_2)$ where $y_1 = x_1 + x_2, y_2 = x_2$ and sketch the region of positive density. Determine the conditional distribution of $x_2$ given $x_1 + x_2 = t$.

18 (continuation of Example 3.2.8) Consider a distribution for $(x_1, x_2)$ with density $g$. Determine the conditional distribution of $x_2$ given $t = x_1 + x_2$.

19 Consider the Pareto($\alpha$) distribution for $(y_1, \ldots , y_k)$ on $R^k$; see Problem 2.3.14. Show that the conditional distribution given $y_1$ can be described as: $(y_2/(1 + y_1), \ldots , y_k/(1 + y_1))$ is Pareto($\alpha + 1$); see Problem 3.2.24.

20 Consider the canonical Student($n$) distribution for $(y_1, \ldots , y_k)$. Show that the conditional distribution given $y_1$ can be described as: $(y_2/(1 + y_1^2)^{1/2}, \ldots , y_k/(1 + y_1^2)^{1/2})$ is canonical Student($n + 1$); see Problem 3.2.25.

21 Consider the Dirichlet($r_1, \ldots , r_{k+1}$) distribution for $(y_1, \ldots , y_k)$. Show that the conditional distribution given $(y_1, \ldots , y_p)$ can be described as: $(y_{p+1}/Y, \ldots , y_k/Y)$ is Dirichlet($r_{p+1}, \ldots , r_{k+1}$), where $Y = 1 - y_1 - \cdots - y_p$; see Problem 3.2.26.

22 Consider a distribution for $(x, y)$ on $R^2$ with density $f$. Determine the conditional density for $y$ given $x = x - y$; see Problem 3.2.29.

23 Consider a distribution for $(x, y)$ on $R^2$ with density $f$, where $f(x, y) = 0$ for $x < 0$ and for $y < 0$. Determine the conditional density for $t = x + y$ given $v = x/y$; see Problem 3.2.30.

24 A distribution for lifetime $t$ on the interval $(0, \infty)$ has a constant failure rate:
\[ \frac{f(t)}{1 - F(t)} = c; \]

(note that there is no "aging." Show that $t$ is exponential (1/c).

H PROBLEMS: ACCEPTANCE SAMPLING

A lot contains $N$ articles, of which $D = Np$ are defective and $N - D = Nq$ are good. The articles are thoroughly mixed and are sampled and tested one by one until $n$ have been tested. Let
Let $r$ be the number of defectives in the sample and $p_1$ be the probability function. From Section 3.5C we have

$$p_1(r) = \binom{n}{r} \frac{(D)(N-D)}{(n-r)} \frac{(N)}{(n)}.$$

25 Verify the following expression directly; see Example 4.1.6.

$$p_1(r) = \binom{n}{r} \frac{D^r(N-D)^{n-r}}{N^n}.$$

26 (continuation) Verify the following expression directly:

$$p_1(r) = \binom{n}{r} \frac{(N-n)}{(D-r)} \frac{D!(N-D)!}{D!(N-D-n)!}.$$

27 (continuation) Verify the following expression directly:

$$p_1(r) = \binom{n}{r} \frac{D^{D-r}(N-D)^{N-n-D+n}}{N^{N-n}}.$$

28 (continuation) Verify the following expression directly:

$$p_1(r) = \binom{r}{D-r} \binom{n-r}{N-D} \binom{n}{n} \binom{N}{N-n}.$$

Consider a sequence of two sets of size $n, N-n$.

Alternatively, the articles are thoroughly mixed and are sampled and tested one by one until $r$ defectives have been obtained. Let $n$ be the number of items tested and $p_2$ be the probability function.

29 In the pattern of Example 4.1.6, verify the following expression:

$$p_2(n) = \binom{n}{r} \frac{D!}{(N-n)!} \frac{(N-D)^{n-r}}{N-n+1}.$$

30 (continuation) Verify the following expression directly:

$$p_2(n) = \binom{n}{r} \frac{(N-n)}{(D-r)} \frac{D!(N-D)!}{N^{N-n}}.$$

Consider a sequence of $N$ articles for testing.

31 Verify the following expressions directly:

$$p_2(n) = \binom{n}{r} \frac{D!}{(N-n)!} \frac{(N-D)^{n-r}}{N!} = \binom{n}{r} \frac{D^{D-r}(N-D)^{N-n-D+n}}{N^{N-n+1}}.$$

32 Verify the following expressions directly:

$$p_2(n) = \binom{n}{r} \frac{D!(N-D)!}{D!(N-D-n+r)!} \frac{(N-n)}{(n-r)} \frac{(N-D-n+r)}{N-n+1}.$$
\[
\frac{D}{(r-1)} \binom{D-r+1}{1} \binom{N-D}{n-r} = \frac{D}{(D-r)} \binom{r}{1} \binom{N-D-n+r}{N-n}.
\]

I  PROBLEMS: COMPUTER USAGE

The elements in a finite population have serial numbers 101, 102, \ldots, 685. A random sample of 25 can be obtained from a computer that produces random digits: the digits are taken in blocks of three; if a three-digit block is a serial number and has not been obtained previously, it is accepted; continue until 25 serial numbers have been accepted. This process is justified on the basis of conditional probability and symmetry.

33 Engines produced by an automobile plant in a particular year have serial numbers: 474400001, 474400002, \ldots, 474438756. Use a random-digit generator to obtain a random sample of 30.

34 The students in a freshman class have student numbers A0001, A0002, \ldots, A2387. Use a random-digit generator to obtain a random sample of 30.

Consider a density function \( f \) on \( R \) which is zero outside the interval \((a, b)\) and has maximum value \( M \). Let \( v_1 \) and \( v_2 \) be independent, \( v_1 \) uniform\((a, b)\), and \( v_2 \) uniform\((0, M)\). Sample values for \((v_1, v_2)\) can be obtained by relocating and rescaling values from the uniform\((0, 1)\); these, in turn, can be approximated by taking batches of random digits. A sample value of \((v_1, v_2)\) is tested to see if \( v_2 \leq K(v_1) \); if so, \( v_1 \) is accepted; if not, \( v_1 \) is rejected. Conditional probability shows that this produces sample values from the distribution with density \( f \).

35 Use random digits to obtain a sample of 25 from the triangular distribution with \( f(x) = 2x \) on \((0, 1)\) and zero otherwise.

36 Use random digits to obtain a sample of 25 from the beta\(3, 3\) distribution.

J  PROBLEMS: NORMAL DISTRIBUTION

37 Consider the bivariate normal\([\mu_1, \mu_2, \sigma_1, \sigma_2, \rho]\) density \( f \) for \((y_1, y_2)\) in Problem 2.4.24. Let \( h_1 \) be the normal\([\mu_1, \sigma_1]\) density and \( h_2((y_1) \) be the normal\([\mu_2 + \rho \sigma_2 y_1 - \mu_1)/\sigma_1, \sigma_2(1 - \rho^2)^{1/2}\] density. Show that

\[ f(y_1, y_2) = h_1(y_1)h_2(y_2; y_1) \]

and deduce that the marginal distribution of \( y_1 \) is normal\([\mu_1, \sigma_1]\) and the conditional of \( y_2 \) is normal\([\mu_2 + \rho \sigma_2 y_1 - \mu_1)/\sigma_1, \sigma_2(1 - \rho^2)^{1/2}\].

38 Consider the multivariate normal\([\mu; \Sigma]\) distribution for \( y \) as given in Section 2.4G and with new notation in Problem 3.6.9. Verify the following factorization of the density function,

\[
f(y) = \frac{[\Sigma^{-1}]^{1/2}}{(2\pi)^{k/2}} \exp \left\{ -\frac{1}{2} (y_1 - \mu_1)^{\Sigma^{-1}} (y_1 - \mu_1) \right\} \\
\frac{[\Sigma^{22}]^{1/2}}{(2\pi)^{k/2}} \exp \left\{ -\frac{1}{2} (y_2 - \mu_2 - \Sigma^{21} \Sigma^{-1} (y_1 - \mu_1) )^{\Sigma^{22}} (y_2 - \mu_2 - \Sigma^{21} \Sigma^{-1} (y_1 - \mu_1)) \right\}
\]

by using matrix identities derived from \( \Sigma^{\Sigma^{-1}} = I \), such as \( \Sigma^{-1} - \Sigma^{21} \Sigma^{22} - \Sigma^{21} = \Sigma^{-1} \).

Deduce that the marginal distribution of \( y_1 \) is normal\([\mu_1; \Sigma_{11}]\), and that the conditional distribution of \( y_2 \) given \( y_1 \) is normal\([\mu_2 + \Sigma^{21} \Sigma^{-1} (y_1 - \mu_1); (\Sigma^{22})^{-1}]\).
SUPPLEMENTARY MATERIAL

Supplement Section 4.3 is on pages 537–544.
Some basic notation for stochastic processes is presented, and a brief introduction is given for Poisson processes, Markov processes, and normal processes. The Polya urn, which has applications in biology and medicine, is also described.

NOTES AND REFERENCES

In this chapter we have given detailed attention to the concept of conditional probability. Conditional probability is usually examined for an isolated probability $P(A; C)$. By contrast we have emphasized the measure $P(\cdot; C)$ or $P_C$: the measure gives a full probability description of a system performance on the basis of information represented by $C$. We have also provided a selection model for the interpretation of this conditional probability measure. Conditional probability is a very fundamental concept for applications.

For some further reading on the supplemental topic of stochastic processes, see the following texts:

We have now examined the basic methods for handling probabilities themselves. In Chapter 1 we introduced the probability model and in Chapter 2 its detailed form on the line and plane. Then in Chapter 3 we examined marginal probability for functions of initial responses, and independence for combining responses. And in Chapter 4 we discussed conditional probability for the case of partial information concerning a response. A model of course has probabilities, but some other characteristics can also be very important.

To investigate the yield of a new variety of wheat, an agricultural statistician will collect data from a moderately large number of plots and will average the values as an estimate of the mean, or long-run average yield. A chemist investigating a new process for producing an isotope will examine a moderately large number of runs on the process and will average the values as an estimate of the mean, or long-run average yield.

In Chapter 5 we discuss the mean value for a probability distribution on the real line. Then by examining the mean for functions on the line and plane we obtain some powerful tools for describing location, scaling, and other properties of distributions.
5

MEAN VALUE FOR REAL AND VECTOR DISTRIBUTIONS

5.1

DEFINITION OF THE MEAN

In this chapter we discuss the mean value—the concept itself and its use to describe and predict important characteristics of probability distributions.

In this Section 5.1 we introduce the mean value and record some simple properties. In Section 5.2 we use the concept to describe the location of a distribution on the real line, and then, by simple vector extension, the location of a distribution on the plane. In Section 5.3 we use it to describe the scaling of a distribution on the line—from very compact through to very diffuse—and then by vector extension the scaling of a distribution on the plane. Matrix notation carries the initial cases on the real line into the vector extension in a simple routine way.

In Section 5.4 we examine location and scaling for a sample from a distribution and we derive important formulas that predict these characteristics from corresponding characteristics of the initial distribution. Again matrix notation provides the vector extension immediately from the initial case on the real line.

Conditional characteristics are examined in Section 5.5, moment sequences in Section 5.6, and moment-generating functions in Section 5.7. These provide a powerful collection of methods for predicting characteristics of various probability distributions.

A DEFINITION

Consider the rolling of a symmetric die and the recording of the number \( y \) of points showing. First suppose that we have found someone who will give us $1 for each point showing. If this game were played a large number of times, say \( N \), then \( y = 1 \) would occur approximately \( N/6 \) times, \( \ldots \), \( y = 6 \) would occur approximately \( N/6 \) times. The total amount received would be approximately

\[
1 \cdot (1/6)N + 2 \cdot (1/6)N + \cdots + 6 \cdot (1/6)N
\]

dollars. Thus the mean or long-run average per play would be

\[
E(y) = 1 \cdot (1/6) + 2 \cdot (1/6) + \cdots + 6 \cdot (1/6) = 3\frac{1}{2}
\]
dollars. Alternatively, suppose that we have found someone who will give us \( y^2 \) dollars on any toss; then, in a similar way, the mean or long-run average per play would be

\[
E(\bar{y}^2) = 1^2 \cdot (1/6) + 2^2 \cdot (1/6) + \ldots + 6^2 \cdot (1/6) = 15\frac{1}{2}.
\]

Now, more generally, consider a random system with sample space \( \mathcal{S} \), with events \( \mathcal{A} \), and with probability measure \( P \). And suppose that we are interested in a real-valued function \( Y \) giving the derived response \( y = Y(s) \) from the basic response \( s \). For example, we might be interested in the mean power \( Y \) in an electronic system in a case where \( Y = x_1^2 + \ldots + x_{10}^2 \) in terms of underlying variables \( x_1, \ldots, x_{10} \).

First suppose that \( Y \) is a simple function, a function taking only a finite number of values; let \( y_1, \ldots, y_m \) be the values. Then, in direct analogy to the preceding example, we define the mean value of \( Y \) written variously as

\[
E(Y) = E(Y(s)) = E(y) = \mu_Y
\]

by the summation

\[
E(Y) = \sum_{i=1}^{m} y_i p_Y(y_i) = \sum_{i=1}^{m} y_i p([s : Y(s) = y_i]).
\]

This is the summation of values \( y_i \) for the function weighted by corresponding probabilities. Note that we have used several different notations for the mean, each being convenient in appropriate contexts; the Greek letter \( \mu \) is often used for mean and a subscript indicates the distribution being examined.

Now suppose that \( Y \) is a bounded real-valued function: \(-M_1 < Y(s) < M_2\) for all \( s \) in \( \mathcal{S} \); see Figure 5.1. And suppose that we subdivide the range \((-M_1, M_2)\) by points \(-M_1 = a_0 < a_1 < \ldots < a_n = M_2\) such that all the differences \( a_i - a_{i-1}\) are less than some small number \( \epsilon \). Then we can approximate \( Y \) from above by the simple function

\[
Y_n = \sum_{j=1}^{n} a_j I_{A_j},
\]

**FIGURE 5.1**

Set \( A_j \), on which the function takes values in \([a_{j-1}, a_j]\): the function \( Y_n \) approximates \( Y \) from above; the function \( X_n \) approximates \( Y \) from below.
where $A_j = Y^{-1}([a_{j-1}, a_j])$ is the preimage of the interval $[a_{j-1}, a_j]$ and $I_{A_j}$ is the corresponding indicator function; see Figure 5.1. This simple function has mean value

$$E(Y_n) = \sum_{j=1}^{n} a_j P(A_j).$$

Similarly, we can approximate $Y$ from below by the simple function

$$X_n = \sum_{j=1}^{n} a_{j-1} I_{A_j},$$

which has mean value

$$E(X_n) = \sum_{j=1}^{n} a_{j-1} P(A_j).$$

If we take a sequence of values of $\epsilon$ going to zero, then the simple functions $Y_n$ and $X_n$ converge to $Y$ from above and below. Consider how the simple functions differ in their mean values:

$$E(Y_n) - E(X_n) \leq \sum_{j=1}^{n} \epsilon P(A_j) = \epsilon,$$

thus they can be at most $\epsilon$ apart. Accordingly, we define $E(Y)$ to be the limiting value as the upper approximation $Y_n$ and lower approximation $X_n$ converge to $Y$:

**DEFINITION 1**

The mean value $E(Y)$ of a bounded function $Y$ is defined by

$$E(Y) = \lim_{\epsilon \to 0} \sum_{j=1}^{n} a_j P(A_j) = \lim_{\epsilon \to 0} \sum_{j=1}^{n} a_{j-1} P(A_j),$$

and denoted by

$$E(Y) = \int Y(s) \, dP(s),$$

where the maximum subdivision interval $\epsilon$ goes to zero.

Now suppose that $Y$ is an arbitrary real-valued function. We obtain a bounded function $Y_{M_1, M_2}$ by flattening $Y$ off at an upper limit $M_2$ and at a lower limit $-M_1$:

$$Y_{M_1, M_2}(s) = M_2 \quad \text{if} \quad M_2 \leq Y(s) < \infty,$$

$$= Y(s) \quad -M_1 < Y(s) < M_2,$$

$$= -M_1 \quad -\infty < Y(s) \leq -M_1.$$

The mean value of $Y_{M_1, M_2}$ is available from Definition 1: we obtain the mean value of $Y$ by taking $M_1$ and $M_2$ to $\infty$:
**DEFINITION 2**

The **mean value** \( E(Y) \) of a function \( Y \) is defined by

\[
E(Y) = \lim_{M_1 \to -\infty} \lim_{M_2 \to -\infty} E(Y_{M_1M_2})
\]

and written

\[
E(Y) = \int Y(s) \, dP(s).
\]

provided the limit exists as \( M_1 \) and \( M_2 \) separately approach \( \infty \); otherwise, we say that \( E(Y) \) does not exist or \( E(Y) \) diverges.

We, of course, assume that any function \( Y \) respects our events and is a measurable function in accord with Definition 3.1.2. Some theory concerning the preceding general definition of an integral will be discussed in Supplement Section 5.8.

**B  CALCULATIONS**

First consider the case of a discrete function \( Y \) or even a more general function \( Y \) provided its distribution is discrete: let \( y_1, y_2, \ldots \) be the points where there is probability. Then directly from Definition 2 we obtain the mean value of \( Y \) as

\[
E(Y) = E(Y(s)) = \sum y_i P(Y^{-1}(y_i))
\]

provided that the summation converges absolutely; otherwise, we say that \( E(Y) \) does not exist or \( E(Y) \) diverges.

**EXAMPLE 1**

Consider 4 tosses of a symmetric coin giving \((x_1, \ldots, x_4)\), where \(x_i\) is the indicator for heads on the \(i\)th toss. And let \( Y = Y(x_1, \ldots, x_4) \) be the number of heads. By Section 3.5A the response \( Y \) is binomial(4, 1/2) with distribution

<table>
<thead>
<tr>
<th>( y )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_Y(y) )</td>
<td>1/16</td>
<td>4/16</td>
<td>6/16</td>
<td>4/16</td>
<td>1/16</td>
</tr>
</tbody>
</table>

The mean value of \( Y \) is

\[
E(Y) = \sum y p_Y(y)
\]

\[
= 0 \cdot (1/16) + 1 \cdot (4/16) + 2 \cdot (6/16) + 3 \cdot (4/16) + 4 \cdot (1/16)
\]

\[
= 2.
\]

Thus with four tosses of a symmetric coin the mean number of heads is 2.
Now consider the case of a function $Y$ defined on $R^1$ or $R^k$ and suppose that the distribution on $R^1$ or $R^k$ has density function $f$. Then Definition 2 (with Proposition 7 of Supplement Section 5.8) gives the mean value of $Y$ as

$$E(Y) = E(Y(x_1, \ldots, x_k)) = \int_{R^k} Y(x_1, \ldots, x_k) f(x_1, \ldots, x_k) \, dx_1 \cdots dx_k$$

provided that the integral converges absolutely; otherwise, we say that $E(Y)$ does not exist or that $E(Y)$ diverges.

EXAMPLE 2

Minimum lifetime: Consider an electronic control and suppose that the engineer has found that the lifetime $y$ of the control is exponential(800). We evaluate the mean lifetime:

$$E(y) = \int_0^\infty y (1/800) e^{-y/800} \, dy$$

$$= 800 \int_0^\infty ze^{-z} \, dz, \quad \text{where } z = y/800$$

$$= 800 \left( -ze^{-z} \bigg|_0^\infty + \int_0^\infty e^{-z} \, dz \right)$$

$$= 800;$$

the mean lifetime is 800.

Now briefly consider the general case of a real-valued function $Y$. Definition 2 gives the mean value of $Y$ as

$$E(Y) = E(Y(s)) = \int_S Y(s) \, dP(s)$$

in terms of the underlying distribution on $S$. Alternatively, formula (20) of Supplement Section 5.8 gives the mean value of $Y$ as

$$E(Y) = E(Y(s)) = \int_S Y \, dP_Y(y) = \int_S y \, dF_Y(y)$$

in terms of the marginal distribution of $Y$. Formulas (8) and (9) apply provided that the integral exists; otherwise, $E(Y)$ does not exist or diverges.

All our examples, however, will involve just the summations (6) or the ordinary integrals (7); some theory for the general case is recorded in Supplement Section 5.8.

Now, more generally, consider a vector-valued function $(Y_1, \ldots, Y_k)$ taking values in $R^k$. Then we define the mean vector to be

$$E(Y_1, \ldots, Y_k) = (E(Y_1), \ldots, E(Y_k)),$$

provided that each coordinate exists.
C PROPERTIES

We now record some properties of the mean value for functions defined on a space $S$ with probability $P$; these important properties will be used freely in succeeding sections.

1. Normed: The mean value of the constant function $c$ is equal to $c$:

$$E(c) = c.$$  

The proof follows trivially by noting that the preimage of the value $c$ is, of course, the whole sample space $S$.

The remaining properties follow directly from propositions in Supplement Section 5.8; for each property the needed proposition is cited.

2. Linear: If $E(Y_1)$ and $E(Y_2)$ exist, then

$$E(aY_1 + bY_2) = aE(Y_1) + bE(Y_2)$$

for real values $a$ and $b$: mean linear combinations equals linear combination means (Proposition 5.8.3).

3. Continuity: If $Y_1, Y_2, \ldots$ is a sequence of functions such that $\lim_{i \to \infty} Y_i(s)$ exists and if $|Y_i(s)| < M(s)$ where $E(M)$ exists, then (Proposition 5.8.8)

$$\lim_{i \to \infty} E(Y_i) = E \left( \lim_{i \to \infty} Y_i \right).$$

4. Independence: If $Y_1$ and $Y_2$ are statistically independent, then

$$E(Y_1Y_2) = E(Y_1)E(Y_2):$$

mean independent product equals product mean (Proposition 5.8.10).

EXAMPLE 3

Suppose that we have found someone who will pay us in dollars the sum plus the product of the points on two good dice; what is a fair price to pay to play this game? Let $Y_1$ and $Y_2$ be the points on the first and second dice, respectively. Then the payoff is $y_1 + y_2 + y_1 y_2$ and it has mean value

$$E(y_1 + y_2 + y_1 y_2) = E(y_1) + E(y_2) + E(y_1)E(y_2) = 3.5 + 3.5 + (3.5)^2 = 19.25.$$  

If a person pays $19.25, the net payoff is $y_1 + y_2 + y_1 y_2 - 19.25$, and it has mean value

$$E(y_1 + y_2 + y_1 y_2 - 19.25) = 19.25 - 19.25 = 0.$$  

Thus if a person pays $19.25 to play the game, the mean exchange on a play is 0 and the game is called fair.

D EXERCISES

1 A modified slot machine pays $50 with probability 1/100, $1 with probability 10/100, and zero otherwise. Calculate the mean payoff. If a person pays $1 to play, determine the mean net payoff.
2 A person rolls a die and receives 5 if the ace comes up; how much should he pay otherwise so that the game is fair?
3 A distribution for $y$ has density function $3y^2$ on $(0, 1)$ and zero elsewhere. Determine $E(y)$ and $E(y^2)$.
4 A distribution for $y$ has density $12y^2(1 - y)$ on $(0, 1)$ and zero elsewhere. Determine $E(y)$ and $E(y^2)$.
5 A symmetrical coin is tossed until the first head occurs; let $x$ be number of tosses. Consider the payoff $y = 2^x$. Show that $E(y)$ does not exist.
6 Consider the standard exponential distribution for $x$. Show that $E(y)$ does not exist for the payoff $y = e^x$.

5.2

THE LOCATION OF A DISTRIBUTION

A statistician has a distribution that describes the lifetime of high-pressure sodium lamps. Of course, he is interested in the location of the distribution, a location point about which realized values tend to be distributed. Also, he is interested in the scaling of the distribution, a distance from the location point that represents the variability in realized values. In this section we consider the mean value as a possible measure of the location of a distribution on the real line $R^1$ or on a real space $R^k$. We then examine scaling in the next section.

A LOCATION ON $R^1$

Consider a distribution for $x$ on the real line $R$ as given by the probability measure $P$, or as given by the probability function $p$ in the discrete case or the probability density function $f$ in the absolutely continuous case; for example, $x$ could be the lifetime for a certain type of printed circuit. Consider the mean value for such a variable.

DEFINITION 1

The mean of the distribution is the point $\mu$,

\[
\mu = \mu_x = E(x) = \int_R x \, dP(x)
\]

\[
= \sum x p(x) \quad \text{if discrete},
\]

\[
= \int_{-\infty}^{\infty} x f(x) \, dx \quad \text{absolutely continuous}
\]

provided that the integral exists (absolute convergence; Definition 5.1.2); otherwise, the mean diverges or does not exist; several notations are recorded.

EXAMPLE 1

The uniform $[a, b]$ distribution: The density function for this uniform distribution is equal to $1/(b - a)$ on $(a, b)$ and zero elsewhere. The mean of the distribution is
\[
\mu = \int_{-\infty}^{\infty} x f(x) \, dx = \int_{a}^{b} x \frac{1}{b - a} \, dx = \frac{b^2/2 - a^2/2}{b - a} = \frac{a + b}{2}.
\]

Note that the distribution is symmetric and that the mean is at the point of symmetry.

The mean value of a distribution is presented in Figure 5.2 as a point \( \mu \) on the space of the distribution. If we consider the deviations \( x - \mu \) of values \( x \) from \( \mu \), then some deviations will be positive and some negative, but the mean deviation from \( \mu \) is zero:

\[
(2) \quad E(x - \mu) = E(x) - E(\mu) = \mu - \mu = 0;
\]

use properties 1 and 2 in Section 5.1C. If we think of probability as unit mass distributed on \( R \), then \( \mu \) is the center of gravity, the point about which the mass distribution balances; the equation states that the moment of the force of gravity about \( \mu \) is zero.

Consider a distribution that is symmetrical about some point \( m \), and suppose that the mean \( \mu \) of the distribution exists. Symmetry can be expressed by saying that \( x - m \) and \(- (x - m)\) have the same distribution and thus the same mean: \( E(x - m) = E(- (x - m)) \). Thus (properties 1 and 2)

\[
E(x) - m = -E(x) + m,
\]

\[
E(x) = m.
\]

Hence the mean \( \mu \) is at the point \( m \) of symmetry.

Consider an affine transformation \( y = a + cx = Y(x) \) carrying the initial response \( x \) into the new response \( y \); for example, such a transformation might correspond to a change from Celsius to Fahrenheit. We have (properties 1 and 2)

\[
(3) \quad E(y) = E(a + cx) = a + cE(x),
\]

\[
\mu_Y = a + c \mu.
\]

FIGURE 5.2
Mean of a discrete distribution; mean of an absolutely continuous distribution; mean after a transformation \( y = a + cx \); illustration has \( c > 0 \).
Thus the transformation also carries the initial mean into the new mean; this is illustrated in Figure 5.2 in terms of a relabeling of the points on $R$.

Now consider a distribution for $(x_1, x_2)$ on the plane and suppose that we form a new response $y = x_1 + x_2 = Y(x_1, x_2)$; then

$$
(4) \quad \mu_Y = E(y) = E(x_1 + x_2) = E(x_1) + E(x_2) = \mu_1 + \mu_2.
$$

Thus, if we add the variables, then we add the means.

Now suppose that the probability distribution has come from some initial space $\mathcal{S}$ with class $\mathcal{E}$ and measure $P$. Let $X$ be a real-valued measurable function; then

$$
(5) \quad \mu_X = E(X) = \int_{\mathcal{S}} X(s) \, dP(s) = \int_R x \, dP_X(x),
$$

using formula (5.1.9). And if $Y = \sigma + cX$, then $\mu_Y = \sigma + c\mu_X$.

\section*{B SOME EXAMPLES ON $R^1$}

\subsection*{EXAMPLE 2}

The normal distribution: Let us examine a standard normal distribution for $z$. The distribution is symmetric about 0; and the mean clearly exists because of the exponential tails: thus $E(z) = 0$. Indeed, we have

$$
(6) \quad E(z) = \int_{-\infty}^{\infty} z \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \, dz = 0,
$$

as clearly the integrand is an odd function. Now suppose that we generate the general normal($\mu, \sigma$) distribution by the transformation $y = \mu + az$; then

$$
E(y) = E(\mu + az) = \mu + aE(z) = \mu.
$$

Thus the $\mu$ we have used in defining the general normal is, in fact, the mean of the distribution.

\subsection*{EXAMPLE 3}

The standard Cauchy distribution: This distribution is symmetrical, but the integral for the mean

$$
\int_{-\infty}^{\infty} z \frac{1}{\pi} \frac{1}{1 + z^2} \, dz
$$

diverges (recall Definition 5.1.2); thus the mean of the Cauchy does not exist. In Section A we saw some of the attractive properties of the mean as a measure of location, and we now see one of the unattractive properties—that it does not exist for some mathematically reasonable distributions. We need other measures of location for the Cauchy; see Problems 14 and 17.

\subsection*{EXAMPLE 4}

The Poisson($\lambda$) distribution: For a discrete example we examine the Poisson($\lambda$) distribution:
\[ E(y) = \sum_{y=0}^{\infty} y \frac{\lambda^y e^{-\lambda}}{y!} = \sum_{y=1}^{\infty} \frac{\lambda^y e^{-\lambda}}{(y-1)!} = \lambda \cdot \sum_{y=1}^{\infty} \frac{\lambda^{y-1} e^{-\lambda}}{(y-1)!} = \lambda, \]

where the sum of Poisson probabilities is 1. We have now verified that the mean of the Poisson(\(\lambda\)) is just \(\lambda\); recall the large-sample data on \(\alpha\)-particles in Section 2.1E.

Note: The first term with \(y = 0\) can be omitted; the \(y\) cancels with the first factor in \(y!\); the adjusted summation is a summation of Poisson probabilities and gives 1. This calculation method recurs with many discrete distributions: the expression being averaged cancels into the functional form of the distribution; with appropriate factors removed, the residual summation is a sum of probabilities (= 1) for some value of the parameter of the distribution. The binomial\((n, p)\) distribution provides another illustration and we obtain \(E(y) = np\).

\[ \mu \]

\[ \int_0^{\infty} y \, dP(y) \geq \int_1^{\infty} y \, dP(y) \geq \int_t^{\infty} t \, dP(y) = tP([t, \infty)). \]

Note: The integrand is nonnegative and we reduce the range of integration; we replace \(y\) by the lower bound \(t\); see Figure 5.3. Thus we obtain Chebyshev’s inequality:

\[ P(y \geq t) \leq \frac{\mu}{t} \quad \text{or} \quad P(y \geq k\mu) \leq \frac{1}{k} \]

for a distribution on \([0, \infty)\). Note, for example, that there is at most 1/3 probability beyond 3\(\mu\).

D LOCATION ON \(R^k\)

Now we examine the parallel development on \(R^k\). Consider a distribution for \(x = (x_1, \ldots, x_k)'\) on \(R^k\) as given by the probability measure \(P\), or as given by the

FIGURE 5.3
On \([0, \infty)\), the identity function is greater than or equal to the function \(t(1, \infty)\).
probability function $p$ in the discrete case or the probability density function $f$ in the absolutely continuous case. For example, $x_1$, $x_2$, and $x_3$ could be the time, intensity, and direction of the reaction to a stimulus in psychological testing.

**DEFINITION 2**

The mean of the distribution is the point $\mu = (\mu_1, \ldots, \mu_k)'$.

$E(x) = (E(x_1), \ldots, E(x_k))'$,

in $\mathbb{R}^k$ provided that all the coordinate means exist; otherwise, we say that the mean diverges or does not exist.

See Figure 5.4. In general, if $E$ is applied to a vector or to a matrix, then by definition it gives the vector or matrix formed by applying $E$ to each element; the linearity properties of $E$ make this a fruitful definition. Note that a coordinate mean can be calculated in a variety of ways:

$E(x_i) = \int_{\mathbb{R}^k} x_i \, dP(x)$

$= \sum p(x) = \sum x_i p_i(x_i) \quad \text{if discrete}$

$= \int_{\mathbb{R}^k} x_i f(x) \, dx = \int_{\mathbb{R}^k} x_i f_i(x_i) \, dx_i \quad \text{absolutely continuous}$

**EXAMPLE 5**

A distribution on $\mathbb{R}^2$: Consider the absolutely continuous distribution for $(x, y)$ with density $f$ in Example 3.2.6:
FIGURE 5.5
Mean \( \mu = (1/2, 1/4)' \) of the distribution in Example 5.

\[ f(x, y) = 6x \quad \text{if } 0 < x < 1, \ 0 < y < 1 - x \]

and zero elsewhere. The marginal densities are

\[ f_1(x) = 6x(1 - x) \quad \text{if } 0 < x < 1, \]
\[ f_2(y) = 3(1 - y)^2 \quad 0 < y < 1 \]

and zero elsewhere. Thus we calculate

\[
\mu_1 = \int_{\mathbb{R}^2} xf(x, y) \, dx \, dy = \int_{\mathbb{R}} xf_1(x) \, dx = \int_0^1 x \cdot 6x(1 - x) \, dx = 1/2.
\]

\[
\mu_2 = \int_{\mathbb{R}^2} yf(x, y) \, dx \, dy = \int_{\mathbb{R}} yf_2(y) \, dy = \int_0^1 y \cdot 3(1 - y)^2 \, dy = 1/4.
\]

Thus the mean \( \mu \) is \((1/2, 1/4)'\), as indicated in Figure 5.5. Note that the integrals are available from Problem 2.2.22 or indirectly from Problem 2.2.17.

The mean of a distribution is presented in Figure 5.4 as a point in \( \mathbb{R}^k \). If we consider the deviation vector \( \mathbf{x} - \mu \), then its values center at \( \mathbf{0} \):

\[ E(\mathbf{x} - \mu) = E(\mathbf{x}) - E(\mu) = \mu - \mu = \mathbf{0}. \]

If we think of probability as unit mass distributed on the plane \( \mathbb{R}^2(k = 2 \text{ case}) \), then the mass balances about the line \( x_1 = \mu_1 \); it also balances about \( x_2 = \mu_2 \); and it balances about any line \( \Sigma l_i x_i = \Sigma l_i \mu_i \) through the point \( \mu \); in fact, it balances on the point \( \mu \).

Consider an affine transformation \( \mathbf{y} = \mathbf{a} + \mathbf{Cx} \), where \( \mathbf{x} \) in \( \mathbb{R}^k \) is mapped to \( \mathbf{y} \) in \( \mathbb{R}^r \) and \( \mathbf{a} \) is \( r \times 1 \) and \( \mathbf{C} \) is \( r \times k \); then the mean \( \nu \) of \( \mathbf{y} \) is

\[ \nu = E(\mathbf{y}) = E(\mathbf{a} + \mathbf{Cx}) = \mathbf{a} + C \mathbf{E(x)} = \mathbf{a} + C \mu. \]

Thus the transformation also carries the original mean into the new mean.

Now consider a distribution for \( (x_1, x_2) \), where each \( x_i \) is in \( \mathbb{R}^k \), and suppose that we form a new response \( \mathbf{y} = x_1 + x_2 \). Then

\[ E(\mathbf{y}) = E(x_1 + x_2) = E(x_1) + E(x_2). \]

Thus, if we add vector variables, we add the vector means.
E SOME EXAMPLES ON $\mathbb{R}^k$

EXAMPLE 6

A sample from the Bernoulli($\rho$) distribution: Consider a sample $(x_1, \ldots, x_n)$ from the Bernoulli($\rho$) distribution. Then

$$E(x_i) = 1 \cdot \rho + 0 \cdot q = \rho;$$

thus the mean of the distribution is

$$E(x) = (\rho, \ldots, \rho)' = \rho 1,$$

where $1 = (1, \ldots, 1)'$ is the 1-vector.

Now suppose that we form the binomial variable $y = \sum_i x_i$; then

$$E(y) = E(x_1 + \cdots + x_n) = \rho + \cdots + \rho = np;$$

thus the mean of the binomial($n, \rho$) distribution is $np$. Note, for example, that with 100 tosses of a symmetric coin, the mean number of heads is 50.

EXAMPLE 7

The normal distribution on $\mathbb{R}^k$: Consider a standard normal distribution for $z$ on $\mathbb{R}^k$; note that we can consider $z = (z_1, \ldots, z_k)'$ as a sample from the standard normal on $\mathbb{R}^1$. Using Example 2, we obtain the mean

$$E(z) = (E(z_1), \ldots, E(z_k))' = (0, \ldots, 0)' = 0.$$

Now suppose that we generate the general normal by the transformation $y = \mu + \Gamma z$, where $\mu$ is $k \times 1$ and $\Gamma$ is $k \times k$. Then the multivariate normal($\mu$; $\Sigma$) distribution has mean

$$E(y) = E(\mu + \Gamma z) = \mu + \Gamma E(z) = \mu.$$

Thus the $\mu$ we have used to locate the general normal is, in fact, the mean of the distribution.

F EXERCISES

1. The binomial($4$, $1/2$) distribution is tabulated in Section 3.5A. Calculate the mean from the tabulation.
2. The binomial($4$, $3/4$) distribution is tabulated in Section 3.5A. Calculate the mean from the tabulation.
3. The hypergeometric($8$, $4$, $1/2$) distribution is tabulated in Section 3.5C. Calculate the mean from the tabulation.
4. The hypergeometric($8$, $4$, $3/4$) distribution is tabulated in Section 3.5C. Calculate the mean from the tabulation.
5. Consider a distribution for $y$ with probability function $p(y) = 2^{-y}$ on $y = 1, 2, \ldots$ and zero elsewhere. Calculate $E(y)$.
6. A distribution on $\mathbb{R}$ has density function $f$ given by $f(y) = 1 - y/2$ on $(0, 2)$ and zero elsewhere. Calculate the mean.
7. A distribution on $\mathbb{R}$ has density function $f$ given by $f(y) = 12y(1 - y)^2$ on $(0, 1)$ and zero elsewhere. Calculate the mean.
8 The hypergeometric(6, 3, 1/6, 2/6, 3/6) distribution is tabulated in Section 3.5D. Calculate the mean \( \mu_1, \mu_2, \mu_3 \) from the tabulation.

9 The multinomial(5, 1/6, 2/6, 3/6) distribution for \( y_1, y_2, y_3 \) is tabulated in Section 3.5B and in Section 4.2A. Calculate \( E(y_1) \) and \( E(y_3) \) from the tabulation; calculate \( E(y_2) \) from \( y_1 + y_2 + y_3 = 5 \); record the mean of the distribution for \( y_1, y_2, y_3 \).

10 Calculate the mean of the uniform distribution on the unit square \((0, 2)^2\).

G PROBLEMS

Additional problems concerning means will be found at the end of Section 5.3.

11 For a discrete distribution on the nonnegative integers, show that

\[
E(y) = \sum_{j=0}^{\infty} P(y > j).
\]

12 Consider a distribution on \( R \) with \( E(|y|^r) < \infty (r > 0) \); show that for \( t > 0 \),

\[
P(|y| \geq t) \leq \frac{E(|y|^r)}{t^r}.
\]

13 Consider a distribution on \( R \) with \( E(\phi(y)) < \infty \), where \( \phi(y) \geq 0 \) and \( \phi(y) \geq k > 0 \) for \( y \) in \( A \); show that

\[
P(A) \leq \frac{E(\phi(y))}{k}.
\]

14 The median is often used as a location measure; the median \( \tilde{y} \) is a solution of the relation \( F(x - \tilde{y}) = 1/2 \leq F(x) \) when \( F \) is the distribution function. Note that if \( F \) is strictly increasing, then \( \tilde{y} \) is unique. If a new response is obtained by \( y = h(x) \), where \( h \) is strictly increasing and continuous, then show that the median \( \tilde{y} \) of \( y \) is given by the same transformation: \( \tilde{y} = h(\tilde{y}) \).

15 Let \( a_1 < a_2 < \cdots < a_n \) be real numbers. Show that \( \sum |a_i - t| \) is a minimum if \( t \) is a median \( \bar{a} \) of the \( n \) numbers; that is, the number of \( a_i < \bar{a} \) is equal to the number \( a_i > \bar{a} \).

(continuation) Consider a distribution for \( y \) on the line \( R \). Show that \( E(|y - t|) \) is a minimum if \( t \) is a median of the distribution.

17 The mode of a distribution with density \( f \) on the real line is a value \( \mu^0 \) such that \( f(\mu^0) \geq f(x) \) for all \( x \). If a new response \( y \) is obtained by \( y = a + cx \) (\( c \neq 0 \)), then show that the mode \( \mu^0 \) of \( y \) is given by the same transformation \( \mu^0 = a + c\mu^0 \).

5.3

THE SCALING OF A DISTRIBUTION

We have mentioned the statistician who has a distribution describing the lifetime of high-pressure sodium lamps, and is interested in its location and scaling. In the preceding section we discussed the mean as a measure of the location of a distribution on the real line. In this section we examine the scaling of a distribution on the real line, scaling as measured by the variance or by its square root the standard deviation. We also examine the scaling of a distribution on the plane or on \( R^k \); the formulas for this follow by a straightforward matrix extension of those presented for the simple case on the real line.
The scaling can be of interest in itself. For example, how variable is the lifetime for a certain type of printed circuit? Or, more frequently, it can be of interest as a measure of precision—the degree with which a response value will approximate the location of the distribution; but more on this in the later chapters on statistics.

**A SCALING ON \( \mathbb{R}^1 \)**

Consider a distribution for \( x \) on the real line \( \mathbb{R} \) as given by the probability measure \( P \), or as given by the probability function \( p \) in the discrete case or the probability density function \( f \) in the absolutely continuous case.

**DEFINITION 1**

The **variance** of the distribution is

\[
\sigma^2 = \sigma_x^2 = \text{var}(x) = E((x - \mu)^2) = \int_{\mathbb{R}} (x - \mu)^2 \, dP(x)
\]

\[
= \sum (x - \mu)^2 p(x) \quad \text{if discrete,}
\]

\[
= \int_{-\infty}^{\infty} (x - \mu)^2 f(x) \, dx \quad \text{absolutely continuous}
\]

provided that the integral exists; otherwise, the variance diverges or does not exist; alternative notations are recorded.

The variance is the mean-squared deviation from the location or center of the distribution. The variance is measured in squared units. The square root of the variance is measured in the same units as the original variable for the distribution; it is called the **standard deviation** (or RMS, root mean square).

**DEFINITION 2**

The **standard deviation** of the distribution is

\[
\sigma = \sigma_x = \text{SD}(x) = \sqrt{\text{var}(x)} = [E((x - \mu)^2)]^{1/2}.
\]

**EXAMPLE 1**

The uniform\((0, 1)\) distribution: The uniform\((0, 1)\) distribution has mean \( \mu = 1/2 \); see Example 5.2.1. Thus we can calculate the variance as follows:

\[
\sigma^2 = \int_{-\infty}^{\infty} (x - 1/2)^2 f(x) \, dx = \int_{0}^{1} (x - 1/2)^2 \, dx = \frac{1}{12}.
\]

The standard deviation is \( \sigma = 1/\sqrt{12} = 0.2887 \).

The standard deviation is presented in Figure 5.6 as a **scaling unit** or as a **statistical distance** from the mean \( \mu \). For the preceding uniform distribution we see
that $\mu \pm \sigma$ contains about 58 percent of the probability; for the normal the corresponding figure is 68.26 percent. In general, for a distribution with a central hump of probability and not-too-long tails, we find that $\mu \pm \sigma$ contains about $2/3$ of the probability.

Some alternative formulas for the variance can often be more convenient for calculations. The formula

\[(3) \quad \sigma^2 = E((x - \mu)^2) = E(x^2 - 2x\mu + \mu^2) = E(x^2) - 2\mu^2 + \mu^2
\]
\[= E(x^2) - (E(x))^2 = MS - SM,
\]
which presents the variance as mean square minus square mean, can be useful with continuous distributions. And the formula

\[(4) \quad \sigma^2 = E(x(x - 1)) - \mu(\mu - 1) = E(x^{(2)}) - \mu^{(2)}
\]
can be useful for discrete distributions; recall the calculation method in Example 5.2.4.

Consider an affine transformation $y = a + cx$ carrying the initial response $x$ with mean $\mu$ into the new response $y$ with mean $\nu$. Then

$$y = a + cx, \quad \nu = a + c\mu,$$

$$y - \nu = c(x - \mu),$$

$$E((y - \nu)^2) = c^2E((x - \mu)^2).$$

Thus

\[(5) \quad \text{var}(a + cx) = c^2 \text{var}(x), \quad \text{SD}(a + cx) = |c| \text{SD}(x).
\]

The transformation changes the scale by the factor $|c|$ and correspondingly changes the standard deviation by the same factor; see Figure 5.6. In particular, note that relocation alone does not change the variance or standard deviation.
Now consider a distribution for \( (x_1, x_2) \) on the plane with mean \((\mu_1, \mu_2)\) and suppose that we form a new response \( y = x_1 + x_2 \) with mean \( \nu \):

\[
\begin{align*}
\nu &= x_1 + x_2, \\
\gamma &= x_1 - x_2, \\
\gamma - \nu &= x_1 - \mu_1 + x_2 - \mu_2 \\
\text{var} (\gamma) &= E((x_1 - \mu_1 + x_2 - \mu_2)^2) \\
&= E((x_1 - \mu_1)^2 + 2(x_1 - \mu_1)(x_2 - \mu_2) + (x_2 - \mu_2)^2) \\
&= \text{var} (x_1) + 2E((x_1 - \mu_1)(x_2 - \mu_2)) + \text{var} (x_2).
\end{align*}
\]

Now suppose that \( x_1 \) and \( x_2 \) are statistically independent. Then by formula (5.1.13), the middle term,

\[
E((x_1 - \mu_1)(x_2 - \mu_2)) = E(x_1 - \mu_1)E(x_2 - \mu_2) = 0 \cdot 0 = 0,
\]

is a product of mean deviations and thus is equal to zero. Hence for independent variables we have

\[
\text{var} (x_1 + x_2) = \text{var} (x_1) + \text{var} (x_2).
\]

Thus, if we add independent variables, we add the variances.

Now suppose that the probability distribution has come from some initial space \( S \) with class \( \mathcal{G} \) and measure \( P \). Let \( X \) be a real-valued measurable function; then

\[
\sigma_X^2 = E((X - \mu_X)^2) = \int (X(s) - \mu_X)^2 \, dP(s).
\]

And if \( Y = a + cX \), then \( \sigma_Y^2 = c^2 \sigma_X^2 \).

**B SOME EXAMPLES ON \( \mathbb{R}^1 \)**

**EXAMPLE 2**

The uniform\((a, b)\) distribution: The uniform\((0, 1)\) distribution has mean \( \mu = 1/2 \); we recalculate the variance:

\[
E(x^2) = \int_{-\infty}^{\infty} x^2 f(x) \, dx = \int_{0}^{1} x^2 \, dx = 1/3;
\]

\[
\text{var} (x) = (1/3) - (1/2)^2 = 1/12.
\]

Now suppose that we generate the uniform\((a, b)\) distribution by the transformation \( y = a + (b - a)x \). Then

\[
\mu_y = a + (b - a) \cdot (1/2) = \frac{a + b}{2},
\]

\[
\sigma_y^2 = (b - a)^2 \cdot (1/12), \quad \sigma_y = \frac{b - a}{\sqrt{12}} = 0.2887(b - a).
\]

**EXAMPLE 3**

The normal distribution: The standard normal distribution for \( z \) has \( E(z) = 0 \). We calculate the variance:
\[ E(z^2) = \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \, dz = \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} z \cdot e^{-z^2/2} \, dz \]
\[ = \frac{-2}{\sqrt{2\pi}} \left( \int_{0}^{\infty} e^{-z^2/2} \, dz \right)^2 = 1; \]
thus
\[ \text{var}(z) = E(z^2) - (E(z))^2 = 1 - 0 = 1. \]

Now, suppose that we generate the general normal(\(\mu, \sigma\)) distribution by the transformation \(y = \mu + \sigma z\); then
\[ \text{var}(y) = \sigma^2 \cdot 1 = \sigma^2. \]
Thus the \(\sigma\) we have used in defining the general normal is, in fact, the standard deviation of the distribution. Recall from Section 2.2C that \(\mu \pm \sigma\) contains 68.26 percent of the probability.

**EXAMPLE 4**

**The Poisson(\(\lambda\)) distribution:** In Example 5.2.4 we saw that the mean is \(\lambda\). For the variance we first calculate
\[ E(y^{(2)}) = \sum_{y=0}^{\infty} y(y - 1) \frac{\lambda^y e^{-\lambda}}{y!} = \sum_{y=2}^{\infty} \frac{\lambda^y e^{-\lambda}}{(y - 2)!} = \lambda^2, \]
following the pattern in the earlier example. Thus the variance is
\[ \text{var}(y) = E(y^{(2)}) - (E(y)) = \lambda^2 - \lambda(\lambda - 1) = \lambda. \]
And the standard deviation is \(\sqrt{\lambda}\); note that this is the square root of the mean.

**EXAMPLE 5**

**A sample from the Bernoulli(\(p\)) distribution:** In Example 5.2.6 we saw that the mean of the Bernoulli(\(p\)) distribution is \(p\). We calculate the variance:
\[ E(x^2) = 1^2 \cdot p + 0^2 \cdot q = p, \quad \text{or} \quad E(x^{(2)}) = 0 \cdot p + 0 \cdot q = 0, \]
\[ \text{var}(x) = p - p^2 = p(1 - p) = pq, \quad \text{or} \quad \text{var}(x) = 0 - p^2 = pq. \]
Now consider a sample \((x_1, \ldots, x_n)\) from the Bernoulli(\(p\)) distribution. We form the binomial variable \(y\) and from formula (7) obtain
\[ \text{var}(y) = \text{var}(x_1) + \cdots + \text{var}(x_n) = npq. \]
Thus the variance of the binomial(\(n, p\)) distribution is \(npq\). Note, for example, that with 100 tosses of a symmetric coin the variance of the number of heads is 25 and the standard deviation is 5.

**WHAT DOES \(\sigma\) SAY ABOUT \(p\)?**

If we know that a distribution has variance \(\sigma^2\), then what constraints does this place on how the probability is distributed on the line \(R\)?
The variance has a rather powerful null property: the variance \( \text{var}(x) = 0 \) if and only if \( x \) is equal to a constant with probability 1 (all probability at a point). For consider the variance

\[
\text{var}(x) = \int_{-\infty}^{\infty} (x - \mu)^2 \, dP(x).
\]

By formula (24) of Supplement Section 5.8 this variance can be zero if and only if \((x - \mu)^2 = 0\) with probability 1, that is \(x = \mu\) with probability 1. This can be a very useful criterion for determining if a distribution is concentrated with all probability at a point.

Now consider a distribution for \( x \) with mean \( \mu \) and variance \( \sigma^2 \) and apply the result in Section 5.2C with \( y \) replaced by \((x - \mu)^2\) and \( t \) replaced by \( t^2 \):

\[
(8) \quad P((x - \mu)^2 \geq t^2) \leq \frac{\sigma^2}{t^2}.
\]

The alternative form,

\[
(9) \quad P(|x - \mu| \geq k\sigma) \leq \frac{1}{k^2},
\]

is another version of Chebyshev’s inequality. Thus there is at most 1/9 probability outside the interval \((\mu - 3\sigma, \mu + 3\sigma)\).

In practice the variance is more specific. For distributions with a central pile of probability and not too much probability in the tails, the interval \((\mu - \sigma, \mu + \sigma)\) contains about 2/3 of the probability.

### D SCALING ON \( R^k \)

Consider a distribution for \( x = (x_1, \ldots, x_k)' \) on \( R^k \) as given by the probability measure \( P \)—or as given by the probability function \( p \) in the discrete case, or the probability density function \( f \) in the absolutely continuous case. For example, \( x_1, x_2, \) and \( x_3 \) could be time, intensity, and direction for the reaction to a stimulus in the psychological testing mentioned earlier. Let \( \mu = (\mu_1, \ldots, \mu_k)' \) be the mean. We first define the covariance between two coordinates:

#### DEFINITION 3

The covariance between \( x_i \) and \( x_j \) is

\[
(10) \quad \sigma_{ij} = \text{cov}(x_i, x_j) = E((x_i - \mu_i)(x_j - \mu_j)) = \int_{R^k} (x_i - \mu_i)(x_j - \mu_j) \, dP(x)
\]

\[
= \sum_{x} (x_i - \mu_i)(x_j - \mu_j)p_{ij}(x_i, x_j) \quad \text{if discrete,}
\]

\[
= \int\int_{-\infty}^{\infty} (x_i - \mu_i)(x_j - \mu_j)f_{ij}(x_i, x_j) \, dx_i \, dx_j \quad \text{if absolutely continuous}
\]

if the integral exists; otherwise, we say that the covariance diverges or does not exist.

**Note:** If \( x_i \) and \( x_j \) are statistically independent, then \( \text{cov}(x_i, x_j) = 0 \); see formula (6).
The covariances can be assembled to form a variance matrix:

**DEFINITION 4**

The variance matrix of a distribution is

\[
\Sigma = \text{VAR}(x) = \begin{pmatrix}
\sigma_{11} & \cdots & \sigma_{1k} \\
\vdots & \ddots & \vdots \\
\sigma_{k1} & \cdots & \sigma_{kk}
\end{pmatrix}
\]

\[
= E \begin{pmatrix}
(x_1 - \mu_1)(x_1 - \mu_1) & \cdots & (x_1 - \mu_1)(x_k - \mu_k) \\
\vdots & \ddots & \vdots \\
(x_k - \mu_k)(x_1 - \mu_1) & \cdots & (x_k - \mu_k)(x_k - \mu_k)
\end{pmatrix}
\]

The variance matrix can be expressed more compactly using matrix multiplication:

\[
\Sigma = E\left( \begin{pmatrix} x_1 - \mu_1 \\ \vdots \\ x_k - \mu_k \end{pmatrix} (x_1 - \mu_1, \ldots, x_k - \mu_k) \right) = E((x - \mu)(x - \mu)').
\]

The matrix may seem unwieldy, but we will see later that it is very useful. Note that \(\Sigma\) is a symmetric matrix; also it can be shown that \(\Sigma\) is positive semidefinite (Problem 32).

**EXAMPLE 6**

The uniform distribution on \((0, 1)^2\): Consider the uniform distribution for \((x_1, x_2)\) over the unit square \((0, 1)^2\). The marginal distribution of \(x_1\) is uniform\((0, 1)\); thus \(E(x_1) = 1/2\), \(\text{var}(x_1) = 1/12\) by Example 1. Similarly, \(E(x_2) = 1/2\), \(\text{var}(x_2) = 1/12\). We calculate

\[
\text{cov}(x_1, x_2) = \int_{x_1} (x_1 - 1/2)(x_2 - 1/2)f(x_1, x_2) \, dx_1 \, dx_2 = \left( \int_0^1 (x - 1/2) \, dx \right)^2 = 0
\]

by the factorization formula (5.1.13). Thus the mean and variance matrix are

\[
\mu = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1/12 & 0 \\ 0 & 1/12 \end{pmatrix}.
\]

Some alternative formulas can often be convenient for calculations. For the covariance we have

\[
\sigma_{ij} = E((x_i - \mu_i)(x_j - \mu_j)) = E(x_ix_j - \mu_i\mu_j - x_i\mu_j + \mu_i\mu_j) = E(x_ix_j) - \mu_i\mu_j = \text{MP} - \text{PM},
\]

which presents the covariance as mean product minus product mean. The analogous result for the variance matrix itself is equally simple:

\[
\Sigma = \text{VAR}(x) = E((x - \mu)(x - \mu)') = E(xx' - \mu x' - x\mu' + \mu\mu') = E(xx') - \mu x' = \text{MP} - \text{PM},
\]

which is mean product minus product mean, where the product is given by matrix multiplication.
Now consider a distribution for \( x' = (x_1, \ldots, x_k, x_{k+1}, \ldots, x_{k+l}) = (x'_1, x'_2) \) on \( \mathbb{R}^{k+l} \); note that the first \( k \) coordinates are grouped as a vector \( x_1 \) and the last \( l \) as \( x_2 \). The variance matrix \( \Sigma \) for \( x \) can be partitioned into the variance matrices \( \Sigma_{11} \) for \( x_1 \) and \( \Sigma_{22} \) for \( x_2 \) and the covariance matrix \( \Sigma_{12} = \text{COV} (x_1, x_2) \) between \( x_1 \) and \( x_2 \); we have

\[
\Sigma = \begin{pmatrix}
\sigma_{11} & \cdots & \sigma_{1k} & \sigma_{1k+1} & \cdots & \sigma_{1k+l} \\
\vdots & & \vdots & \vdots & & \vdots \\
\sigma_{k1} & \cdots & \sigma_{kk} & \sigma_{kk+1} & \cdots & \sigma_{kk+l} \\
\sigma_{k+1,1} & \cdots & \sigma_{k+1,k} & \sigma_{k+1,k+1} & \cdots & \sigma_{k+1,k+l} \\
\vdots & & \vdots & \vdots & & \vdots \\
\sigma_{k+l,1} & \cdots & \sigma_{k+l,k} & \sigma_{k+l,k+1} & \cdots & \sigma_{k+l,k+l}
\end{pmatrix} = \begin{pmatrix}
\Sigma_{11} \\
\Sigma_{21} \\
\Sigma_{12} \\
\Sigma_{22}
\end{pmatrix}.
\]

Note that the covariance matrix

\[
\Sigma_{12} = \text{COV} (x_1, x_2) = E((x_1 - \mu_1)(x_2 - \mu_2)') = E(x_1, x_2) - \mu_1 \mu_2'
\]

records the covariances between coordinates of \( x_1 \) and coordinates of \( x_2 \); and \( \Sigma_{21} = \Sigma_{12}' \).

Consider an affine transformation \( y = a + Cx \), where \( x \) in \( \mathbb{R}^k \) is mapped to \( y \) in \( \mathbb{R}^r \) and \( a \) is \( r \times 1 \) and \( C \) is \( r \times k \). Let \( \mu \) and \( \Sigma \) be the mean and variance matrix of \( x \); and let \( \nu \) be the mean for \( y \). Then

\[
y = a + Cx, \quad \nu = a + C\mu, \quad y - \nu = C(x - \mu).
\]

Thus the variance matrix for \( y \) is

\[
\text{VAR} (y) = E((y - \nu)(y - \nu)') = E(C(x - \mu)(x - \mu)'C') = CE((x - \mu)(x - \mu)'C')C' = C\Sigma C' = C \text{VAR} (x) C'.
\]

In particular, we obtain

\[
\text{var} \left( a_1 + \sum_j c_{1j}x_j \right) = \sum_{j, j'=1}^k c_{1j}c_{1j'}\sigma_{jj'},
\]

\[
\text{cov} \left( a_1 + \sum_j c_{1j}x_j, a_2 + \sum_j c_{2j}x_j \right) = \sum_{j, j'=1}^k c_{1j}c_{2j'}\sigma_{jj'}
\]

for the \((1, 1)\) and \((1, 2)\) elements in \( \text{VAR} (y) \).

Now consider a distribution for \((x_1, x_2)\) where each \( x_i \) is in \( \mathbb{R}^k \), and assume that \( x_1 \) is statistically independent of \( x_2 \). Let \( \mu_1 \) and \( \Sigma_{11} \) be the mean and variance matrix for \( x_1 \), and \( \mu_2 \) and \( \Sigma_{22} \) for \( x_2 \); note that the covariance matrix \( \Sigma_{12} = 0 \) by independence. Suppose that we form a new variable \( y = x_1 + x_2 \) with mean \( \nu \):

\[
y = x_1 + x_2, \quad \nu = \mu_1 + \mu_2, \quad y - \nu = x_1 - \mu_1 + x_2 - \mu_2.
\]

Thus the variance matrix for \( y \) is

\[
\text{VAR} (y) = E((x_1 - \mu_1 + x_2 - \mu_2)(x_1 - \mu_1 + x_2 - \mu_2)')
\]

\[
= E((x_1 - \mu_1)(x_1 - \mu_1)' + (x_1 - \mu_1)(x_2 - \mu_2)'
\]

\[
+ (x_2 - \mu_2)(x_1 - \mu_1)' + (x_2 - \mu_2)(x_2 - \mu_2)').
\]

\[
= \Sigma_{11} + \Sigma_{12} + \Sigma_{21} + \Sigma_{22} = \Sigma_{11} + \Sigma_{22}.
\]

Thus if we add independent vectors, we add the variance matrices.
FIGURE 5.7
Mean $\mu$ and variance $\Sigma$ for the uniform distribution on a parallelogram.

**E SOME EXAMPLES**

**EXAMPLE 7**

**Uniform distribution on a parallelogram:** First consider a uniform distribution for $(x_1, x_2)$ on the unit square $(0, 1)^2$; the mean and variance matrix are recorded in Example 6. Now suppose that we form a new variable $(y_1, y_2)$ by the transformation

$$y_1 = 1 + x_1 - x_2, \quad y_2 = 2 + 2x_1 + x_2,$$

so that $(y_1, y_2) = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} + \begin{pmatrix} 1 & -1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$;

it is straightforward to see that $y$ is uniformly distributed over the parallelogram with vertices $(1, 2), (2, 4), (0, 3), (1, 5)$; see Example 2.4.6. The mean and variance matrix of $y$ are

$$\mu = \begin{pmatrix} 1 \\ 2 \end{pmatrix} + \begin{pmatrix} 1 & -1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix} = \begin{pmatrix} 1 \\ 3/2 \end{pmatrix}$$

and

$$\Sigma = \begin{pmatrix} 1 & -1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 1/12 \\ 1/12 \end{pmatrix} = \begin{pmatrix} 1/6 & 1/12 \\ 1/12 & 5/12 \end{pmatrix}.$$

See Figure 5.7.

**EXAMPLE 8**

**A distribution on $\mathbb{R}^2$:** Consider the absolutely continuous distribution for $(x, y)$ with density $f$ as given in Example 3.2.6:

$$f(x, y) = 6x \quad \text{if } 0 < x < 1, 0 < y < 1 - x$$

and zero elsewhere. The marginal densities are
\[ f_1(x) = 6x(1 - x) \quad \text{if } 0 < x < 1 \]
\[ f_2(y) = 3(1 - y)^2 \quad 0 < y < 1 \]

and zero elsewhere. The mean \( \mu = (1/2, 1/4)' \) was calculated in Example 5.2.5.
Now consider calculations toward obtaining the variance matrix:

\[ E(x^2) = \int_0^1 x^2 \cdot 6x(1 - x) \, dx = 6/20, \quad \sigma_{11} = (6/20) - (1/2)^2 = 1/20, \]
\[ E(y^2) = \int_0^1 y^2 \cdot 3(1 - y)^2 \, dy = 1/10, \quad \sigma_{22} = (1/10) - (1/4)^2 = 3/80, \]
\[ E(xy) = \int_0^1 \left( \int_0^{1-x} xy \cdot 6x \, dy \right) \, dx = \int_0^1 6x^2 \frac{(1 - x)^2}{2} \, dx = 1/10, \]
\[ \sigma_{12} = (1/10) - (1/2) \cdot (1/4) = -1/40. \]

Thus the mean and variance matrix are
\[ \mu = \begin{pmatrix} 1/2 \\ 1/4 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1/20 & -1/40 \\ -1/40 & 3/80 \end{pmatrix}. \]

**EXAMPLE 9**

A sample from the Bernoulli(\( p \)) distribution: Let \( x_1, \ldots, x_n \) be a sample from the Bernoulli(\( p \)) distribution. From Example 5.2.6 we have \( E(x) = p \cdot 1 \). From Example 5 we have \( \text{var} (x_i) = pq \) and from the independence of coordinates we have \( \text{cov} (x_i, x_j) = 0 \) for \( i \neq j \); this gives us the variance matrix
\[ \Sigma = \begin{pmatrix} pq & 0 \\ 0 & pq \end{pmatrix}. \]

Now suppose that we form the binomial variable \( y = \Sigma x_i = 1'x \). We have calculated its variance in Example 5; we now calculate it again using formula (17):
\[ \text{var} (y) = (1, \ldots, 1) \begin{pmatrix} pq & 0 \\ 0 & pq \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = (pq, \ldots, pq) \begin{pmatrix} 1 \\ 1 \end{pmatrix} = npq. \]

**EXAMPLE 10**

The normal distribution: Consider a standard normal distribution for \( z \) on \( \mathbb{R}^k \).
The mean \( E(z) = 0 \) was calculated in Example 5.2.7. We note that the \( z_1, \ldots, z_n \) are statistically independent and each is standard normal; thus
\[ E(z_i^2) = 1, \quad E(z_i z_j) = E(z_i)E(z_j) = 0, \quad i \neq j. \]

It follows that the variance matrix is the \( k \times k \) identity matrix \( I \).

Now suppose that we generate the general normal by the transformation \( y = \mu + \Gamma z \), where \( \mu \) is \( k \times 1 \) and \( \Gamma \) is \( k \times k \). We have calculated the mean \( E(y) = \mu \) in Example 5.2.7. We now calculate the variance matrix using formula (17):

\[ \text{VAR} (y) = \Gamma \text{VAR} (z) \Gamma' = \Gamma \Gamma' = \Sigma; \]
this is the matrix $\Sigma$ that was defined in Section 2.4G for a special case, $|\Sigma| \neq 0$. Thus the normal $(\mu, \Sigma)$ distribution has mean $\mu$ and variance matrix $\Sigma$.

**F WHAT DOES $\Sigma$ SAY ABOUT $\rho$?**

Consider the case $k = 2$: the variance matrix is

$$
\Sigma = \begin{pmatrix}
\sigma_{11} & \sigma_{12} \\
\sigma_{21} & \sigma_{22}
\end{pmatrix}.
$$

The diagonal element $\sigma_{11}$ is the variance $\sigma_1^2$ of the first coordinate $x_1$; it describes the scaling of the marginal distribution of $x_1$, that is, of the probability on the plane as projected on the first axis. Similarly, $\sigma_{22}$ is the variance $\sigma_2^2$ of the second coordinate $x_2$; it describes the scaling of the probability as projected onto the second axis. Now consider the covariance between the coordinates:

$$
\sigma_{12} = \sigma_{21} = \mathbb{E}((x_1 - \mu_1)(x_2 - \mu_2)).
$$

Suppose that we standardize it with respect to the scaling $\sigma_1$ for $x_1$ and the scaling $\sigma_2$ for $x_2$ (assume that $\sigma_1$ and $\sigma_2 > 0$); we then obtain the correlation coefficient

$$
(22) \quad \rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \mathbb{E}\left(\frac{x_1 - \mu_1}{\sigma_1} \frac{x_2 - \mu_2}{\sigma_2}\right) = \int \frac{x_1 - \mu_1}{\sigma_1} \frac{x_2 - \mu_2}{\sigma_2} \, dP(x) = \text{cov}\left(\frac{x_1 - \mu_1}{\sigma_1}, \frac{x_2 - \mu_2}{\sigma_2}\right);
$$

this is the covariance between the standardized variables. The correlation coefficient is a measure of how two variables tend to move in the same direction.

From the Schwarz inequality (Problem 17 of Supplement Section 5.8), we obtain the correlation inequality,

$$
(23) \quad -1 \leq \rho_{12} \leq 1,
$$

with equality on the right (left) if and only if

$$
\frac{x_2 - \mu_2}{\sigma_2} = \frac{x_1 - \mu_1}{\sigma_1} \quad \left(\frac{x_2 - \mu_2}{\sigma_2} = -\frac{x_1 - \mu_1}{\sigma_1}\right)
$$

holds with probability 1. Thus $\rho_{12} = +1$ means that all probability is on a line with positive slope, and $\rho_{12} = -1$ means that all probability is on a line with negative slope. In practice with reasonable variables, a large positive correlation means a distribution with a compressed shape tilted up to the right and a large negative correlation means a distribution with a compressed shape tilted down to the right.

Note that independence implies that $\sigma_{12} = 0$ and $\rho_{12} = 0$. But $\sigma_{12} = 0$ or $\rho_{12} = 0$ does not imply independence (see Problem 33).

The Schwarz inequality (Problem 17 of Supplement Section 5.8) can be applied directly to $(x_1 - \mu_1)$ and $(x_2 - \mu_2)$, giving the covariance inequality

$$
(24) \quad (\sigma_{12})^2 \leq \sigma_{11} \sigma_{22}
$$

with equality if and only if $x_1 - \mu_1$ and $x_2 - \mu_2$ are linearly related with probability 1.
**EXERCISES**

1. The binomial(4, 1/2) distribution is tabulated in Section 3.5A. Calculate the variance for the tabulation.
2. The binomial(4, 3/4) distribution is tabulated in Section 3.5A. Calculate the variance for the tabulation.
3. The hypergeometric(8, 4, 1/2) distribution is tabulated in Section 3.5C. Calculate the variance from the tabulation.
4. The hypergeometric(8, 4, 3/4) distribution is tabulated in Section 3.5C. Calculate the variance from the tabulation.
5. A distribution on \( R \) has density \( f(y) = 1 - y/2 \) on (0, 2) and zero elsewhere. Calculate the variance.
6. A distribution on \( R \) has density \( f(y) = 12y(1 - y)^2 \) on (0, 1) and zero elsewhere. Calculate the variance.
7. The hypergeometric(6, 3, 1/6, 2/6, 3/6) distribution is tabulated in Section 3.5D. Calculate the variance matrix \( \Sigma \); record the covariance matrix \( \Sigma_{12} = \text{COV}(\gamma_1, \gamma_2, \gamma_3) \) between \( \gamma_1 \) and \( \gamma_2, \gamma_3 \).
8. Calculate the mean \( \mu \) and variance matrix \( \Sigma \) for the uniform distribution on the square with vertices \((-1, 0), (0, -1), (1, 0), (0, 1)\).

**PROBLEMS FOR DISCRETE DISTRIBUTIONS**

For each of the following distributions, verify the corresponding expressions for
\[
E(\gamma), \quad E(\gamma^2), \quad \text{var}(\gamma).
\]

9. The uniform distribution on the integers 1, 2, \ldots, \( N \):
\[
\frac{N + 1}{2}, \quad \frac{N^2 - 1}{3}, \quad \frac{N^2 - 1}{12}.
\]

10. Poisson(\( \lambda \)) distribution:
\[
\lambda, \quad \lambda^2, \quad \lambda.
\]

11. Geometric(\( p \)) distribution:
\[
\frac{q}{p}, \quad 2 \left( \frac{q}{p} \right)^2, \quad \frac{q}{p^2}.
\]

12. Binomial(\( n, p \)) distribution:
\[
np, \quad n^2p^2, \quad npq.
\]

13. Negative binomial(\( k, p \)) distribution:
\[
k \frac{q}{p}, \quad (k + 1)k \left( \frac{q}{p} \right)^2, \quad k \frac{q}{p^2}.
\]

14. Hypergeometric distribution(\( N, n, p \)):
\[
np, \quad n^2p \frac{Np - 1}{N - 1}, \quad np(1 - p) \frac{N - n}{N - 1}.
\]

For each of the following distributions, verify the corresponding expressions for
\[
E(\gamma_i), \quad E(\gamma_i^2), \quad \text{var}(\gamma_i), \quad E(\gamma_i \gamma_j), \quad \text{cov}(\gamma_i, \gamma_j), \quad i \neq j.
\]
15 Multinomial\((n, p_1, \ldots, p_k)\) distribution:

\[ np_i, \quad n^2p_i^2, \quad np_i(1 - p_i), \quad n^2p_ip_j, \quad -np_ip_j. \]

16 Multivariate hypergeometric\((N, n, p_1, \ldots, p_k)\) distribution:

\[ np_i, \quad n^2p_i \frac{Np_i - 1}{N - 1}, \quad np_i(1 - p_i) \frac{N - n}{N - 1}, \quad n^2p_ip_j \frac{N}{N - 1}, \quad -np_ip_j \frac{N - n}{N - 1}. \]

**I PROBLEMS: FOR ABSOLUTELY CONTINUOUS DISTRIBUTIONS**

For each of the following distributions, verify the corresponding expressions for 

\[ E(y), \quad E(y^2), \quad \text{var} (y). \]

The more difficult problems are marked with an asterisk.

17 **Standard exponential distribution:**

\[ 1, \quad 2, \quad 1. \]

18 **Pareto\((\alpha)\) distribution:**

\[ \frac{1}{\alpha - 1} (\alpha > 1), \quad \frac{2}{(\alpha - 2)(\alpha - 1)} (\alpha > 2), \quad \frac{\alpha}{(\alpha - 1)^2(\alpha - 2)} (\alpha > 2). \]

19 **Standard Laplace distribution:**

\[ 0, \quad 2, \quad 2. \]

*20 **Standard logistic distribution:**

\[ 0, \quad \sigma^2/3, \quad \sigma^2/3. \]

21 **Standard log-normal\((\tau)\) distribution:**

\[ e^{-\tau^2/2}, \quad e^{\tau^2}, \quad e^{\tau^2}(e^{\tau^2} - 1). \]

22 **Gamma\((\alpha)\) distribution:**

\[ p, \quad (p + 1)p, \quad p. \]

23 **Chi-square\((m)\) distribution:**

\[ m, \quad (m + 2)m, \quad 2m. \]

24 **Beta\((p, q)\) distribution:**

\[ \frac{p}{p + q}, \quad \frac{p(p + 1)}{(p + q)(p + q + 1)}, \quad \frac{pq}{(p + q)^2(p + q + 1)}. \]

25 **Canonical \(F(m, n)\) distribution:**

\[ \frac{m}{n - 2} (n > 2), \quad \frac{m(m + 2)}{(n - 2)(n - 4)} (n > 4), \quad \frac{2m(m + n - 2)}{(n - 2)^2(n - 4)} (n > 4). \]

26 **Canonical Student\((n)\) distribution:**

\[ 0 (n > 1), \quad \frac{1}{n - 2} (n > 2), \quad \frac{1}{n - 2} (n > 2). \]
27 Standard Weibull(\(\beta\)) distribution:

\[ \Gamma(\beta^{-1} + 1), \quad \Gamma(2\beta^{-1} + 1), \quad \Gamma(2\beta^{-1} + 1) - \Gamma^2(\beta^{-1} + 1). \]

For the following distribution, verify the corresponding expressions for

\[ E(y_i), \quad E(y_i^2), \quad \text{var}(y_i), \quad E(y_i y_j), \quad \text{cov}(y_i, y_j), \quad i \neq j. \]

28 Dirichlet(\(r_1, \ldots, r_{k+1}\)) distribution:

\[
\frac{r_i}{\sum r_a} = \frac{(r_i + 1)r_i}{(\sum r_a + 1)\sum r_a}, \quad \frac{r_i(r_{a} - r_i)}{(\sum r_a + 1)\sum r_a},
\]

\[
\frac{r_i r_j}{(\sum r_a + 1)\sum r_a} = \frac{-r_i r_j}{(\sum r_a + 1)\sum r_a}.
\]

J SOME GENERAL PROBLEMS

29 Show that the mean square deviation from \(t\), \(E((x - \mu)^2)\), is a minimum if \(t = \mu\).

30 For a sample from the Bernoulli(\(p\)) distribution, show that the proportion \(\hat{p} = \frac{\sum x_i}{n} = \frac{y}{n}\) has mean \(p\) and variance \(pq/n\). Consider a possibly biased coin: how many times should the coin be tossed to be 95 percent sure that the proportion \(\hat{p}\) is within 0.01 of the probability \(p\)? (Use Chebyshev and note that \(pq\) has maximum value 1/4.)

31 Bivariate normal in Problem 2.4.24: Show that the bivariate normal(\(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho\)) has mean and variance matrix,

\[
\begin{pmatrix}
\mu_1 \\
\mu_2
\end{pmatrix}
\quad \text{and} \quad
\begin{pmatrix}
\sigma_1^2 & \sigma_1 \sigma_2 \rho \\
\sigma_1 \sigma_2 \rho & \sigma_2^2
\end{pmatrix},
\]

and correlation coefficient \(\rho\). See Problem 4.2.37.

32 Consider a distribution for \(x\) on \(R^k\) with variance matrix \(\Sigma\). Show that the variance of \(\sum t_i x_i = t^\top x\) is given by \(t^\top \Sigma t\). Hence deduce that \(\Sigma\) is positive semidefinite; use the results in Supplement Section 2.5.C.

33 Consider a standard normal distribution for \(z\) and define a distribution for \((z_1, z_2)\) on \(R^2\) by \((z_1, z_2) = (z, |z|)\). Calculate the mean and variance matrix for \((z_1, z_2)\). Calculate the correlation coefficient \(\rho_{z_1 z_2}\); are \(z_1\) and \(z_2\) statistically independent?

34 For the uniform distribution over the disk \(x_1^2 + x_2^2 < 4\), show that the mean and variance matrix are

\[
\begin{pmatrix}
0 \\
0
\end{pmatrix}, \quad \begin{pmatrix}
1 & 0 \\
0 & 1
\end{pmatrix}.
\]

Note that these are the same as the standard normal on \(R^2\).

35 (continuation) Show that the bivariate normal in Problem 31 has the same mean and covariance matrix as the uniform distribution over the elliptical disk

\[
\frac{1}{1 - \rho^2} \left[ \frac{(y_1 - \mu_1)^2}{\sigma_1^2} - 2\rho \frac{(y_1 - \mu_1)(y_2 - \mu_2)}{\sigma_1 \sigma_2} + \frac{(y_2 - \mu_2)^2}{\sigma_2^2} \right] < 4.
\]

36 For the uniform distribution over the ball \(\sum x_i^2 < k + 2\), show that the mean and variance matrix are \(0\) and \(I\), where \(0\) is the 0 vector in \(R^k\) and \(I\) is the \(k \times k\) identity matrix. Method:

Use the intuitively reasonable fact that \(r = (x_1^2 + \cdots + x_k^2)^{1/2}\) has a marginal distribution with density \(cr^{k-1}\) on \((0, \sqrt{k + 2})\).
37 \textit{(continuation)} Show that the multivariate normal($\mu, \Sigma$) with $|\Sigma| \neq 0$ as given in Section 2.4.G has the same mean and variance matrix as the uniform distribution on the elliptical ball $(y - \mu)^T \Sigma^{-1} (y - \mu) < k + 2$.

5.4

THE LOCATION AND SCALING OF A SAMPLE

Consider a chemist with a sample of $n$ determinations of the yield for a new isotope process. He would certainly be interested in the mean yield for the process and perhaps in the variance of the process. With the sample of $n$ values in hand he would naturally wonder to what degree the location and scaling for the sample represented the true mean and variance for the process.

In this section we develop the powerful formulas mentioned at the beginning of this chapter, formulas that relate the location and scaling of a sample to the location and scaling of the distribution being sampled. The applications are far-reaching, both in probability theory and in statistics.

We first examine some functions that describe the location and scaling of a sample of $n$ values on the real line $R$, then functions that describe location and scaling on $R^k$. We then consider the distribution of a sample from an initial distribution and investigate some distribution properties of these location and scaling functions.

A SAMPLE LOCATION ON $R$

Consider a sequence $(x_1, \ldots, x_n)$ of values on the real line $R$. A simple measure of the location of this sample is given by the \textit{sample average}

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n} \sum x_i.$$ \hfill (1)

If we consider unit masses at each of the points $x_i$, then $\bar{x}$ is the center of gravity, the point about which the array balances; see Figure 5.8. For we have

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum x_i - n\bar{x} = 0.$$ \hfill (2)

![FIGURE 5.8](image)

Sample $(x_1, x_2, x_3, x_4)$ with average $\bar{x}$; corresponding sample $(y_1, y_2, y_3, y_4)$ with average $\bar{y} = a + c\bar{x}$. 
If we obtain a new sequence \((y_1, \ldots, y_n)\) from an initial sequence \((x_1, \ldots, x_n)\) by an affine transformation \(y = a + cx_i\), then

\[
\bar{y} = \frac{1}{n} \sum (a + cx_i) = a + c\bar{x}.
\]

Now let \((x_1, \ldots, x_n)\) be a sample from a distribution on \(R\) with mean \(\mu\) and variance \(\sigma^2\). We then have a distribution for \(x = (x_1, \ldots, x_n)^T\) on \(R^n\) in which the coordinates are statistically independent and each has the same given distribution with mean \(\mu\) and variance \(\sigma^2\). Consider some characteristics of the marginal distribution of the function \(\bar{x}\); note that \(\bar{x}\) is a function from \(R^n\) to \(R^1\).

From formulas (5.2.4) and (5.3.7) we obtain

\[
E \left( \sum x_i \right) = n\mu, \quad \var \left( \sum x_i \right) = n\sigma^2.
\]

Then from formulas (5.2.3) and (5.3.5), we obtain

\[
E(\bar{x}) = \mu, \quad \var(\bar{x}) = \frac{\sigma^2}{n}.
\]

Note that the distribution of \(\bar{x}\) has the same location \(\mu\) as the distribution being sampled, but its variance is reduced by a factor \(1/n\), and its standard deviation is reduced by a factor \(1/\sqrt{n}\). For an example, see Figures 5.9 and 5.10.

**B Sample Scaling on \(R\)**

Consider again a sequence \((x_1, \ldots, x_n)\) of values on the real line \(R\). The residual vector \((x_1 - \bar{x}, \ldots, x_n - \bar{x})\) shows how the values are spread out with respect to the sample average \(\bar{x}\). First, some geometrical motivation.
Sec. 5.4: The Location and Scaling of a Sample

\[ n = 1 \]
\[ 0 \quad 1 \quad 2 \quad 3 \quad \Sigma x_i \]
\[ 0 \quad \frac{1}{2} \quad \bar{x} \]

\[ n = 2 \]
\[ 0 \quad 1 \quad 2 \quad 3 \quad \Sigma x_i \]
\[ 0 \quad \frac{1}{2} \quad \bar{x} \]

\[ n = 3 \]
\[ 0 \quad 1 \quad 2 \quad 3 \quad \Sigma x_i \]
\[ 0 \quad \frac{1}{2} \quad \bar{x} \]

**FIGURE 5.10**
Distribution of \( \Sigma x_i \) and \( \bar{x} \) for samples of \( n = 1, 2, 3 \) from the Bernoulli(1/2) distribution.

We consider \( x = (x_1, \ldots, x_n)' \) as a point in \( \mathbb{R}^n \); see Figure 5.11. The point \( \bar{x} = (\bar{x}_1, \ldots, \bar{x}_n)' \) is the fitted point with all coordinates equal and yet with coordinates that add to the same total as for the original vector; that is, \( \bar{x} \) lies on \( \mathcal{L}(1) \) and also lies on the plane \( \Sigma x_i = c \) that passes through the original \( x \). The point \( x - \bar{x} \) is the

**FIGURE 5.11**
Vector \( x \): its projection on the extended 1 vector \( \mathcal{L}(1) \) gives the fitted vector \( \bar{x} \); its projection on the orthogonal complement \( \mathcal{L}^\perp(1) \) of the vector 1 gives the residual vector \( x - \bar{x} \). Note that the sum of these two components reconstructs the original vector \( x \).
residual vector or the deviation vector from the fitted \( \bar{x} \). Note that these vectors \( \bar{x} \) and \( x - \bar{x} \) are orthogonal: the inner product is

\[
(\bar{x}, x - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x}) = \bar{x} \sum_{i=1}^{n} (x_i - \bar{x}) = 0.
\]

Also note that they add to give the original vector,

\[
\bar{x} + (x - \bar{x}) = x;
\]

and in addition that their squared lengths add to give the squared length of the original vector,

\[
(6) \quad n\bar{x}^2 + \sum (x_i - \bar{x})^2 = \sum x_i^2,
\]

an easily checked consequence of the orthogonality. Thus we see that \( \bar{x} \) is the projection of \( x \) onto the line \( \mathcal{L}(1) \); that is, it has an orthogonal residual. And we see that \( x - \bar{x} \) is the projection of \( x \) onto the plane \( \mathcal{L}(1) = \{ y : \sum y_i = 0 \} \) consisting of all vectors orthogonal to \( 1 \); that is, it has an orthogonal residual which is just \( \bar{x} \).

We will see that an "average" squared deviation, called the sample variance,

\[
(7) \quad s_x^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = \frac{1}{n-1} \left( \sum x_i^2 - n\bar{x}^2 \right),
\]

provides a very reasonable measure of scaling in squared units, and that root-average-squared deviation, the sample standard deviation, \( s_x \), provides a reasonable measure of scaling in given units; see Figure 5.12.

If we obtain a new sequence \( (y_1, \ldots, y_n) \) from a given sequence \( (x_1, \ldots, x_n) \) by an affine transformation \( y = a + cx \), it is easily checked that

\[
(8) \quad s_y^2 = c^2 s_x^2, \quad s_y = |c| s_x.
\]

Now let \( (x_1, \ldots, x_n) \) be a sample from a distribution on \( R \) with mean \( \mu \) and variance \( \sigma^2 \). We then have a distribution for \( x = (x_1, \ldots, x_n) \) on \( R^n \) with independent coordinates each having the given distribution with mean \( \mu \) and variance \( \sigma^2 \). Consider some characteristics of the marginal distribution of the function \( s_x^2 \): note that \( s_x^2 \) is a function from \( R^n \) to \( R^1 \).

If we replace \( x_i \) by \( x_i - \mu \) in the identity (6) and rearrange, we obtain

\[
\sum (x_i - \bar{x})^2 = \sum (x_i - \mu)^2 - n(\bar{x} - \mu)^2.
\]

Then using \( \text{E}(x_i - \mu)^2 = \text{var}(x_i) = \sigma^2 \) and \( \text{E}(\bar{x} - \mu)^2 = \text{var}(\bar{x}) = \sigma^2/n \), we obtain

\[
(9) \quad \text{E} \left( \sum (x_i - \bar{x})^2 \right) = n\sigma^2 - \frac{n\sigma^2}{n} = (n - 1)\sigma^2.
\]

Thus the mean value of the sample variance is equal to the variance of the distribution being sampled.

\[
\begin{array}{cccccccc}
\bullet & x_1 & \bullet & x_3 & \bullet & x_4 & \bullet & x_5 \\
\bar{x} & & & & & & & \end{array}
\]

**Figure 5.12**
Sample average \( \bar{x} \) and sample standard deviation \( s_x \).
C SAMPLE LOCATION ON $\mathbb{R}^k$

Consider a sequence $(x_1, \ldots, x_n)$ of $n$ points in $\mathbb{R}^k$. For example, $x$ might record (length, 'girth, age') for a fish in a sample of $n$ fish. We can average the $j$th coordinates of these vectors,

$$\bar{x}_j = \frac{x_{j1} + \cdots + x_{jn}}{n} = \frac{1}{n} \sum_{i=1}^{n} x_{ji},$$

and then form the sample average vector,

$$\bar{x} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_k \end{pmatrix} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

as a point in $\mathbb{R}^k$, the sample average vector $\bar{x}$ gives the center of gravity of the $n$ points $x_1, \ldots, x_n$; see Figure 5.13. For the example mentioned earlier $\bar{x}$ records (average length, average girth, average age).

Now let $(x_1, \ldots, x_n)$ be a sample from a distribution on $\mathbb{R}^k$ with mean $\mu$ and variance matrix $\Sigma$. Consider some characteristics of the marginal distribution of the function $\bar{x}$; note that $\bar{x}$ is a function from $(\mathbb{R}^k)^n = \mathbb{R}^{kn}$ into $\mathbb{R}^k$.

From formulas (5.2.12) and (5.3.20) we obtain

$$\begin{align*}
E \left( \sum_{i=1}^{n} x_i \right) &= n \mu, \\
\text{VAR} \left( \sum_{i=1}^{n} x_i \right) &= n \Sigma.
\end{align*}$$

Then from formulas (5.2.11) and (5.3.17), we obtain

$$\begin{align*}
E(\bar{x}) &= \mu, \\
\text{VAR}(\bar{x}) &= n^{-1} \Sigma.
\end{align*}$$

Thus, in particular, $E(\bar{x}_i) = \mu_i$, $\text{var}(\bar{x}_i) = \sigma_{ii}/n$, and $\text{cov}(\bar{x}_1, \bar{x}_2) = \sigma_{12}/n$. Thus the distribution of $\bar{x}$ has the same location as the distribution being sampled, but its variance matrix is reduced by a factor $1/n$.

![Figure 5.13](image-url) Sample $(x_1, x_2, x_3, x_4)$ on the plane. The sample average vector $\bar{x}$. 
D SAMPLE SCALING ON $\mathbb{R}^k$

Consider again a sequence $(x_1, \ldots, x_n)$ of $n$ points in $\mathbb{R}^k$, say the measurements for the $n$ fish. We follow the pattern in Section B; for this we briefly reverse our geometry and consider $k$ points in $\mathbb{R}^n$. From the original first coordinates we form

$$(x_{11}, \ldots, x_{1n}),$$

and then the residual

$$(x_{11} - \bar{x}_1, \ldots, x_{1n} - \bar{x}_1).$$

And from the original second coordinates we form

$$(x_{21}, \ldots, x_{2n})$$

and then the residual

$$(x_{21} - \bar{x}_2, \ldots, x_{2n} - \bar{x}_2);$$

see Figure 5.14. The identity

$$(13) \quad \sum x_{1i}x_{2i} = \sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) + n\bar{x}_1\bar{x}_2$$

is easily checked algebraically. A measure of how the first coordinate varies with respect to the second coordinate is given by an "average" product deviation, the sample covariance,

$$(14) \quad s_{12} = \frac{1}{n - 1} \sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) = \frac{1}{n - 1} \left( \sum x_{1i}x_{2i} - n\bar{x}_1\bar{x}_2 \right).$$

**FIGURE 5.14**

Vectors $(x_{11}, \ldots, x_{1n})$ and $(x_{21}, \ldots, x_{2n})$ in $\mathbb{R}^n$. The corresponding residuals $(x_{11} - \bar{x}_1, \ldots, x_{1n} - \bar{x}_1)$ and $(x_{21} - \bar{x}_2, \ldots, x_{2n} - \bar{x}_2)$ in $\mathbb{L}^\perp(1)$. The angle $\alpha$ between the residuals; the correlation $r_{12}$ is the cosine of the angle $\alpha$. 
Note that the sample covariance of the first coordinate with the first coordinate is just
the sample variance, say $s_{11}$, for that coordinate. The sample covariances can be
assembled as a \textit{sample variance matrix},

\begin{equation}
S = \begin{pmatrix}
  s_{11} & \cdots & s_{1k} \\
  \vdots & & \vdots \\
  s_{k1} & \cdots & s_{kk}
\end{pmatrix}
\end{equation}

\begin{equation}
= \frac{1}{n - 1} \begin{pmatrix}
  \sum (x_{1i} - \bar{x}_1)^2 & \cdots & \sum (x_{1i} - \bar{x}_1)(x_{ki} - \bar{x}_k) \\
  \vdots & & \vdots \\
  \sum (x_{ki} - \bar{x}_k)(x_{1i} - \bar{x}_1) & \cdots & \sum (x_{ki} - \bar{x}_k)^2
\end{pmatrix}
\end{equation}

Note that $(n - 1)S$ is the inner product matrix for the residual vectors.

Now let $(x_1, \ldots , x_n)$ be a sample from a distribution on $R^k$ with mean $\mu$ and
variance matrix $\Sigma$. Consider some characteristics of the marginal distribution of the
sample variance matrix.

If we replace each $x_{1i}$ by $x_{1i} - \mu_1$ and each $x_{2i}$ by $x_{2i} - \mu_2$ in the identity \eqref{eq:13} and
rearrange, we obtain

\begin{equation}
\sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) = \sum (x_{1i} - \mu_1)(x_{2i} - \mu_2) - n(\bar{x}_1 - \mu_1)(\bar{x}_2 - \mu_2).
\end{equation}

Then from $E((x_{1i} - \mu_1)(x_{2i} - \mu_2)) = \sigma_{12}$ and $E((\bar{x}_1 - \mu_1)(\bar{x}_2 - \mu_2) = \text{cov} (\bar{x}_1, \bar{x}_2) = \sigma_{12}/n$, we obtain

\begin{equation}
E(S_{12}) = \sigma_{12}.
\end{equation}

Accordingly, we have

\begin{equation}
E(S) = \Sigma;
\end{equation}

the mean value of the sample variance matrix is the variance matrix of the distribution being sampled.

\section*{Sample Correlation}

Consider a sequence $(x_1, \ldots , x_n)$ of points in $R^k$, and for simplicity take $k = 2$. We follow the pattern in Sections B and D and briefly reverse our geometry and consider two points in $R^2$. From the first coordinates we obtain the deviation vector

$(x_{11} - \bar{x}_1, \ldots , x_{1n} - \bar{x}_1)$,

and from the second coordinates we obtain the deviation vector

$(x_{21} - \bar{x}_2, \ldots , x_{2n} - \bar{x}_2)$.

The cosine of the angle between these two deviation vectors (see Figure 5.14) is called the \textit{sample correlation coefficient} between the first two coordinates:

\begin{equation}
r_{12} = \frac{\sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\left( \sum (x_{1i} - \bar{x}_1)^2 \right)^{1/2} \left( \sum (x_{2i} - \bar{x}_2)^2 \right)^{1/2}} = \frac{s_{12}}{(s_{11}s_{22})^{1/2}}.
\end{equation}

If the two deviation vectors are pointing in almost the same direction, then $r_{12}$ will be close to $+1$; if they are pointing in almost opposite directions, then $r_{12}$ will be close to
FIGURE 5.15
Samples of 100 plotted on the plane $R^2$ with correlations as marked. Samples were obtained by computer from bivariate normal distributions with $\sigma_1 = \sigma_2$ and $\rho = 0, 0.3, 0.5, 0.7, 0.9$. 
The Schwarz inequality (Problem 5.8.17) can be applied to the deviation vectors using a unit measure for each of the \( n \) coordinate positions; we obtain
\[
\left( \sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)^2 \right)^2 \leq \sum (x_{1i} - \bar{x}_1)^2 \sum (x_{2i} - \bar{x}_2)^2
\]
with equality if and only if one deviation vector is a multiple of the other; note equivalent forms of the inequality
\[
s_{12}^2 \leq s_{11}s_{22}, \quad -1 \leq r_{12} \leq 1.
\]

Now we return to the original geometry of a sequence \((x_1, \ldots, x_n)\) of \( n \) points in \( \mathbb{R}^2 \). In practice, with reasonable response vectors, a large positive correlation means a rather compressed array of points tilted up to the right, and a large negative correlation means a distribution with a compressed array of points tilted down to the right; see Figure 5.15.

**F Exercises**

1. For a new sample \((y_1, \ldots, y_n)\) obtained from a given sample \((x_1, \ldots, x_n)\) by the affine transformation \( y = a + cx \), show that \( s_y^2 = c^2s_x^2 \), \( s_y = |c|s_x \).

2. Show that the sum of squares of deviations from \( t \), \( \sum (y_i - t)^2 \), is a minimum if \( t = \bar{y} \).

3. The following sample data were obtained in a chemical determination of sodium content: 17.1, 14.9, 15.2, 18.1, 16.4, 19.1, 18.5, 14.6, 17.3. Calculate the sample mean \( \bar{y} \), the sample variance \( s_y^2 \), and the sample standard deviation \( s_y \).

4. Ten determinations were made in a spectral line, giving the data
\[
65.175 \quad 65.169 \quad 65.172 \quad 65.183 \quad 65.179
\]
\[
65.178 \quad 65.185 \quad 65.180 \quad 65.181 \quad 65.180
\]
Calculate the sample mean \( \bar{y} \), the sample variance \( s_y^2 \), and the sample standard deviation \( s_y \); for easy arithmetic use formulas (3) and (8).

5. The following sample of 5 was obtained on the vector variable \( \begin{pmatrix} x \end{pmatrix} \) where \( x \) is the amount of fertilizer and \( y \) is the yield
\[
\begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}
\]
Calculate the sample mean and sample variance matrix
\[
\begin{pmatrix} x \end{pmatrix} \quad S = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix}
\]
Calculate the sample correlation coefficient \( r_{xy} \). In this example the values of \( x \) were chosen by design; thus there would not be a distribution for \( (x, y) \) on \( \mathbb{R}^2 \).

6. The following sample of 7 was obtained on the vector \( \begin{pmatrix} y_1, y_2 \end{pmatrix} \), where \( y_1 \) is the tensile strength and \( y_2 \) the Brinell hardness of cold-drawn copper.
\[
\begin{pmatrix} 106.1 \\ 106.3 \\ 104.4 \\ 103.8 \\ 101.5 \\ 100.6 \\ 106.2 \end{pmatrix}
\]
Calculate the sample mean, the sample variance matrix, and the sample correlation coefficient. For easier arithmetic, relocate and rescale the \( y \) values, relocate and rescale the
y_2 values, then use appropriate correction formulas—for example, simplified versions of those in Problem 7.

G PROBLEMS

7 For a new sample \((y_1, \ldots, y_n)\) obtained from a given sample \((x_1, \ldots, x_n)\) by the affine transformation \(y = a + Cx\), show that

\[
\bar{y} = a + C\bar{x}, \quad S_{yy} = CS_{xx}C',
\]

where \(S_{yy} (S_{xx})\) is the sample variance matrix for the \(y_i\)'s (\(x_i\)'s).

8 The sequence \((x_1, \ldots, x_n)\) of \(n\) points in \(R^a\) can be presented as a matrix,

\[
X = \begin{pmatrix}
    x_{11} & \cdots & x_{1n} \\
    \vdots & \ddots & \vdots \\
    x_{n1} & \cdots & x_{nn}
\end{pmatrix}
\]

(a) Check that the residual vectors described in Section D are given by the row vectors of

\[
X - \bar{x}1',
\]

where \(1\) is the 1-vector in \(R^n\).

(b) Show that the sample variance matrix is given by

\[
S = (n - 1)^{-1}(X - \bar{x}1')(X - \bar{x}1)' = (n - 1)^{-1}(XX' - n\bar{x}\bar{x}')
\]

9 Show that the sum of squared deviations \(\Sigma (y_i - bx_i)^2\) of \(y\) from a point \(bx\) in \(L(x)\) is minimized by choosing \(b = \Sigma x_iy_i/\Sigma x_i^2\). Method: The expression is quadratic in \(b\); complete the square in \(b\), obtaining an expression \(a(b - b_0)^2 + d\).

10 Let \((u_1, \ldots, u_n)\) be a sample from the uniform\((0, 1)\) distribution. And let \(u\) and \(v\) be the smallest and largest in the sample. The results in Problem 3.6.12 can be specialized to give the following probability density for \((u, v)\):

\[
f(u, v) = n(n - 1)(v - u)^{n-2} \quad \text{if } 0 < u < v < 1
\]

and zero otherwise. Determine the mean and variance matrix of this distribution.

11 (continuation) Show that the interorder statistic distances \(c_1, c_2, \ldots, c_{n+1}\) all have mean \(1/(n + 1)\); a very short proof is possible using Problem 3.6.10. Thus show that the order statistic \(u_{(i)}\) has mean \(i/(n - 1)\).

12 (continuation) The joint density of \(u_{(i)}, u_{(r+s)}\) can be obtained from Problem 3.6.12:

\[
f(u_{(i)}, u_{(r+s)}) = \frac{n!}{(r - 1)!(s - 1)!(n - r - s)!} u_{(i)}^{r-1}(u_{(r+s)} - u_{(i)})^{s-1}(1 - u_{(r+s)})^{n-r-s}
\]

on \(0 < u_{(i)} < u_{(r+s)} < 1\) and zero elsewhere. Determine \(E(u_{(i)}u_{(r+s)})\) and deduce \(cov(u_{(i)}, u_{(r+s)})\). Record the mean and variance matrix for \((u_{(i)}, \ldots, u_{(n)})\).

H PROBLEMS: COMPUTER USAGE

Program the computer to calculate as follows: (a) from an incoming sample of \(n\), calculate \(\bar{y}\) and \(s_y^2\); (b) from a large number \(N\) of repetitions on \(\bar{y}\) and \(s_y^2\), calculate the long-run average \(\hat{\mu}_y = \Sigma \bar{y}/N\) for \(\bar{y}\), the long-run variance \(\hat{\sigma}_y^2 = \Sigma (y - \hat{\mu}_y)^2/(N - 1)\) for \(\bar{y}\), and the long-run average \(\hat{\mu}_{s_y^2} = \Sigma s_y^2/N\) for \(s_y^2\).

13 Obtain a sample of \(n = 10\) from the uniform distribution on \((0, 1, \ldots, 9)\), and calculate \(\bar{y}\) and \(s_y^2\). For \(N = 1000\) repetitions, calculate the long-run averages, \(\hat{\mu}_y, \hat{\sigma}_y^2, \hat{\mu}_{s_y^2}\).
Compare with the theoretical values $E(\bar{y})$, Var $(\bar{y})$, and $E(s^2)$. 

Obtain a sample of 5 from the triangular distribution with $f(y) = 2y$ on $(0, 1)$ and zero elsewhere (use Problem 4.2.35) and calculate $\bar{y}$ and $s^2$. For $N = 1000$ repetitions, calculate $\hat{\mu}_y$, $\hat{\sigma}_y^2$, $\hat{\mu}_s$. Compare with the theoretical values $E(\bar{y})$, var $(\bar{y})$, and $E(s^2)$.

### 5.5

**CONDITIONAL LOCATION AND SCALING**

We noted in Chapter 4 that partial information is sometimes available concerning a realization and that conditional probability then provides the appropriate description. Certainly the location and scaling of such a conditional distribution would be of prime importance. Consider some examples.

An early example involved the measurement of $(x, y)$, where $x$ is the height of a father and $y$ the height of the eldest son. A large batch of data led to a bivariate normal as a reasonable approximation. For a father of height $x$ with a young son, the conditional distribution of $y$ given $x$ would describe the future adult height of the son. The mean and variance of this distribution would be of particular interest.

Consider a distribution describing $(x, y)$, where $x$ is the length and $y$ the age for a fish of a certain species. Length is easy to determine and $y$ is difficult to determine. For a fish of length $x$, the conditional distribution of $y$, given $x$, provides an estimating procedure for age. The mean and variance of this distribution would be of particular importance.

### A CONDITIONAL MEANS AND VARIANCES ON $R^2$

Consider a distribution for $(x, y)$ on the plane $R^2$ with probability function $p$ in the discrete case or probability density function $f$ in the absolutely continuous case. Let $p_1$ or $f_1$ describe the marginal distribution for $x$; and $p_2(\cdot : x)$ or $f_2(\cdot : x)$ describe the conditional distribution for $y$ given $x$.

We examine the conditional distribution of $y$ given $x$; see Figure 5.16. For this distribution, let $\mu_2(x)$ be the mean and $\sigma_2^2(x)$ be the variance; assuming they exist we have

1. $\mu_2(x) = E(y : x) = \sum_y y p_2(y : x)$ if discrete, 
   
   $= \int_{-\infty}^{\infty} y f_2(y : x) \, dy$ absolutely continuous,

2. $\sigma_2^2(x) = \text{var} (y : x) = \sum_y (y - \mu_2(x))^2 p_2(y : x)$ if discrete, 

   $= \int_{-\infty}^{\infty} (y - \mu_2(x))^2 f_2(y : x) \, dy$ absolutely continuous.
Figure 5.16
Mean $\mu_2(x)$ and standard deviation $\sigma_2(x)$ of the conditional distribution of $y$ given $x$. The regression curve of $y$ on $x$ is given by $y = \mu_2(x)$.

Example 1

A distribution on $\mathbb{R}^2$: Consider the distribution for $(x, y)$ with density $f$ as examined in Example 4.2.3:

$$f(x, y) = 6x \quad \text{if } 0 < x < 1, \ 0 < y < 1 - x$$

and zero otherwise. The mean and variance matrix have been calculated in Example 5.3.8. The conditional density for $y$ given $x$ is

$$f_2(y|x) = (1 - x)^{-1} \quad \text{if } 0 < y < 1 - x$$

and zero otherwise; this is the uniform$(0, 1 - x)$ distribution. The mean and variance for the uniform distribution are available from Example 5.3.2; thus

$$\mu_2(x) = \frac{1 - x}{2}, \quad \sigma_2^2(x) = \frac{(1 - x)^2}{12} \quad \text{for } 0 < x < 1.$$  

Suppose that we view $\mu_2(x)$ and $\sigma_2^2(x)$ as functions of $x$. Then we define:

Definition 1

The regression of $y$ on $x$ is the function $\mu_2(x)$ recording the conditional mean for each value of $x$.  

The variance about regression on $x$ is the function $\sigma_2^2(x)$ recording the conditional variance for each value of $x$.

See Figure 5.16.

Now suppose that we view $\mu_2(x)$ as a function of $(x, y)$, a function on $\mathbb{R}^2$ that happens not to depend on the second coordinate $y$. Then we have:

**Lemma 3**

The marginal mean is equal to the mean value of the conditional mean:

(3) \[ \mu_2 = E(\mu_2(x)) \quad \text{or} \quad E(y) = E(E(y \mid x)). \]

**Proof** We record the proof for the absolutely continuous case:

\[
E(\mu_2(x)) = \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} y f_2(y \mid x) \, dy \right) f_1(x) \, dx = \int_{-\infty}^{\infty} y f_2(y \mid x) f_1(x) \, dx \, dy
\]

\[
= \int_{-\infty}^{\infty} y f(x, y) \, dx \, dy = \mu_2.
\]

This lemma says that we can calculate a mean value in two stages: first the conditional mean over a section of the distribution; then the mean of that conditional mean.

Now suppose that we view $\sigma_2^2(x)$ as a function of $(x, y)$, a function on $\mathbb{R}^2$ that happens not to depend on the second coordinate. Then we have:

**Lemma 4**

The marginal variance is equal to the mean conditional variance plus the variance of the conditional mean:

(4) \[ \sigma_2^2 = E(\sigma_2^2(x)) + \text{var}(\mu_2(x)) \]

\[ \text{var}(y) = E(\text{var}(y \mid x)) + \text{var}(E(y \mid x)). \]

**In brief**, variance is mean variance about regression plus variance of regression.

**Proof** By formula (5.3.3) we have $E(w^2) = \text{var}(w) + (E(w))^2$; let $w$ have the conditional distribution of $y - \mu_2$, given $x$:

\[ E((y - \mu_2)^2 \mid x) = \text{var}(y \mid x) + (E(y \mid x) - \mu_2)^2. \]

Then, using Lemma 3, we obtain

\[ \text{var}(y) = E((y - \mu_2)^2) = E(E((y - \mu_2)^2 \mid x)) = E(\text{var}(y \mid x)) + \text{var}(E(y \mid x)). \]

For an interpretation, see Figure 5.16 again. The term $\text{var}(y)$ is the variance along the second axis when all the probability is projected onto that axis. Now suppose that the probability is collapsed along $x$ sections and placed on the curve $\mu_2(x)$; in doing this, we relinquish $E(\text{var}(y \mid x))$, the mean variance about regression. And, indeed, we are left with just the variance along the second axis of this collapsed distribution.
B SOME EXAMPLES ON $R^2$

EXAMPLE 2

The multinomial($n, p_1, p_2, p_3$) distribution: Consider the multinomial($n, p_1, p_2, p_3$) distribution for ($y_1, y_2, y_3$). As noted before $y_3 = n - y_1 - y_2$; thus, effectively, we have a distribution for ($y_1, y_2$) on $R^2$. The marginal distribution for $y_1$ is binomial($n, p_1$); see Problem 3.5.37. The conditional distribution for $y_2$ given $y_1$ is binomial($n - y_1, p_2/(p_2 + p_3)$); see Example 4.2.2. The mean and variance of the binomial distribution are given in Examples 5.2.6 and 5.3.5; thus

\[ \mu_2(y_1) = (n - y_1) \frac{p_2}{p_2 + p_3}, \quad \sigma_2^2(y_1) = (n - y_1) \frac{p_2}{p_2 + p_3} \frac{p_3}{p_2 + p_3}. \]

The regression of $y_2$ on $y_1$ is linear, decreasing to zero at $y_1 = n$. The variance about regression decreases to zero as $y_1$ increases to the maximum value $n$. See Figure 5.17.

![Figure 5.17](image)

Regression of $y_2$ on $y_1$ for the multinomial($n, p_1, p_2, p_3$).

EXAMPLE 3

The bivariate normal distribution: Consider the bivariate normal($\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$) for ($y_1, y_2$); see Problem 4.2.37. The marginal for $y_1$ is normal($\mu_1, \sigma_1$); and the conditional of $y_2$ given $y_1$ is normal($\mu_2 + \rho \sigma_2 (y_1 - \mu_1) / \sigma_1, \sigma_2 (1 - \rho^2)^{1/2}$). Thus we obtain

\[ \mu_2(y_1) = \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (y_1 - \mu_1), \quad \sigma_2^2(y_1) = \sigma_2^2 (1 - \rho^2); \]

see Figure 5.18. The regression of $y_2$ on $y_1$ is linear, tilted up to the right if $\rho$ is positive and down to the right if $\rho$ is negative. The variance about regression is constant; note that $\rho^2$ is the proportion of $\sigma_2^2$ attributable to regression and $1 - \rho^2$ is the proportion attributable to variation about regression. Thus if $\rho^2$ is large, the distribution is very tight about the regression line; compare with the remarks in Section 5.3F; see Figure 5.18.
C  CONDITIONAL MEANS AND VARIANCES ON $\mathbb{R}^{k+l}$

Consider a distribution on $\mathbb{R}^{k+l}$. For notation, let

$$y' = (y_1, \ldots, y_k, y_{k+1}, \ldots, y_{k+l}) = (y'_1, y'_2),$$

where we have partitioned the vector into two component vectors corresponding to the first $k$ and the last $l$ coordinates.

We examine the conditional distribution of the last $l$ coordinates as given by $y_2$ given the first $k$ coordinates as given by $y_1$. For this distribution let $\mu_2(y_1)$ be the mean and $\Sigma_{22}(y_1)$ be the variance matrix, assuming they exist:

$$\mu_2(y_1) = E(y_2 : y_1) = \sum_{y_2} y_2 \rho_{k+1, \ldots, k+l}(y_2 : y_1) \quad \text{if discrete},$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} y_2 \rho_{k+1, \ldots, k+l}(y_2 : y_1) \, dy_2 \quad \text{absolutely continuous},$$

$$\Sigma_{22}(y_1) = \text{VAR}(y_2 : y_1) = E((y_2 - \mu_2(y_1))(y_2 - \mu_2(y_1))' : y_1).$$

Suppose that we view $\mu_2(y_1)$ and $\Sigma_{22}(y_1)$ as functions of $y_1$. Then we define

**DEFINITION 5**

The **regression** of $y_2$ on $y_1$ is the function $\mu_2(y_1)$ recording the conditional mean for each value of $y_1$.

**DEFINITION 6**

The **variance matrix about regression** on $y_1$ is the matrix-valued function $\Sigma_{22}(y_1)$ recording the conditional variance matrix for each value of $y_1$.

Now suppose that we view $\mu_2(y_1)$ as a function of $y$, a function on $\mathbb{R}^{k+l}$ that happens not to depend on the last $l$ coordinates. Then we have:
LEMMA 7

The marginal mean is equal to the mean value of the conditional mean:

\[ \mu_2 = E(\mu_2(y_1)) \text{ or } E(y_2) = E(E(y_2 : y_1)). \]

Proof: Essentially the same as for Lemma 3.

Now suppose that we view \( \Sigma_{22}(y_1) \) as a function of \( y \), a function on \( \mathbb{R}^{k+l} \) that happens not to depend on the last \( l \) coordinates. Then we have

LEMMA 8

The marginal variance matrix is equal to the mean conditional variance matrix plus the variance matrix of the conditional mean:

\[ \Sigma_{22} = E(\Sigma_{22}(y_1)) + \text{VAR} (\mu_2(y_1)) \]

\[ \text{VAR} (y_2) = E(\text{VAR} (y_2 : y_1)) + \text{VAR} (E(y_2 : y_1)). \]

In brief, \text{VARiance is mean \text{VARI}ance about regression plus \text{VAR}iance of regression.}

Proof: By formula (5.3.14) we have

\[ E(ww') = \text{VAR} (w) + E(w)E(w'); \]

let \( w \) have the conditional distribution of \( y_2 - \mu_2 \) given \( y_1 \):

\[ E((y_2 - \mu_2)(y_2 - \mu_2)' : y_1) = \text{VAR} (y_2 : y_1) + (\mu_2(y_1) - \mu_2)(\mu_2(y_1) - \mu_2)'. \]

Then, using Lemma 7, we obtain

\[ \text{VAR} (y_2) = E((y_2 - \mu_2)(y_2 - \mu_2)') = E(\Sigma_{22}(y_1)) + \text{VAR} (\mu_2(y_1)). \]

D AN EXAMPLE ON \( \mathbb{R}^{k+l} \)

EXAMPLE 4

Multivariate normal(\( \mu; \Sigma \)) distribution: Consider the multivariate normal distribution on \( \mathbb{R}^{k+l} \) with mean and variance matrix,

\[ \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \]

partitioned corresponding to the first \( k \) and the last \( l \) coordinates. By Problem 4.2.38 the multivariate normal density can be factored into a normal(\( \mu_1; \Sigma_{11} \)) density for \( y_1 \) and a normal(\( \mu_2 + \Sigma_{21} \Sigma_{11}^{-1} (y_1 - \mu_1); (\Sigma_{22}^{-1}) \)) density for \( y_2 \) given \( y_1 \); note that \( \Sigma_{22} \) is a component of the partitioned inverse matrix

\[ \Sigma^{-1} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}. \]

It follows then that

\[ \mu_2(y_1) = \mu_2 + \Sigma_{21} \Sigma_{11}^{-1} (y_1 - \mu_1), \quad \Sigma_{22}(y_1) = (\Sigma_{22}^{-1}). \]
Note that the regression surface is an affine function of $y_1$, and that the variance about regression is constant.

Lemma 7 gives the obvious

$$\mu_2 = \mu_2 + \Sigma_{21}^{-1} (\mu_1 - \mu_1).$$

Lemma 8 gives the less-than-obvious

\begin{equation}
(12) \quad \Sigma_{22} = (\Sigma^{22})^{-1} + \text{VAR}(\mu_2(y_i)) = (\Sigma^{22})^{-1} + \Sigma_{21}^{-1} \Sigma_{11}^{-1} \Sigma_{12}
\end{equation}

$$= (\Sigma^{22})^{-1} + \Sigma_{21}^{-1} \Sigma_{11}^{-1} \Sigma_{12}$$

or the equivalent,

\begin{equation}
(13) \quad (\Sigma^{22})^{-1} = \Sigma_{22} - \Sigma_{21}^{-1} \Sigma_{11}^{-1} \Sigma_{12}.
\end{equation}

This is a property of partitioned positive definite matrices and can be checked directly in a mechanical way; see Problem 7.

---

**E EXERCISES**

1. Calculate $\mu_1(y)$ and $\sigma_1^2(y)$ for Example 1.
2. Verify that $\mu_2 = E(\mu_2(x))$ and $\sigma_2^2 = E(\sigma_2^2(x)) + \text{var}(\mu_2(x))$ hold for Example 1.
3. Consider the uniform distribution on the disk $x_1^2 + x_2^2 < 1$; see Problem 4.2.12. Determine $\mu_2(x_i)$ and $\sigma_2^2(x_i)$.
4. Let $u$ and $v$ be the smallest and largest in a sample $(u_1, \ldots, u_n)$ from the uniform $(0, 1)$; see Problem 5.4.10. For the distribution of $(u, v)$, calculate $\mu_2(u)$ and $\sigma_2^2(u)$.

---

**F PROBLEMS**

5. Consider the order statistic $(u_{(1)}, \ldots, u_{(n)})$ for a sample from the uniform $(0, 1)$. The distribution of $(u_{(1)}, u_{(n)})$ is recorded in Problem 5.4.12. Determine the conditional mean and variance of $u_{(n)}$ given $u_{(1)}$.

6. Consider the Dirichlet distribution; see Problems 4.2.21 and 5.3.28. Determine the conditional mean and variance matrix given $(y_1, \ldots, y_p)$.

7. Consider $\Sigma$ and $\Sigma^{-1}$ as partitioned in Example 4. Multiply out $\Sigma^{-1} = I$ by partitioned matrix multiplication. Use the resulting equations to show that

\begin{equation}
(\Sigma^{22})^{-1} = \Sigma_{22} - \Sigma_{21}^{-1} \Sigma_{11}^{-1} \Sigma_{12}.
\end{equation}

8. **Kolmogorov inequality**: An extension of the Chebyshev inequality in Section 5.3C. Let $(y_1, \ldots, y_n)$ be nonnegative and suppose that $E(y_i; y_1, \ldots, y_i) \geq y_r$ for $k > r$; then

$$P(\max \{y_1, \ldots, y_n\} \geq t) \leq \frac{E(y_n)}{t}$$

for any $t > 0$. Method: Let $A_j$ be the event $\{y_1 \geq t, y_i < t, \ldots, y_{j-1} < t\}$; then $\{A_1, \ldots, A_n\}$ is a partition of $\{\max \{y_1, \ldots, y_n\} \geq t\}$; examine the integral of $y_n$ over $U A_j$.

9. **(continuation)** An extension of the Chebyshev inequality in Section 5.3C. Let $x_1, \ldots, x_n$ be statistically independent with $E(x_i) = 0$ and $\text{var}(x_i) = \sigma_i^2$; and let $S_k = x_1 + \cdots + x_k$.

$$L_n = \max \{|S_1|, \ldots, |S_n|\}.$$ Then

$$P(L_n \geq t) \leq \frac{E(S_n^2)}{t^2} = \frac{\sum_{i=1}^{n} \sigma_i^2}{t^2}.$$
Method: From Problem 8, in the way that the second Chebyshev inequality was obtained from the first Chebyshev inequality in Section 5.3.

10 Consider a distribution for \( y' = (y_1', y_2') = (y_1, \ldots, y_k, y_{k+1}, \ldots, y_{k+l}) \) on \( R^{k+l} \). Let \( \mu' = (\mu'_1, \mu'_2) \) be the mean and

\[
\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}
\]

be the variance matrix as partitioned in Example 4. Assume that \( \Sigma_{11} \) is positive definite. The linear regression of \( y_2 \) on \( y_1 \) is defined by the affine function,

\[ m_2 = \mu_2 + B(y_1 - \mu_1), \]

of \( y_1 \) where \( B \) is \( k \times k \) and is chosen to "minimize" the variance matrix of the residual \( y_2 - m_2 \); the residual is given by

\[ y_2 - m_2 = y_2 - \mu_2 - B(y_1 - \mu_1). \]

(a) Verify that the variance of the residual is

\[
\text{VAR} (y_2 - m_2) = \Sigma_{22} - B\Sigma_{12} - \Sigma_{21}B' + B\Sigma_{11}B'
\]

\[ = (B - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})(B - \Sigma_{21}\Sigma_{11}^{-1})' + (\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}). \]

(b) The first matrix expression on the last line is positive semidefinite but can be made zero by choosing \( B = \Sigma_{21}\Sigma_{11}^{-1} \); this gives the following variance for the residual from linear regression:

\[
\text{VAR} (y_2 - m_2) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.
\]

Deduce that this variance matrix is the zero matrix if and only if \( y_2 \) is an affine function of \( y_1 \) with probability 1.

5.6

MOMENTS OF A DISTRIBUTION

The mean and standard deviation give some indication as to how probability is distributed along the real line \( R \). Perhaps by having further characteristics, we can pin down more precisely how the probability is distributed—perhaps even pin it down exactly and determine the distribution. In this section we consider some common sequences of characteristics. The results can be useful in many areas of applications—they provide a powerful method for determining needed marginal distributions.

A THE MOMENT SEQUENCE

The mean \( \mu = E(x) \) of a distribution on the real line is sometimes called the first moment, by analogy with the moment of the force of gravity acting on unit mass. In a similar way, \( E(x^2) \) is called the second moment and corresponds to the moment of inertia. We generalize this notion:

**DEFINITION 1**

The \( r \)th moment of a distribution for \( x \) with probability measure \( P \) is
\[ \mu_r = E(x^r) = \int_{-\infty}^{\infty} x^r \, dP(x) \quad r = 0, 1, 2, \ldots, \]
\[ = \sum_{-\infty}^{\infty} x^r \rho(x) \quad \text{if discrete,} \]
\[ = \int_{-\infty}^{\infty} x^rf(x) \, dx \quad \text{absolutely continuous,} \]

provided that the integral exists; otherwise, we say that the moment does not exist.

Note that \( \mu_0 = 1 \) and \( \mu_1 = \mu \), the ordinary mean. Also note that if \( \mu_r \) exists, then \( \mu_s \) exists for \( s \leq r \); this follows from \( |x^s| \leq \max(1, |x^r|) \) and from Problem 5.8.16 in Supplement Section 5.8K.

It is convenient to assemble the moments and form a moment sequence or vector:

**Definition 2**

The moment sequence of a distribution on the real line is given by

\[ \mathbf{m} = (1, \mu_1, \mu_2, \ldots)' \]

and the truncated moment sequence by

\[ \mathbf{m}_r = (1, \mu_1, \mu_2, \ldots, \mu_r)' \]

**Example 1**

The uniform\((0, 1)\) distribution: For the uniform distribution, we have

\[ \mu_r = \int_{-\infty}^{\infty} x^r f(x) \, dx = \int_{0}^{1} x^r \, dx = \frac{1}{r + 1}; \]

thus

\[ \mathbf{m} = (1, 1/2, 1/3, \ldots)', \quad \mathbf{m}_r = (1, 1/2, 1/3, \ldots, 1/(r + 1))'. \]

Suppose that a distribution for \( y \) is obtained from a distribution for \( x \) by the affine transformation \( y = a + cx \). Then the moments \( \nu_r \) for \( y \) are easily calculated from the moments \( \mu_r \) for \( x \); for example,

\[ \nu_3 = E(y^3) = E((a + cx)^3) = E(a^3 + 3a^2cx + 3ac^2x^2 + c^3x^3) \]
\[ = a^3 + 3a^2c\mu_1 + 3ac^2\mu_2 + c^3\mu_3. \]

These relations can be presented easily by means of matrix multiplication:

\[ \begin{pmatrix} 1 \\ \nu_1 \\ \nu_2 \\ \nu_3 \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots \\ a & c & 0 & 0 & \cdots \\ a^2 & \left( \begin{array}{c} 2 \\ 1 \end{array} \right) & ac & c^2 & 0 \\ a^3 & \left( \begin{array}{c} 3 \\ 1 \end{array} \right) & a^2c & \left( \begin{array}{c} 3 \\ 2 \end{array} \right) & ac^2 & c^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \vdots \end{pmatrix}. \]
note the terms of the binomial expansions of \((a + c), (a + c)^2, (a + c)^3, \ldots\).

Now suppose that \(x\) and \(y\) are independent and we want the moment sequence for the sum \(x + y\). Let \(\mu\) and \(\nu\) refer to the moments of \(x\) and \(y\), respectively; then

\[
E((x + y)^r) = E\left( x^r + \binom{r}{1} x^{r-1} y + \cdots + y^r \right)
= \mu^r \nu_0 + \binom{r}{1} \mu_{r-1} \nu_1 + \cdots + \mu_0 \nu_r;
\]

note that we have used "mean product is product mean" given independence; see formula (5.1.13).

**B THE CENTRAL MOMENT SEQUENCE**

The ordinary moments change drastically under a simple location change of a distribution. We can avoid this by examining moments of the deviation \(x - \mu\) from the center of the distribution:

**DEFINITION 3**

The \(r\)th central moment of a distribution for \(x\) with probability measure \(P\) is

\[
\bar{\mu}_r = E((x - \mu)^r) = \int_{-\infty}^{\infty} (x - \mu)^r dP(x) \quad r = 0, 1, 2, \ldots,
\]

\[
= \sum (x - \mu)^r \rho(x) \quad \text{if discrete},
\]

\[
= \int_{-\infty}^{\infty} (x - \mu)^r f(x) \, dx \quad \text{absolutely continuous},
\]

where \(\mu = E(x)\), provided that the integral exists.

Note that \(\bar{\mu}_0 = 1, \bar{\mu}_1 = 0\), and \(\bar{\mu}_2 = \sigma^2\).

**DEFINITION 4**

The central moment sequence of a distribution on the real line is given by

\[
\bar{m} = (1, 0, \bar{\mu}_2, \bar{\mu}_3, \ldots)'
\]

and the truncated central moment sequence by

\[
\bar{m}_r = (1, 0, \bar{\mu}_2, \bar{\mu}_3, \ldots, \bar{\mu}_r)'.
\]

**EXAMPLE 2**

Uniform(0, 1) distribution: The uniform(0, 1) distribution has mean 1/2; thus the deviation \(y = x - 1/2\) is uniform(-1/2, 1/2). The \(r\)th central moment of \(x\) is

\[
E((x - 1/2)^r) = \int_{-1/2}^{1/2} y^r \, dy = \frac{(1/2)^{r+1}}{r + 1} = \frac{1}{(r + 1)2^r} \quad r \text{ even},
\]

\[
= 0 \quad r \text{ odd}.
\]
Thus,
\[ \overline{m} = (1, 0, 1/3 \cdot 2^2, 0, 1/5 \cdot 2^4, \ldots). \]

We can easily see that a location change from \( x \) to \( x + a \) does not affect the central moments, and that a scale change from \( x \) to \( cx \) changes \( \overline{\mu}_r \) to \( c^r \overline{\mu}_r \).

The ordinary moments can be calculated from the central moments plus the mean. We use the transformation \( x = \mu + (x - \mu) \) and equation (4):

\[
\begin{pmatrix}
1 \\
\mu \\
\mu_2 \\
\mu_3 \\
\vdots
\end{pmatrix}
= 
\begin{pmatrix}
1 & 0 & 0 & 0 & \ldots \\
\mu & 1 & 0 & 0 & 0 \\
\mu^2 & \binom{2}{1} \mu & 1 & 0 \\
\mu^3 & \binom{3}{1} \mu^2 & \binom{3}{2} \mu & 1 \\
\vdots & \vdots & \vdots & \vdots & \ddots
\end{pmatrix}
\begin{pmatrix}
1 \\
\overline{\mu}_2 \\
\overline{\mu}_3 \\
\vdots
\end{pmatrix}.
\]

The central moments can be calculated from the ordinary moments. We use the transformation \( x - \mu = -\mu + x \) and equation (4):

\[
\begin{pmatrix}
1 \\
0 \\
-\mu_2 \\
-\mu_3 \\
\vdots
\end{pmatrix}
= 
\begin{pmatrix}
1 & 0 & 0 & 0 & \ldots \\
-\mu & 1 & 0 & 0 \\
\mu^2 & -\binom{2}{1} \mu & 1 & 0 \\
-\mu^3 & \binom{3}{1} \mu^2 & -\binom{3}{2} \mu & 1 \\
\vdots & \vdots & \vdots & \vdots & \ddots
\end{pmatrix}
\begin{pmatrix}
1 \\
\mu_2 \\
\mu_3 \\
\vdots
\end{pmatrix}.
\]

Note that the truncated moment sequence \( m_r \) can be obtained from the mean \( \mu \) and the truncated central moment sequence \( \overline{m}_r \), and conversely. Thus we see that \( m_r \) and \( (\mu, \overline{m}_r) \) provide the same information concerning the distribution.

Also note the simple coefficients in the two matrices, which indeed are inverses of each other.

**EXAMPLE 3**

**Standard normal distribution:** Consider the standard normal distribution for \( z \):

\[ \mu_r = E(z^r) = \int_{-\infty}^{\infty} z^r \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \, dz. \]

If \( r \) is odd, then \( \mu_r = 0 \). If \( r \) is even, then the integral can be adjusted into a gamma integral; alternatively, integration by parts gives

\[ \int_0^\infty z^{2m} e^{-z^2/2} \, dz = -\int_0^\infty z^{2m-1} \, de^{-z^2/2} = (2m - 1) \int_0^\infty z^{2m-2} e^{-z^2/2} \, dz; \]

thus \( \mu_{2m} = (2m - 1)(2m - 3) \ldots 3.1 \). Hence

\[ m = (1, 0, 1, 0, 3, 0, 5 \cdot 3, 0, 7 \cdot 5 \cdot 3, \ldots). \]
C THE FACTORIAL MOMENT SEQUENCE

A modification of the ordinary moments has substantial computational advantage for the common discrete distributions that have factorials in the denominator of the probability function.

DEFINITION 5

The $r$th factorial moment of a distribution for $x$ with probability measure $P$ is

\[
\mu^{(r)} = E(x^{(r)}) = \int_{-\infty}^{\infty} x(x - 1) \cdots (x - r + 1) \, dP(x), \quad r = 0, 1, \ldots,
\]

\[
= \sum x^{(r)} p(x) \quad \text{if discrete},
\]

\[
= \int_{-\infty}^{\infty} x^{(r)} f(x) \, dx \quad \text{absolutely continuous},
\]

provided that the integral exists.

DEFINITION 6

The factorial moment sequence of a distribution on the real line is given by

\[
m^* = (1, \mu^{(1)}, \mu^{(2)}, \mu^{(3)}, \ldots).
\]

The factorial moments can be calculated from the ordinary moments; we use the relations $x^{(1)} = x$, $x^{(2)} = x(x - 1) = x^2 - x$, $x^{(3)} = x^3 - 3x^2 + 2x$, \ldots:

\[
\begin{pmatrix}
1 \\
\mu \\
\mu^{(2)} \\
\mu^{(3)} \\
\vdots
\end{pmatrix}
= 
\begin{pmatrix}
1 & 0 & 0 & 0 & \ldots \\
0 & 1 & 0 & 0 \\
0 & -1 & 1 & 0 \\
0 & 2 & -3 & 1 \\
\vdots & \vdots & \vdots & \vdots
\end{pmatrix}
\begin{pmatrix}
1 \\
\mu \\
\mu^{(2)} \\
\mu^{(3)} \\
\vdots
\end{pmatrix}
\]

The ordinary moments can be calculated from the factorial moments:

\[
\begin{pmatrix}
1 \\
\mu \\
\mu^{(2)} \\
\mu^{(3)} \\
\vdots
\end{pmatrix}
= 
\begin{pmatrix}
1 & 0 & 0 & 0 & \ldots \\
0 & 1 & 0 & 0 \\
0 & 1 & 1 & 0 \\
0 & 1 & 3 & 1 \\
\vdots & \vdots & \vdots & \vdots
\end{pmatrix}
\begin{pmatrix}
1 \\
\mu \\
\mu^{(2)} \\
\mu^{(3)} \\
\vdots
\end{pmatrix}
\]

the coefficients can be obtained by simple rules (Problem 9).

Note that the truncated moment sequence $m_r$ can be obtained from the truncated factorial moment sequence $m^{*}_r$, and conversely.

Thus, in summary, we have that $m_r$ ($\mu, \overline{m}_r$), and $m^{*}_r$ provide the same information concerning the distribution.
EXAMPLE 4

The Poisson(\(\lambda\)) distribution: Consider the factorial moments of the Poisson(\(\lambda\)) distribution:

\[
\mu^{(r)} = E(y^{(r)}) = \sum_{y=0}^{\infty} y^{(r)} \frac{\lambda^y e^{-\lambda}}{y!} = \sum_{y=r}^{\infty} \frac{\lambda^y e^{-\lambda}}{(y-r)!} = \lambda^r,
\]

\(m^* = (1, \lambda, \lambda^2, \lambda^3, \ldots)\).

\(\sigma^2 = \mu_{(2)} - \lambda^{(2)} = \lambda^2 - \lambda(\lambda - 1) = \lambda\).

D. MOMENTS ON \(R^k\)

Consider a distribution for \(x\) on \(R^k\). The \((r_1, \ldots, r_k)\)th ordinary moment is

\[
\mu_{r_1, \ldots, r_k} = E(x_{r_1}^{1} \cdots x_{r_k}^{k}), \quad r_i = 0, 1, 2, \ldots;
\]

in particular, \(\mu_{10\cdots0} = \mu_1, \ldots, \mu_{0\cdots01} = \mu_k\). The \((r_1, \ldots, r_k)\)th central moment is

\[
\bar{\mu}_{r_1, \ldots, r_k} = E((x_1 - \mu_1)^{r_1} \cdots (x_k - \mu_k)^{r_k}), \quad r_i = 0, 1, 2, \ldots;
\]

in particular, \(\bar{\mu}_{10\cdots0} = \cdots = \bar{\mu}_{0\cdots01} = 0\), and \(\bar{\mu}_{20\cdots0} = \sigma_1^2, \ldots, \bar{\mu}_{0\cdots02} = \sigma_k^2\), and \(\bar{\mu}_{110\cdots0} = \sigma_{12}\). And the \((r_1, \ldots, r_k)\)th factorial moment is

\[
\mu_{r_1, \ldots, r_k}^{*} = E(x_{r_1}^{(1)} \cdots x_{r_k}^{(k)}), \quad r_i = 0, 1, 2, \ldots.
\]

The ordinary moments provide the same information concerning the distribution as the central moments plus the mean, and the same as the factorial moments.

E. EXERCISES

1. Calculate the \(r\)th moment of the triangular distribution with density \(f(x) = 2x\) on \((0, 1)\), and zero otherwise.
2. Calculate the \(r\)th moment of the uniform\((a, b)\) distribution.
3. Calculate the \(r\)th moment and the \(r\)th central moment for the Bernoulli\(\(p\)) distribution.
4. Calculate the central moment sequence for the normal\((0, \sigma)\).

F. PROBLEMS

Additional problems on the calculation of moments will be found at the end of Section 5.7.

5. For the standard normal distribution, calculate \(E(x^{2m})\) by adjusting the integral into the form of a gamma function.

6. Determine the \(r\)th factorial moment for the uniform distribution on the integers 1, \ldots, \(N\). Method: Use the easily verified relation

\[
\sum_{x=a}^{b} x^m = \frac{1}{r+1} \left((b+1)^{r+1} - a^{r+1}\right).
\]

7. Let \((x_1, \ldots, x_n)\) be a sample from a distribution with moments 1, \(\mu_1, \mu_2, \ldots\). The sample mean \(\bar{x}\) is an analog of \(\mu_1\); and the sample \(r\)th moment \(m_r = \Sigma x_i^r/n\) is an analog of \(\mu_r\). We know that \(E(\bar{x}) = \mu_1, \text{var}(\bar{x}) = \sigma^2/n\). Show that \(E(m_r) = \mu_r\) and \(\text{var}(m_r) = (\mu_{2r} - \mu_r^2)/n\).
(continuation) The sample rth central moment is \( \bar{m}_r = \Sigma(x_i - \bar{x})^r/n \). Express the ordinary moments \( 1, m_1, m_2, \ldots \) in terms of the central moments \( 1, \bar{m}_1, \bar{m}_2, \ldots \), and \( \bar{x} \), and conversely. Show that \( E\bar{m}_2 \neq \mu_2 \), except in a degenerate case.

9

The connection between ordinary and factorial moments. Let \( \Delta \) be the finite-difference operator and \( \bar{E} = 1 + \Delta \) be the advancing operator:

\[ \Delta f(x) = f(x+1) - f(x) = (\bar{E} - 1)f(x), \]
\[ \Delta^2 f(x) = f(x+2) - 2f(x+1) + f(x) = (\bar{E} - 1)^2f(x), \]
\[ \Delta^r f(x) = \Delta \cdot \Delta \ldots \Delta f(x) = (\bar{E} - 1)^rf(x) = f(x + r) - \left( \begin{pmatrix} r \\ 1 \end{pmatrix} \right)f(x + r - 1) + \ldots + \left( \begin{pmatrix} r \\ 1 \end{pmatrix} \right)f(x). \]

\[ f(x) = (1 + \Delta)^r f(0) = f(0) + x^{(1)} \Delta f(0) + \frac{x^{(2)}}{2!} \Delta^2 f(0) + \frac{x^{(3)}}{3!} \Delta^3 f(0) + \ldots. \]

If \( f \) is a polynomial of the \( r \)th degree, then the series will terminate at the \( r \)th difference; the series is a polynomial that agrees with the original polynomial at the integers, and hence for all \( x \). Thus

\[ x^r = x^{(1)} \Delta 0^r + \frac{x^{(2)}}{2!} \Delta^2 0^r + \ldots + \frac{x^{(r)}}{r!} \Delta^r 0^r, \]

where, for example, \( \Delta^2 0^r = 2^r - 2.1^r + 0^r \) is the second difference of \( x^r \) at \( x = 0 \); and we obtain

\[ \mu_r = \Delta 0^r \mu_{(1)} + \frac{\Delta^2 0^r}{2!} \mu_{(2)} + \ldots + \frac{\Delta^r 0^r}{r!} \mu_{(r)}. \]

Show that

\[ \Delta^r 0^r = n(\Delta^{r-1} 0^{r-1} + \Delta^r 0^{r-1}). \]

and extend the following table to \( r = 4 \):

<table>
<thead>
<tr>
<th>( r )</th>
<th>( \Delta 0^r )</th>
<th>( \Delta^2 0^r )</th>
<th>( \Delta^3 0^r )</th>
<th>( \ldots )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

5.7

MOMENT-GENERATING AND CHARACTERISTIC FUNCTIONS

As we have noted, the mean and standard deviation give some indication as to how probability is distributed along the real line \( R \). Moment sequences were defined with the hope that additional characteristics could pin down more precisely how the probability is distributed. We now develop an analytically more useful way of presenting the information content of a moment sequence. And we discuss how this information content pins down exactly how the probability is distributed on the line \( R \). Again, the results can be useful in many areas of applications—they provide a powerful method for determining needed marginal distributions.
A MOMENT-GENERATING FUNCTIONS

Consider a distribution for $x$ on the real line $\mathbb{R}$ with probability measure $P$. To calculate the moments, we take the mean, $E$, of

$$1 \quad x \quad x^2 \quad x^3 \ldots$$

To do this collectively, we could take the mean, $E$, of

$$e^{xt} = 1 + xt + x^2 \frac{t^2}{2!} + x^3 \frac{t^3}{3!} + \cdots$$

DEFINITION 1

The moment-generating function $m$ of a distribution for $x$ with probability measure $P$ is given by

$$m(t) = E(e^{xt}) = \int_{-\infty}^{\infty} e^{xt} \, dP(x)$$

$$= \sum_{x} e^{xt} p(x) \quad \text{if discrete,}$$

$$= \int_{-\infty}^{\infty} e^{xt} f(x) \, dx \quad \text{absolutely continuous}$$

for those values of $t$ for which the integral exists. The function $m$ is said to exist if the integral (1) exists for $t$ in an open interval containing the origin $0$.

The integral can be shown to exist for an interval containing $0$; if the distribution has a lot of probability in the tails, then the integral may exist only for the trivial $t = 0$. In analysis, $m$ is called the Laplace transform of the distribution $P$.

A closely related complex-valued function always exists:

DEFINITION 2

The characteristic function $c$ of a distribution for $x$ with probability measure $P$ is given by

$$c(t) = E(e^{itx}) = E(\cos xt + i \sin xt)$$

$$= \int_{-\infty}^{\infty} \cos xt \, dP(x) + i \int_{-\infty}^{\infty} \sin xt \, dP(x)$$

for real values of $t$.

The integrands $\cos xt$, $\sin xt$ are bounded by the function 1, a function that can be integrated with respect to the probability measure $P$. Then from Supplement Problem 5.8.16 it follows that the characteristic function exists for all real values of $t$. In analysis, the function $c$ is called the Fourier transform of the distribution $P$. 
EXAMPLE 1

The binomial\((n, p)\) distribution: The moment-generating function for the binomial distribution is

\[ m(t) = \sum_{0}^{n} e^{yt} \binom{n}{y} p^{y} q^{n-y} = \sum_{0}^{n} \binom{n}{y} (pe^{t})^{y} q^{n-y} = (pe^{t} + q)^{n}. \]

Similarly, the characteristic function is

\[ c(t) = (pe^{it} + q)^{n}. \]

EXAMPLE 2

The standard normal: The moment-generating function is

\[ m(t) = \int_{-\infty}^{\infty} e^{zt} \frac{1}{\sqrt{2\pi}} e^{-z^{2}/2} \, dz = e^{t^{2}/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(z-t)^{2}/2} \, dz = e^{t^{2}/2}. \]

The characteristic function is more difficult to derive; it can be obtained by contour integration in the complex plane:

\[ c(t) = e^{-t^{2}/2}. \]

Now, suppose that the distribution on \( R \) has come from some initial space \( S \) with class \( \mathfrak{C} \) and measure \( P \). Let \( X \) be a real-valued function; then

\[ m_{X}(t) = E(e^{X(t)t}) = \int_{S} e^{X(t)s} \, dP(s), \]

\[ c_{X}(t) = E(e^{itX(t)}) = \int_{S} e^{itX(s)} \, dP(s). \]

It is of some formal interest that the functions \( m \) and \( c \) defined by (1) and (2) are two sections of the function \( E(e^{zw}) \) for \( w \) in the complex plane.

B PROPERTIES OF MOMENT-GENERATING FUNCTIONS

Two different distributions cannot have the same moment-generating function or the same characteristic function. Thus the moment-generating function or the characteristic function contains enough information to fully determine the corresponding distribution. This uniqueness result is established in Supplement Section 5.8.

The uniqueness result can be presented in terms of one-to-one correspondences as illustrated in Figure 5.19. Certain operations on distribution functions or on density functions may be difficult to complete, and yet the corresponding operations on moment-generating functions may be very simple. Thus if we can move back and forth easily between distributions and generating functions, then we have available a simple round-about route for completing the operations. We now consider some operations that are simple in terms of the generating functions.

Consider a distribution for \( x \) with moment-generating function \( m \) and characteristic
function $c$. And suppose that we obtain a distribution for $y$ by the affine transformation

$$y = a + kx = Y(x).$$

Then the corresponding functions for $y$ are

\begin{align*}
(5) \quad m_y(t) &= E(e^{(a+k^2)t}) = e^{at}E(e^{k^2t}) = e^{at}m(kt), \\
&= e^{at}E(e^{k^2t}) = e^{iat}c(kt).
\end{align*}

Consider a distribution for $(x_1, x_2)$ and suppose that $x_1$ is independent of $x_2$ and that $x_i$ has moment-generating function $m_i$ and characteristic function $c_i$. Then the corresponding functions for the sum $y = x_1 + x_2 = Y(x_1, x_2)$ are

\begin{align*}
(6) \quad m_y(t) &= E(e^{yt}) = E(e^{yt_1}e^{yt_2}) = m_1(t)m_2(t), \\
&= c_1(t)c_2(t),
\end{align*}

where we have used formula (5.1.13). Thus to add independent variables, we multiply the moment-generating functions and we multiply the characteristic functions. Note that we can write $y = x_1 + x_2$, but to calculate its distribution directly we must perform an integration; recall the convolution formulas (3.2.9) and (3.2.25)!

Now consider a moment-generating function that exists for the closed interval $[-\delta, +\delta]$. Then we can write

\begin{align*}
(7) \quad m(t) &= E(e^{xt}) = E\left(1 + xt + x^2 \frac{t^2}{2!} + \cdots \right) = 1 + \mu_1 t + \mu_2 \frac{t^2}{2!} + \cdots.
\end{align*}

using formula (5.1.12) and the fact that a partial sum $1 + xt + \cdots + x^2t^2/2!$ is bounded by the integrable function $e^{x^2} + e^{-x^2}$.

The preceding formula gives grounds for the name "moment-generating function."

And, in a corresponding way, if we follow the pattern for a Taylor expansion and take the $r$th derivative at $t = 0$, we obtain the $r$th moment:

\begin{align*}
(8) \quad \frac{d^r m(t)}{dt^r} \bigg|_{t=0} &= \mu_r.
\end{align*}

Now suppose that the $r$th moment $\mu_r$ exists. Then by Propositions 8 and 9 of Supplement Section 5.8, we obtain
\[
\begin{align*}
(9) \quad c(t) &= 1 + i\mu t - \frac{\mu^2 t^2}{2!} - \frac{i\mu^3 t^3}{3!} + \cdots + (i)\mu_\tau \frac{t^\tau}{\tau!} + o(t) \\
(10) \quad \left. \frac{d^\tau c(t)}{dt^\tau} \right|_{t=0} &= E(i^\tau x^\tau e^{it}) \left|_{t=0} = i^\tau \mu_\tau. \right.
\end{align*}
\]

C SOME EXAMPLES INVOLVING MOMENT-GENERATING FUNCTIONS

EXAMPLE 3

The normal \((\mu, \sigma)\) distribution: Consider the standard normal distribution and suppose that we expand the moment-generating function in a power series in accord with formula (7):

\[
m(t) = e^{t^2/2} = 1 + \left(\frac{t^2}{2}\right) + \left(\frac{t^2}{2}\right)^2 \frac{1}{2!} + \left(\frac{t^2}{2}\right)^3 \frac{1}{3!} + \cdots
\]

\[
= 1 + \frac{t^2}{2!} + 3 \cdot 1 \frac{t^4}{4!} + 5 \cdot 3 \cdot 1 \frac{t^6}{6!} + \cdots.
\]

Then we obtain

\[
(11) \quad \mu_{2m+1} = 0, \quad \mu_{2m} = (2m - 1)(2m - 3) \cdots 3 \cdot 1.
\]

Now suppose that we form the general normal \((\mu, \sigma)\) distribution by the affine transformation \(y = \mu + \sigma z = Y(x).\) Then by formula (5) we obtain

\[
(12) \quad m_Y(t) = e^{it\mu + (\sigma t)^2/2} = e^{it\mu + \sigma^2 t^2/2},
\]

\[
c_Y(t) = e^{it\mu - (\sigma t)^3/6} = e^{it\mu - \sigma^2 t^3/6}.
\]

EXAMPLE 4

Convolution of the normal: Consider a distribution for \((x_1, \ldots, x_n)\) on \(R^n\) and suppose that the \(x_1, \ldots, x_n\) are statistically independent and \(x_i\) is normally distributed \((\mu_i, \sigma_i).\) By Example 3 the moment-generating function of \(x_i\) is

\[
m_i(t) = e^{it\mu_i + \sigma_i^2 t^2/2}.
\]

Now suppose that we form the sum variable \(y = x_1 + \cdots + x_n = Y(x_1, \ldots, x_n).\) Then by formula (6) we obtain

\[
m_Y(t) = m_1(t) \cdots m_n(t) = \prod_{i=1}^{n} e^{it\mu_i + \sigma_i^2 t^2/2}
\]

\[
= e^{it\mu + \Sigma \sigma_i^2 t^2/2}.
\]

Note that this is the moment-generating function for the normal \((\Sigma \mu_i, (\Sigma \sigma_i^2)^{1/2})\) distribution. By the uniqueness property discussed in Section B, it follows that \(y\) is normal \((\Sigma \mu_i, (\Sigma \sigma_i^2)^{1/2}).\) Thus if we add independent normal variables, we obtain a normal variable.
D MOMENT-GENERATING FUNCTIONS ON $R^k$

Consider a distribution for $x$ on $R^k$:

**DEFINITION 3**

The moment-generating function $m$ of a distribution for $x$ on $R^k$ with probability measure $P$ is given by

\[(13) \quad m(t) = m(t_1, \ldots, t_k) = E(e^{\sum x_i t_i}) = E(e^{x^t t}) = \int e^{x^t} dP(x)\]

provided that the integral exists. The function $m$ is said to exist if the integral (13) exists for $t$ in an open rectangle containing the origin $0$. The characteristic function is defined by

\[(14) \quad c(t) = c(t_1, \ldots, t_k) = E(e^{i\sum x_i t_i}) = \int e^{i x^t} dP(x).\]

Again, on $R^k$, two different distributions cannot have the same moment-generating function or the same characteristic function. The moment-generating function or the characteristic function thus contains enough information to fully determine the corresponding distribution. See the earlier remarks in Section B; the formal uniqueness results will be recorded in Supplement Section 5.8E.

Consider a distribution for $x$ with moment-generating function $m$. Then the moment-generating function for $y = a + Cx = Y(x)$ is

\[(15) \quad m_Y(t) = E(e^{y^t}) = E(e^{(a^t + xc^t)}) = e^{a^t}m(C^t).\]

Consider a distribution for $(x_1, \ldots, x_n)$ and suppose that the $x_i$ are independent and $x_i$ has moment-generating function $m_i$. Then the sum $y = \sum x_i = Y(x_1, \ldots, x_n)$ has moment-generating function

\[(16) \quad m_Y(t) = E(e^{\sum x_i t_i}) = m_1(t) \cdots m_n(t).\]

Now suppose that a moment-generating function $m$ is expanded in a series about 0:

\[(17) \quad m(t) = \sum \frac{t_1^{r_1} \cdots t_k^{r_k}}{r_1! \cdots r_k!}.\]

The moments can be obtained by differentiation:

\[(18) \quad \frac{\partial r_1}{\partial t_1^{r_1}} \cdots \frac{\partial r_k}{\partial t_k^{r_k}} m \bigg|_{t=0} = \mu_{r_1} \cdots r_k ;\]

**EXAMPLE 5**

The multivariate normal($\mu; \Sigma$): Consider $z = (z_1, \ldots, z_k)'$ with a standard normal distribution on $R^k$. From the independence among the coordinates we obtain

\[\begin{align*}
    m(t_1, \ldots, t_k) &= E(e^{z_1 t_1 + \cdots + z_k t_k}) = e^{t_1^2/2} \cdots e^{t_k^2/2} = e^{t^t/2}.
\end{align*}\]
Suppose that we form the general normal\((\mu; \Sigma)\) distribution by the affine transformation \(y = \mu + \Gamma z = Y(z)\), where \(Y\) is in \(R^l\). Then by (15) we obtain the moment-generating function for the general normal\((\mu; \Sigma)\),
\[ m_Y(t) = e^{\mu^t + \Sigma \Gamma^t / 2} = e^{\mu^t + \Sigma_{1/2}}, \]
where \(\Sigma = \Gamma \Gamma'\). Formulas (15) and (17) can then show that \(\mu\) is the mean and \(\Sigma\) is the variance matrix; see Problem 42.

**E MARGINAL DISTRIBUTIONS**

In Chapter 2 we developed a variety of methods for determining marginal distributions. With the uniqueness theorems in this section we now have a new range of methods for determining marginal distributions. Often we can find fairly easily the characteristic function, or the moment-generating function, or the moments of some marginal variable of interest. Then if we recognize the characteristic function, or use the inversion formula, we can obtain the corresponding distribution. These new methods are largely restricted to linear operations on the initial variables. See Example 4 in Section C; also the problems in Section K below.

**F PROBABILITY-GENERATING FUNCTIONS**

For discrete distributions on the nonnegative integers we can define a very useful function closely related to the moment-generating function. It provides yet another way of determining a marginal distribution—for responses involving counts.

**DEFINITION 4**

The **probability-generating function** \(r\) of a distribution for \(x\) on the nonnegative integers is given by
\[ r(t) = E(t^x) = t^0p(0) + t^1p(1) + t^2p(2) + \cdots \]
for those values of \(t\) for which the summation converges; the summation converges for an interval that includes \(t = 0\) and \(t = 1\).

For certain purposes connected with moments, we find it convenient to say that \(r\) **exists** if the summation (19) exists for an **open** interval containing \(t = 1\).

Note the obvious connection with the moment-generating function:
\[ m(t) = r(e^t); \quad r(t) = m(\ln t), \quad t > 0. \]
Also note the fundamental property that the coefficients of the powers of \(t\) are the corresponding probabilities of the distribution.

For an affine transformation \(y = a + kx = Y(x)\), we have
\[ r_Y(t) = t^a r(t^k). \]

For the sum \(y = x_1 + x_2 = Y(x_1 + x_2)\) where \(x_1\) and \(x_2\) are independent with probability-generating functions \(r_1\) and \(r_2\), we have
(22) \[ r_X(t) = r_1(t)r_2(t). \]

Now suppose that a probability-generating function is differentiated \( s \) times and evaluated at \( t = 1 \):

\[
\frac{d}{dt} r(t) = \sum x t^{x-1} p(x), \quad \frac{d^s r(t)}{dt^s} = \sum x^{(s)} t^{x-s} p(x);
\]

(23) \[
\left. \frac{d}{dt} r(t) \right|_{t=1} = \mu, \quad \left. \frac{d^s r(t)}{dt^s} \right|_{t=1} = E(x^{(s)}) = \mu^{(s)};
\]

thus we obtain the factorial moments.

---

EXAMPLE 6

Sample from the Bernoulli(\( p \)) distribution: Consider a sample from the Bernoulli(\( p \)) distribution. For an individual \( x_i \) we have

\[ r_i(t) = q + pt. \]

Then for the sum \( Y = x_1 + \cdots + x_n \), we have

\[ r_Y(t) = (q + pt)^n = q^n + \binom{n}{1} p q^{n-1} t + \cdots + \binom{n}{r} p^r q^{n-r} t^r + \cdots + p^n t^n. \]

The distribution of \( Y \) is of course binomial(\( n, p \)); note that we have obtained the binomial probabilities as coefficients in the expansion of \( r_Y \).

The factorial moments for the binomial distribution can be calculated by evaluating the derivatives at \( t = 1 \). We can accomplish this equivalently by expanding in a series about \( t = 1 \). For this let \( u = t - 1 \); then

\[ r(t) = r(1 + u) = (q + p(1 + u))^n = (1 + pu)^n \]

\[ = 1 + n pu + n^2 p^2 \frac{u^2}{2!} + n^3 p^3 \frac{u^3}{3!} + \cdots + p^n u^n; \]

thus

\[ \mu^{(s)} = E(y^{(s)}) = n^{(s)} p^s. \]

---

The analogous definition for the probability-generating function of a distribution for \( x \) on \( \mathbb{R}^k \) is given by

\[ r(t_1, \ldots, t_k) = E(t_1^{x_1} \cdots t_k^{x_k}) = m(t_1, \ldots, t_k), \quad t_i > 0. \]

Differentiation at the point 1 gives

\[ \left. \frac{\partial^{s_1}}{\partial t_1^{s_1}} \cdots \frac{\partial^{s_k}}{\partial t_k^{s_k}} r \right|_{t=1} = E(x_1^{(s_1)} \cdots x_k^{(s_k)}) = \mu^{(s_1)\cdots(s_k)}; \]

the \((s_1, \ldots, s_k)\)th factorial moment.

---

G EXERCISES

1. Calculate the moment-generating function for the uniform distribution on the integers 1, 2.
2. 3. Use the moment-generating function to deduce the mean and second moment.
2. The binomial(4, 1/2) distribution is tabulated in Section 3.5A. Calculate the moment-generating function.

3. Show that the moment-generating function for the uniform(0, 1) distribution has \( m(t) = (e^t - 1)/t \). Use the moment-generating function to deduce the mean.

4. Calculate the moment-generating function for the triangular distribution with density \( f(x) = 2x \) on (0, 1) and zero elsewhere.

5. Show that the moment-generating function for the uniform distribution on the square (0, 1)^2 has \( m(t_1, t_2) = (e^{t_1} - 1)(e^{t_2} - 1)/t_1t_2 \).

**PROBLEMS: STATISTICAL INDEPENDENCE**

6. Consider a distribution for \((x_1, x_2)\) on the plane \(R^2\) with moment-generating function \( m(t_1, t_2) \) and characteristic function \( c(t_1, t_2) \). Show that the moment-generating function for \(x_1\) is \( m_1(t_1) = m(t_1, 0) \) and for \(x_2\) is \( m_2(t_2) = m(0, t_2) \). Correspondingly, show that \( c_1(t_1) = c(t_1, 0) \), \( c_2(t_2) = c(0, t_2) \).

7. (continuation) Show that \(x_1\) and \(x_2\) are statistically independent if and only if \( c(t_1, t_2) = g(t_1)h(t_2) \) factors so that the variables separate. Use the uniqueness discussed in Section B.

8. (continuation) Suppose that the moment-generating function \( m \) exists. Then show that \(x_1\) and \(x_2\) are statistically independent if and only if \( m(t_1, t_2) = g(t_1)h(t_2) \) factors so that the variables separate. Use the uniqueness described in Section B.

**PROBLEMS FOR DISCRETE DISTRIBUTIONS**

For each of the following distributions, verify the corresponding expressions for

- \( \mu_n \), \( r(t) \), \( m(t) \).

9. The uniform distribution on the integers 1, 2, \ldots, \( N \):

\[
\frac{(N + 1)^{r+1}}{N(r+1)} \quad (r > 0), \quad \frac{N^{r+1} - t}{N(t - 1)}, \quad \frac{e^{Nt} - e^t}{N(e^t - 1)}.
\]

see Problem 5.6.6.

10. Poisson(\( \lambda \)) distribution:

\[
\lambda^r, \quad \exp \{ \lambda t - \lambda \}, \quad \exp \{ \lambda e^t - \lambda \}.
\]

11. Geometric(p) distribution:

\[
\frac{r!q^r}{p^r}, \quad \left( \frac{1 - qt}{p} \right)^{-1}, \quad \left( \frac{1 - qe^t}{p} \right)^{-1}.
\]

12. Binomial(n, p) distribution:

\[
\frac{n^r p^r}{(q + p)^n}, \quad (q + pe^t)^n.
\]

13. Negative binomial(k, p) distribution:

\[
\frac{(k + r - 1)^{r}q^r}{p^r}, \quad \left( \frac{1 - qt}{p} \right)^{-k}, \quad \left( \frac{1 - qe^t}{p} \right)^{-k}.
\]

14. Hypergeometric(N, n, p) distribution:

\[
\frac{n^r(Np)^n}{N^n}.
\]

For each of the following distributions, verify the corresponding expressions for

- \( \mu_{t_1, \ldots, t_k} \), \( r(t_1, \ldots, t_k) \), \( m(t_1, \ldots, t_k) \).
Sec. 5.7: Moment-Generating and Characteristic Functions

15 Multinomial\((n, p_1, \ldots, p_k)\) distribution:
\[ r(p_1^{t_1} \cdots p_k^{t_k}) \left( p_1^{t_1 + \cdots + p_k^{t_k}} \right)^{n}, \quad (p_1^{t_1 + \cdots + p_k^{t_k}})^n. \]

16 Multivariate hypergeometric\((N, n, p_1, \ldots, p_k)\) distribution:
\[ \frac{r(N_1^{t_1} \cdots (N_k^{t_k})^{n})}{N^{t_1} \cdots (N_k^{t_k})^{n}}. \]

J PROBLEMS FOR ABSOLUTELY CONTINUOUS DISTRIBUTIONS

For each of the following distributions, verify the corresponding expressions for 
\[ \mu, \quad m(t), \quad c(t). \]

More difficult problems are marked with an asterisk.

17 Standard exponential distribution:
\[ r, \quad (1 - t)^{-1}, \quad (1 - it)^{-1}. \]

18 Pareto\((\alpha)\) distribution for \(z\); the distribution for \(1 + z\) has moments
\[ \frac{\alpha}{\alpha - r} \quad (r < \alpha). \]

19 Standard Laplace distribution:
\[ 0 \quad \text{(r odd)} \quad (1 - t^2)^{-1}, \quad (1 + t^2)^{-1}. \]
\[ r! \quad \text{(r even)} \]

**20 Standard logistic distribution:
\[ 0 \quad \text{(r odd)} \quad 2 \Gamma(r + 1) (1 - 2^{1-r})^{\xi(r)} \quad \text{(r even)} \quad B(1 - t, 1 + t), \quad \pi t \cosech \pi t, \]

where \(B\) is the beta function (Problem 2.2.22) and \(\xi(x) = \sum_{n=1}^{\infty} i^{-x}\) is the Riemann zeta function.

21 Standard lognormal\((\tau)\):
\[ \exp \left\{ \frac{1}{2} \tau^2 \right\}. \]

22 Gamma\((p)\) distribution:
\[ \frac{\Gamma(p + \tau)}{\Gamma(p)}, \quad (1 - t)^{-p}, \quad (1 - it)^{-p}. \]

23 Chi-square\((m)\) distribution:
\[ m(m + 2) \cdots (m + 2r - 2), \quad (1 - 2t)^{-m/2}, \quad (1 - 2it)^{-m/2}. \]

24 Beta\((p, q)\) distribution:
\[ \frac{p^{[r]}}{(p + q)^{[r]}}, \quad \sum_{r=0}^{\infty} \frac{p^{[r]}}{(p + q)^{[r]}}, \quad \frac{p^{[r]}}{r!}, \quad \sum_{r=0}^{\infty} \frac{(it)^r}{r!}, \]

where \(p^{[r]} = p(p + 1) \cdots (p + r - 1)\) is \(p\) ascending \(r\) factorial.

25 Canonical \(F(m, n)\) distribution:
\[ \frac{(m/2)^{[r]}}{(n/2 - 1)^{[r]}}, \quad \sum_{r=0}^{\infty} \frac{(m/2)^{[r]}}{(n/2 - 1)^{[r]}}, \quad r. \]
26 Canonical Student(\(n\)) distribution:

\[
\begin{align*}
0 & \quad (r \text{ odd}) \\
\frac{(r - 1)(r - 3) \cdots 3 \cdot 1}{(n - 2)(n - 4) \cdots (n - r)} & \quad (r \text{ even})
\end{align*}
\]

27 Standard Weibull(\(\beta\)) distribution:

\[\Gamma(\beta^{\frac{1}{\beta}} + 1).\]

28 Standard Cauchy distribution has characteristic function

\[c(t) = \exp\{-|t|\}.\]

Method: Use Problem 19 and the inversion formula in Supplement Section 5.8.

K PROBLEMS INVOLVING UNIQUENESS
AND OTHER PROPERTIES
OF GENERATING FUNCTIONS

29 Let \(x_1, \ldots, x_k\) be statistically independent and \(x_i\) be binomial(\(n_i, p\)). Show that \(y = \sum x_i\) is binomial(\(\sum n_i, p\)).

30 Let \(x_1, \ldots, x_k\) be statistically independent and \(x_i\) be Poisson(\(\lambda_i\)). Show that \(y = \sum x_i\) is Poisson(\(\sum \lambda_i\)).

31 Let \(z\) be standard normal. Determine the moment-generating function of \(y = z^2\) directly from the normal density and then show that \(y = z^2\) is chi-square(1).

32 Let \(x_1, \ldots, x_k\) be statistically independent and \(x_i\) be gamma(\(\rho_i\)). Show that \(y = \sum x_i\) is gamma(\(\sum \rho_i\)).

33 Let \(x_1, \ldots, x_k\) be statistically independent and \(x_i^2\) be chi-square(\(f_i\)). Show that \(x^2 = \sum x_i^2\) is chi-square(\(\sum f_i\)).

34 Let \((x_1, \ldots, x_k)\) be a sample from the standard Cauchy distribution. Show that \(y = \sum x_i/k\) has the same standard Cauchy distribution!

35 Let \((x_1, \ldots, x_k)\) be a sample from the normal(\(\mu, \sigma\)) distribution. Use moment-generating functions to show that \(\bar{x} = \sum x_i/n\) is normal(\(\mu, \sigma/\sqrt{n}\)).

36 Let \((x_1, \ldots, x_k)\) be a sample from the multivariate normal(\(\mu, \Sigma\)). Show that \(\bar{x} = \sum x_i/n\) is multivariate normal(\(\mu, n^{-1}\Sigma\)).

37 Let \(x\) be multivariate normal(\(\mu, \Sigma\)). Show that \(y = a + Cx\) is multivariate normal(\(a + C\mu; C \Sigma C^T\)).

38 Consider a distribution for \(x\) on \(R^k\) with probability measure \(P\). If the marginal distribution of \(\sum a_i x_i = a' x\) is given for all vectors \(a\), then show that the probability measure \(P\) is uniquely determined.

39 Consider a distribution for \(x\) with all probability on the nonnegative integers. Let \(r\) be the probability generating function and \(R\) be the distribution-generating function

\[R(t) = \sum F(x) t^x, \quad R(t) = \frac{r(t)}{1 - t};\]

prove the connecting relation between \(R\) and \(r\).

40 Let \(x_1, x_2, \ldots\) be statistically independent with a common discrete distribution with probability generating function \(r(t)\). Let \(n\) be statistically independent with a discrete distribution on \(\{1, 2, \ldots\}\) with probability generating function \(g(t)\). Let \(s = x_1 + \cdots + x_n\); show that its probability generating function is \(g(r(t))\).
FURTHER PROBLEMS

41 The moment-generating function for the multivariate normal \((\mu; \Sigma)\) is recorded in Example 5. Specialize and record the moment-generating function for the bivariate normal \((\mu_1, \mu_2; \sigma_1, \sigma_2, p_l)\).

42 (continuation) For the general case expand in a series and show that \(E(y) = \mu\). Determine the generating function for \(y - \mu\); expand in a series and show that \(\text{VAR}(y) = \Sigma\).

43 The logarithm of the characteristic function is called the cumulant-generating function \(k\).

\[ k(t) = \ln c(t), \quad c(t) = e^{it\mu}. \]

If \(c(t)\) can be expanded to \(r\) terms (if the \(r\)th moment exists), then so can the cumulant-generating function; the coefficients (analogous to moments) are called cumulants and are designated \(k_1, k_2, k_3, \ldots\):

\[ k(t) = k_1(it) + k_2 \frac{(it)^2}{2!} + k_3 \frac{(it)^3}{3!} + \cdots + k_r \frac{(it)^r}{r!} + o(t^r). \]

Use the exponential or logarithmic expansion to show that

\[ k_1 = \mu \quad \mu_1 = k_1 \]
\[ k_2 = \sigma^2 \quad \mu_2 = k_2 + k_1^2 \]
\[ k_3 = \bar{\mu}_3 \quad \mu_3 = k_3 + 3\mu_1 k_2 + k_1^3 \]
\[ \vdots \]

44 If \(x_1\) and \(x_2\) are statistically independent with cumulant-generating functions \(k_1\) and \(k_2\), show that the cumulant-generating function for the convolution distribution of \(x_1 + x_2\) is \(k_1 + k_2\). This additivity often makes cumulants more convenient than moments.

45 Let \(x_1, \ldots, x_n\) be a sample from a distribution with cumulants \(k_1, k_2, \ldots\). Functions of \(x_1, \ldots, x_n\), called \(k\) statistics, have been defined for use with cumulants:

\[ k_1 = \frac{S_1}{n}, \quad k_2 = \frac{nS_2 - S_1^2}{n(n - 1)}, \quad k_3 = \frac{n^2S_3 - 3nS_2S_1 + 2S_1^3}{n(n - 1)(n - 2)}, \]

where \(S_1 = x_1^2 + \cdots + x_n^2\), Show that \(k_1 = \bar{x}, k_2 = s^2\). Show that \(E(k_i) = k_i\) for \(i = 1, 2, 3\).

SUPPLEMENTARY MATERIAL

Supplement Section 5.8 is on pages 544–554. The uniqueness theorem for characteristic functions is proved; also the theorem for moment-generating functions. The uniqueness theorems for moment sequences are recorded; also the various theorems for the multivariate case.

A brief development of the general integral is given, and various properties and theorems are developed or presented as problems; for example, linearity and limiting properties. The Fubini theorem is recorded in a convenient form for later reference.

NOTES AND REFERENCES

This chapter assembles various concepts connected with the mean-value concept: mean, variance, covariance, moments, moment-generating functions, characteristic functions. These concepts all depend on the integral.

The first section presents the mean value in terms of the integral, examining the special cases of summation and the Riemann integral. Supplement Section 5.8 records various properties of the general integral.

The basic development is for distributions on the real line but always in a form suitable for
matrix extension to distributions in several dimensions. This makes available some rather rich results for the vector case.

For some references to other texts, see the Notes and References for Chapter 1 and the following:

We have examined probabilities themselves in Chapters 1–4 and mean-value characteristics in Chapter 5. These results lead to the following: the average \( \bar{y} \) for a sample \((y_1, \ldots, y_n)\) from the normal\((\mu, \sigma)\) distribution has a normal \((\mu, \sigma/\sqrt{n})\) distribution. A proof can be based on routine probability methods (Problem 3.3.11) or on mean-value characteristics (Problem 5.7.35).

Empirically, however, we find a much broader result: The average \( \bar{y} \) for a sample \((y_1, \ldots, y_n)\) from an arbitrary distribution with mean \( \mu \) and standard deviation \( \sigma \) is approximately normal\((\mu, \sigma/\sqrt{n})\) as \( n \) becomes large. A theoretical verification of this can be based on the notion of limiting distribution. In Chapter 6 we examine limiting distributions: the approach to a limit in Section 6.1, and the approach to a normal limit (Central Limit Theorems) in Section 6.2. The concept of limiting function is discussed in Supplement Section 6.3, together with some concluding comments (Section 6.3N) on the development of the probability model.
6
LIMITING DISTRIBUTIONS
AND LIMITING FUNCTIONS

6.1
LIMITING DISTRIBUTIONS

Many response variables in science and industry are approximately normally distributed. A theoretical argument supporting this runs as follows. There are many small sources of variation internal to the system being investigated; these components add to produce the variation in the response. The sum of bounded independent components has a limiting normal distribution. In this section we examine the notion of limiting distribution. Then in Section 6.2 we examine some Central Limit Theorems for limiting normal distributions. Supplement Section 6.3 examines limiting functions and records some concluding comments (Section 6.3N) on the probability model. The limiting distributions provide some very useful approximations.

A EXAMPLE OF A LIMITING DISTRIBUTION

Consider a classical example involving a limiting distribution. The Poisson distribution is often used in applications where we can envisage a large number of independent sources for an event and yet a very small probability that any particular source produces the event. For example, consider radioactive material with a large number of atoms, each having a small probability of producing an \(\alpha\)-particle; recall Section 2.1E. The binomial distribution, however, is the obvious distribution for the total frequency with independent components having common probability for producing the event.

For this example we examine the binomial\( (n, p) \) distribution with \( n \) large and \( p \) small; specifically we take a fixed mean \( \lambda = np \), write \( \rho = \lambda / n \), and investigate the limit as \( n \to \infty \). Let \( y_n \) be used to present events and probabilities for this binomial distribution.

Consider the binomial probability function at a point \( x \):

\[
P(y_n = x) = \rho_n(x) = \binom{n}{x} \left( \frac{\lambda}{n} \right)^x \left( 1 - \frac{\lambda}{n} \right)^{n-x} \quad x = 0, 1, 2, \ldots, n
\]

\[
= \frac{\lambda^x}{x!} \left( 1 - \frac{\lambda}{n} \right)^n \cdot \frac{n^x}{n^x} \left( 1 - \frac{\lambda}{n} \right)^{-x}
\]

\[
= \frac{\lambda^x}{x!} \left( 1 - \frac{\lambda}{n} \right)^n \cdot \frac{1}{1} \left( 1 - \frac{1}{n} \right) \cdots \left( 1 - \frac{x-1}{n} \right) \cdot \left( 1 - \frac{\lambda}{n} \right)^{-x}.
\]
Consider the limit as \( n \to \infty \):

\[
\lim_{n \to \infty} P(y_n = x) = \lim_{n \to \infty} \rho_n(x) = \frac{\lambda^x}{x!} e^{-\lambda} \cdot 1 \cdot 1 = \frac{\lambda^x}{x!} e^{-\lambda} = \rho(x),
\]

where \( \rho \) is the probability function for the Poisson\((\lambda)\) distribution. Thus for large \( n \) the binomial\((n, \lambda/n)\) probability function approaches the Poisson\((\lambda)\) probability function.

We now examine the binomial distribution function at a point \( x \):

\[
P(y_n \leq x) = F_n(x) = \rho_n(0) + \rho_n(1) + \cdots + \rho_n([x]),
\]

where \([x]\) is the largest integer \( \leq x \). Taking the limit of the finite sum we obtain

\[
\lim_{n \to \infty} F_n(x) = \rho(0) + \rho(1) + \cdots + \rho([x]) = F(x),
\]

where \( F \) is the Poisson\((\lambda)\) distribution function. Thus the binomial\((n, \lambda/n)\) distribution function approaches the Poisson\((\lambda)\) distribution function at all points \( x \) on the real line.

For a numerical illustration suppose that we have 50 independent repetitions of a system that has a certain event that occurs with probability 1/50. Let \( y_{60} \) designate the number of occurrences of the event. The distribution of \( y_{60} \) is binomial\((50, 1/50)\); some values for this binomial probability function and for the corresponding Poisson\((1)\) probability function are given in Table 6.1.

<table>
<thead>
<tr>
<th>Table 6.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability function values</td>
</tr>
<tr>
<td>( x )</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>( \rho_{60}(x) )</td>
</tr>
<tr>
<td>( \rho(x) )</td>
</tr>
</tbody>
</table>

**B DEFINITION OF A LIMITING DISTRIBUTION**

In Section A we showed that the binomial distribution function with \( p = \lambda/n \) converges to the Poisson\((\lambda)\) distribution function. This convergence of a sequence of distribution functions will be the basis for defining a limiting distribution. We first examine a small technical complication illustrated by the following example.

**EXAMPLE 1**

Consider a distribution for \( y_n \) with probability 1/2 at each of 1/n and \(-1/n\). As \( n \to \infty \) the two points, each with probability 1/2, converge together toward zero. Let \( F_n \) be the distribution function for \( y_n \):

\[
F_n(x) = \begin{cases} 
0 & \text{if } -\infty < x < -1/n, \\
1/2 & -1/n \leq x < 1/n, \\
1 & 1/n \leq x < \infty,
\end{cases}
\]

\[
\lim_{n \to \infty} F_n(x) = 0 \quad \text{if } -\infty < x < 0,
\]
\[
= 1/2 \quad \quad \quad \quad \quad \quad x = 0,
\]
\[
= 1 \quad \quad \quad \quad \quad \quad 0 < x < \infty.
\]

See Figure 6.1 and note that \( \lim F_n \) is not a distribution function.

Now consider intuitively the distribution for \( y_n \). Clearly the two probability weights of 1/2 are converging to the origin; and in a sense the total probability is piling up at the origin. Consider a distribution for \( y \) with all probability at the origin. Then we would want to say that \( y_n \to y \) in a distribution sense.

Let \( H \) be the distribution function for \( y \) which has all probability at the point 0:

\[ H(x) = 0, 1 \quad \text{according as } x < 0, 0 \leq x. \]

Then we note that \( \lim F_n = H \) except at the point 0, where \( H \) has a discontinuity. Accordingly, we find it convenient to ignore nonconvergence at the points of discontinuity of the limit function.

We now formalize the concept of a sequence of distributions approaching a limiting distribution.

**DEFINITION 1**

A sequence of distributions with distribution functions \( F_n \) has limiting distribution with distribution function \( F \) if

\[ \lim_{n \to \infty} F_n(x) = F(x) \]

at all points \( x \) at which \( F \) is continuous; we write \( F_n \to F \) or \( y_n \xrightarrow{D} y \) or \( P_n \xrightarrow{D} P \) and say that \( y_n \) converges to \( y \) in distribution.
For a substantial example, recall the binomial\((n, \lambda/n)\) distribution converging to the Poisson\((\lambda)\) distribution, as examined in Section A.

Now, in general, consider a sequence of distributions \(F_n\) with limiting distribution \(F\). An interval \(B = \{a, b\}\) on \(R\) is said to be an interval of continuity if the end points \(a\) and \(b\) are points of continuity of the distribution function \(F\). Then, by differencing the distribution function, we obtain

\[
\lim_{n \to \infty} P_n(B) = \lim_{n \to \infty} (F_n(b) - F_n(a)) = F(b) - F(a) = P(B).
\]

Of course, the limit need not hold for other intervals or for most Borel sets. Accordingly, if we know a distribution approximately or if we are concerned with the limiting distribution of a sequence, we are essentially restricted to intervals—in fact, to rather special intervals. Thus we are withdrawing rather substantially from the Borel sets as the logical consequence of examining intervals.

**C Criteria for Limiting Distributions**

Sometimes limiting distributions can be determined directly from probability density functions.

**Proposition 2**

If the distribution function \(F_n\) has density \(f_n\) and the distribution function \(F\) has density \(f\), then

\[
(2) \quad \lim_{n \to \infty} f_n(x) = f(x)
\]

at all points \(x\) implies that \(F_n \to F\).

For a method of proof, see Problem 17.

**Example 2**

**Student distribution:** Consider the standard Student\((n)\) distribution as examined in Problem 2.2.19. Does this distribution have a limiting distribution as \(n \to \infty\)?

We calculate:

\[
(3) \quad \lim_{n \to \infty} f_n(t) = \lim_{n \to \infty} \frac{\Gamma\left(\frac{n + 1}{2}\right)}{\Gamma(1/2)\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-n/2} \frac{1}{\sqrt{n}}
\]

\[
= \lim_{n \to \infty} \frac{1}{\sqrt{2\pi}} \frac{\Gamma\left(\frac{n + 1}{2}\right)}{\Gamma(n/2) \sqrt{n/2}} \left(1 + \frac{t^2}{n}\right)^{-n/2} \left(1 + \frac{t^2}{n}\right)^{-1/2}
\]

\[
= \frac{1}{\sqrt{2\pi}} \cdot 1 \cdot e^{-t^2/2} \cdot 1 = \frac{1}{\sqrt{2\pi}} e^{-t^2/2};
\]

this uses Problem 3.4.11. Thus the standard Student\((n)\) distribution converges to the standard normal distribution as \(n \to \infty\). See Figure 6.2.
FIGURE 6.2
Standard Student(\(n\)) distribution for \(n = 1, 3, 5, 10, 20, 50, \infty\). The limiting distribution is standard normal.

A very powerful way of determining limiting distributions is by means of moment-generating and characteristic functions. For this, we need the following continuity theorem:

**THEOREM 3**

If \(F_n\) has characteristic function \(c_n\) and \(F\) has characteristic function \(c\), then \(F_n \to F\) if and only if

\[(4) \quad \lim_{n \to \infty} c_n(t) = c(t)\]

at all points \(t\). If \(F_n\) has moment-generating function \(m_n\) and \(F\) has moment-generating function \(m\), then \(F_n \to F\) if and only if

\[(5) \quad \lim_{n \to \infty} m_n(t) = m(t)\]

for an open interval containing the origin.

Proof The proof is recorded in Supplement Section 6.3H.

An analogous result holds for moment sequences—with a single qualification that the limiting moment sequence should correspond to a unique distribution.

Figure 6.3 presents the one-to-one mapping between distributions \(\mathcal{D}\) and characteristic functions \(\mathcal{C}\) and the one-to-one mapping between a subset of \(\mathcal{D}\) and the moment-generating functions \(\mathcal{M}\). The theorem says that these one-to-one mappings are continuous; we refer to the theorem as the **continuity theorem**.
EXAMPLE 3

**Poisson(λ) distribution:** Consider the Poisson(λ) distribution for \( y \). The mean of the distribution is \( \lambda \) and the standard deviation is \( \sqrt{\lambda} \). As \( \lambda \to \infty \), the distribution dilates and moves toward \( +\infty \). Suppose that we examine the limiting distribution of the variable standardized with respect to location and scale. The moment-generating function of the Poisson is recorded in Problem 5.7.10. We now use properties (5.7.5) of the moment-generating function and calculate

<table>
<thead>
<tr>
<th>Variable</th>
<th>( y )</th>
<th>( y - \lambda )</th>
<th>( (y - \lambda)/\sqrt{\lambda} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moment-generating function(( t ))</td>
<td>( \exp { \lambda e^t - \lambda } )</td>
<td>( \exp { \lambda e^t - \lambda - \lambda t } )</td>
<td>( \exp { \lambda e^{t/\sqrt{\lambda}} - \lambda - \sqrt{\lambda}t } )</td>
</tr>
</tbody>
</table>

Let \( m_\lambda(t) \) be the moment-generating function of \( (y - \lambda)/\sqrt{\lambda} \); then examining the logarithm we obtain

\[
\lim_{\lambda \to \infty} m_\lambda(t) = \lim_{\lambda \to \infty} (\lambda e^{t/\sqrt{\lambda}} - \lambda - \sqrt{\lambda}t)
\]

\[
= \lim_{\lambda \to \infty} \left( \lambda \left(1 + \frac{t}{\sqrt{\lambda}} + \frac{t^2}{2! \lambda} + \frac{t^3}{3! \lambda^{3/2}} + \cdots \right) - \lambda - \sqrt{\lambda}t \right)
\]

\[
= \lim_{\lambda \to \infty} \left( \frac{t^2}{2} + \frac{t^3}{3! \sqrt{\lambda}} + \frac{t^4}{4! \lambda} + \cdots \right) = \frac{t^2}{2};
\]

that is, \( \lim m_\lambda(t) = e^{t^2/2} \), the moment-generating function for the standard normal. Thus the standardized Poisson distribution converges to the standard normal distribution.

**D Closeness of the Poisson Approximation**

First consider the use of the preceding approximation to calculate a Poisson probability.

---

**FIGURE 6.3**
P\(_n\) converges to \( P \) if and only if \( c_n \) converges to \( c \) if and only if \( m_n \) converges to \( m \); this assumes that each function belongs to the indicated class.
EXAMPLE 4

The number of fatal automobile accidents per year in a large city is known to approximate the Poisson(80) distribution. On the assumption that conditions remain essentially stable, find the probability that the number of accidents $y$ for the coming calendar year is $\leq 110$.

We use the normal approximation in Example 3 and refine it slightly. The frequency $y$ has a discrete distribution on the integers; the normal is continuous. Thus for a frequency $y \leq 110$ it is perhaps reasonable to use the point $110 + 1/2$, half way between 110 and 111, as the dividing point. Let $P_{80}$ designate probability calculated according to the Poisson(80) and $P$ designate probability calculated according to the standard normal; we calculate

$$P_{80}(y \leq 110) = P_{80}\left(\frac{y - 80}{\sqrt{80}} \leq \frac{110 - 80}{\sqrt{80}}\right)$$

$$\approx P\left(z \leq \frac{110.5 - 80}{8.944}\right) = P(z \leq 3.410) = 0.99968.$$

The exact value for the Poisson(80) is 0.99929. Ordinarily, we would not depend on a limiting distribution for a calculation of probability far out on the tail of a distribution.

Now consider an alternative way to approximate Poisson($\lambda$) probabilities on the real line $R$. From Example 3 we know that the limiting distribution of

$$z_1 = \frac{y - \lambda}{\sqrt{\lambda}}$$

is standard normal; note that $y$ is rescaled by a different factor $1/\sqrt{\lambda}$ for each different $\lambda$ value. A somewhat different standardization

$$z_2 = 2(\sqrt{y} - \sqrt{\lambda})$$

can be more convenient: the scaling is the same for each $\lambda$. We now show that $z_2$ has a limiting standard normal distribution. For this let $G_\lambda$, $H_\lambda$ be the distribution functions for $z_1$, $z_2$; then

$$H_\lambda(x) = P(z_2 \leq x) = P(2(\sqrt{y} - \sqrt{\lambda}) \leq x) = P(\sqrt{y} \leq \sqrt{\lambda} + x/2)$$

$$= P(y \leq \lambda + x\sqrt{\lambda} + x^2/4) = P\left(\frac{y - \lambda}{\sqrt{\lambda}} \leq x + \frac{x^2}{4\sqrt{\lambda}}\right)$$

$$= G_\lambda(x + x^2/4\sqrt{\lambda});$$

thus

$$\lim_{\lambda \to \infty} H_\lambda(x) = \lim_{\lambda \to \infty} G_\lambda(x + x^2/4\sqrt{\lambda}) = G(x),$$

where $G$ is the standard normal distribution function. Note: $G_\lambda(x) < H_\lambda(x)$.

EXAMPLE 5

Consider again the Poisson(80) distribution for the accident frequency in Example 4. We calculate the Poisson probability $P_{80}(y \leq 110)$ using the square-root
approximation $z_2$ in (7). Again it is reasonable to go to the appropriate half-integer point:

$$P_{80}(y \leq 110) = P_{80}(2(\sqrt{y} - \sqrt{80}) \leq 2(\sqrt{110} - \sqrt{80}))$$

$$\approx P(z_2 \leq 2(\sqrt{110.5} - \sqrt{80}))$$

$$= P(z_2 \leq 3.135) = 0.99914.$$  

The exact value for the Poisson(80) is 0.99929. For this particular probability the square-root approximation is slightly better.

Now consider a comparison of the two approximations for the Poisson. Let $F_\lambda$ be the distribution function for the Poisson($\lambda$) and let $G$ be the standard normal distribution function. A good approximation should yield a $z$ value that accurately calculates the probability. The correct $z$ value for the Poisson probability $F_\lambda(y)$ is

$$z = G^{-1}(F_\lambda(y)).$$

The two approximations give

$$z_1 = \frac{y + 1/2 - \lambda}{\sqrt{\lambda}}, \quad z_2 = 2(\sqrt{y + 1/2} - \sqrt{\lambda}).$$

Consider the values for the Poisson(80) given in Table 6.2. The values in the last column are those found in Examples 4 and 5. See also Figure 6.4.

**TABLE 6.2**

<table>
<thead>
<tr>
<th>$y$</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
<th>110</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z_1$</td>
<td>-3.30</td>
<td>-2.18</td>
<td>-1.06</td>
<td>0.06</td>
<td>1.17</td>
<td>2.29</td>
<td>3.41</td>
</tr>
<tr>
<td>$z_2$</td>
<td>-3.68</td>
<td>-2.33</td>
<td>-1.10</td>
<td>0.06</td>
<td>1.14</td>
<td>2.16</td>
<td>3.14</td>
</tr>
<tr>
<td>$z$</td>
<td>-3.52</td>
<td>-2.26</td>
<td>-1.07</td>
<td>0.07</td>
<td>1.17</td>
<td>2.22</td>
<td>3.24</td>
</tr>
</tbody>
</table>

In general, $z_2 < z_1$; this follows from $G_\lambda < H_\lambda$. An investigation of graphs such as Figure 6.4 indicates that $z_2 < z < z_1$, with $z_2$ approximating better than $z_1$; exceptions occur near the center, where, however, both approximations are very good. The second approximation $z_2$ given in (7) has, in addition, substantial advantages for statistics: a single reexpression $2\sqrt{y}$ gives the scaled variable regardless of the value for $\lambda$.

**E CONVERGENCE TO A CONSTANT**

If all the probability piles up at a point $c$, then it is reasonable to say that the distribution converges to the constant $c$; see Example 1. Consider a distribution function $F_n$ for $y_n$ on the real line $R$. Then as a special case of Definition 1, we say that $y_n$ converges in distribution to the constant $c$, $y_n \xrightarrow{D} c$, if

$$\lim_{n \to \infty} F_n(x) = 0 \quad x < c,$$

$$\quad = 1 \quad x > c.$$  

An alternative criterion for this convergence to the constant $c$ is that
Approximations $z_1$ and $z_2$ plotted against the true $z$ for the Poisson(20) distribution for $y$. As $\lambda$ increases, the approximations become closer in the pattern indicated.

\[
P(|\gamma_n - c| > \epsilon) \to 0 \quad \text{for each } \epsilon > 0
\]

as $n \to \infty$, or that

\[
P(|\gamma_n - c| \leq \epsilon) \to 1 \quad \text{for each } \epsilon > 0
\]

as $n \to \infty$. These are just other ways of saying that the probability for $\gamma_n$ piles up at the point $c$.

Part of the importance of this special definition is indicated by the following theorem.

**THEOREM 4**

If $\gamma_n \xrightarrow{D} \gamma$ and if $x_n \xrightarrow{D} 0$, then $\gamma_n + x_n \xrightarrow{D} \gamma$.

*Proof* The proof is recorded in Supplement Section 6.31. For some generalizations see Problems 6.1.14–6.1.16.

We can interpret the theorem by thinking of $x_n$ as a perturbation whose effects go to zero as $n \to \infty$. Thus we see that adding the perturbation $x_n$ to $\gamma_n$ does not change the limiting distribution for $\gamma_n$.

**F EXERCISES**

1. The number $y$ of fatal accidents per year in a large city is known to be approximately Poisson(100). Calculate the probabilities

\[
P(90 \leq y \leq 110), \quad P(80 \leq y \leq 120)
\]

using a normal approximation.
2. Use moment-generating or characteristic functions to show that the distribution of
\[ \frac{\chi^2 - m}{\sqrt{2m}} \]
converges to the standard normal distribution where \( \chi^2 \) has the chi-square(m) distribution.
3. Use Exercise 2 to show that the distribution of
\[ \sqrt{2\chi^2 - \sqrt{2m - 1}} \]
converges to the standard normal distribution where \( \chi^2 \) has the chi-square(m) distribution.
4. Prove that formula (9) is equivalent to \( \gamma_n \xrightarrow{} c \).
5. Show that if \( \gamma_n \) converges in distribution to the constant \( c \), then \( g(\gamma_n) \) converges in distribution to the constant \( g(c) \), where \( g \) is a continuous function.

G PROBLEMS

6. Use moment-generating functions or characteristic functions to show that the distribution of
\[ \frac{\gamma - np}{\sqrt{npq}} \]
converges to the standard normal distribution, where \( \gamma \) has the binomial(n, p) distribution.
7. Let \( (u_{1n}, \ldots, u_{nn}) \) be the order statistic for a sample of \( n \) from the uniform(0, 1) distribution.
   (a) Show that \( n u_{1n} \) has a limiting standard exponential distribution. Method: Use the distribution or density function.
   (b) Then deduce that \( n(1 - u_{nn}) \) has a limiting standard exponential distribution.
8. Let \( (y_{1n}, \ldots, y_{nn}) \) be the order statistic of a sample of \( n \) from a distribution with continuous distribution function \( F \). Use Problem 7 to show that \( nF(y_{nn}) \) has a limiting standard exponential distribution; that \( n(1 - F(y_{nn})) \) has a limiting standard exponential distribution.
9. Normal approximation to the Poisson; by density functions. Consider a Poisson(\( \lambda \)) distribution for \( y \) and an independent uniform(-1/2, 1/2) distribution for \( u \). Let \( w = y + u \).
   (a) Show that the density function for \( w \) is
   \[ f_w(w) = \frac{\lambda^{(w)}}{(w)!} e^{-\lambda} \quad \text{on} \ (-1/2, \infty), \]
   where \( (w) \) is the nearest integer to \( w \).
   (b) Use Theorem 4 to show that as \( \lambda \to \infty \)
   \[ \frac{\gamma - \lambda}{\lambda^{1/2}}, \quad \frac{w - \lambda}{\lambda^{1/2}} \]
   have the same limiting distribution, if any.
10. (continuation) Show that density for the standardized continuous Poisson \( (w - \lambda)/\lambda^{1/2} \) converges to the standard normal density function as \( \lambda \to \infty \); use Stirling's formula.
11. Normal approximation to the binomial; by density functions. Consider a binomial(n, p) distribution for \( y \) and an independent uniform(-1/2, 1/2) distribution for \( u \). Let \( w = y + u \). Show that the density function for \( w \) is
   \[ f_w(w) = \binom{n}{(w)} p^{(w)} q^{n-(w)} \quad \text{on} \ (-1/2, n + 1/2), \]
   where \( (w) \) is the nearest integer to \( w \).
12. (continuation) Show that the density function for the standardized continuous binomial \( (w - np)/(npq)^{1/2} \) converges to the standard normal density function as \( n \to \infty \); use Stirling's formula and Theorem 4.
13. Let \( (x_{1n}, \ldots, x_{nn}) \) be the order statistic for a sample of \( n \) from a distribution with a continuous
density function $f$. For a given $p = 1 - q$ in $(0, 1)$, let $x_{(np)}$ be the order statistic with subscript taken to be the largest integer in $np$, and $\xi_p$ be such that $F(\xi_p) = p$, where $F$ is the distribution function. Show that if $f(\xi_p) > 0$, then

$$\frac{x_{(np)} - \xi_p}{(pq/nf^2(\xi_p))^{1/2}}$$

has a limiting standard normal distribution as $n \to \infty$. **Method:** For given $z$, show that

$$x_{(np)} < \xi_p + z \frac{(pq/n)^{1/2}}{f(\xi_p)}$$

is equivalent to $y \geq [np]$, where $y$ is the number of observations less than $\xi_p + z(pq/n)^{1/2}/f(\xi_p)$, and also show that $y$ has a binomial$(n, F(\xi_p + z(pq/nf^2(\xi_p))^{1/2}))$ distribution.

**Assume:** The standardized binomial distribution function approaches the standard normal distribution function uniformly in $p$ with $0 < p_1 < p < p_2 < 1$.

14 If $y_n \xrightarrow{D} y$ and $x_n \xrightarrow{D} a$, show that $y_n + x_n \xrightarrow{D} y + a$. Thus if $y_n$ has limiting distribution function $F(\cdot)$, then $y_n + x_n$ has limiting distribution function $G(\cdot)$, where $G(w) = F(w - a)$.

15 (continuation) If $a > 0$, then $w_n = y_n/x_n$ has limiting distribution $G$, where $G(w) = F(w/a)$. **Hint:** As in Supplement Section 6.31.

16 (continuation) If $a > 0$, then $w_n = y_n/x_n$ has limiting distribution $H$, where $H(w) = F(aw)$. **Hint:** As in Supplement Section 6.31.

17 Prove Proposition 2. **Method:** Apply the Fatou lemma in Supplement Problem 5.8.5 to a Borel set $B$ and to the Borel set $B'$.

---

**Figure 6.5**

Distribution of the sample average $\bar{y}_n$ for a sample of $n$ from the Bernoulli$(1/2)$.
THE WEAK LAW OF LARGE NUMBERS AND THE CENTRAL LIMIT THEOREM

We have noted that many response variables in science and industry are approximately normally distributed—the explanation being that many small sources of variation add to produce normal variation in the response. We find also that many derived variables are very closely normally distributed. Notable among such variables of course is the average of a sample of independent repetitions on a random system; such samples are essential to experimental investigations. In Section A we examine two simple cases to see how the distribution of the sample average varies with the sample size. In Section B we examine the Weak Law of Large Numbers, which shows fairly generally that the sample average converges to the mean as the sample size increases. Then in Section C we examine some Central Limit Theorems that provide an explanation for the approximate normality for sums and averages and for many responses.

A EXAMPLES OF THE DISTRIBUTION OF THE SAMPLE AVERAGE

We now consider two distributions, one discrete and one continuous, and examine the distribution of the sample average \( \bar{y}_n = \frac{\sum y_i}{n} \) as the sample size is increased.

For a discrete distribution we choose the simple Bernoulli(1/2). The distribution

---

**FIGURE 6.6**
Distribution of the sample average \( \bar{y}_n \) for a sample of \( n \) from the uniform(0, 1).
of \( \Sigma y_i \) is binomial(n, 1/2). Correspondingly, the distribution of \( \bar{Y} \) is the binomial(n, 1/2) distribution but with the interval [0, n] scaled into [0, 1]; see Figure 6.5.

For an absolutely continuous distribution we choose the uniform(0, 1). The distribution of \( \Sigma y_i \) can be obtained by induction using the convolution formula (3.2.25); the density for \( \bar{Y} \) on (0, 1) is

\[
f(t) = \frac{n}{(n-1)!} \left( (nt)^{n-1} - \binom{n}{1} (nt - 1)^{n-1} + \binom{n}{2} (nt - 2)^{n-1} \right. \\
- \cdots \left. (-1)^r \binom{n}{r} (nt - r)^{n-1} \right)
\]

if \( \frac{r}{n} < t \leq \frac{r + 1}{n} \), for \( r = 0, 1, \ldots, n - 1 \) and zero otherwise. See Figure 6.6.

As \( n \) increases we notice in each case that the distribution of the sample average \( \bar{Y}_n \) is gradually concentrating around the mean \( \mu \) of the distribution being sampled. This phenomenon in fact occurs in general provided only that the mean \( \mu \) exists; it is covered by the Weak Law of Large Numbers discussed in Section B.

Now consider the shape of the distribution for the sample average \( \bar{Y}_n \). For this we need to remove the concentrating effect we have just described: the deviation from the mean is \( \bar{Y}_n - \mu \); the standard deviation of this deviation is \( \sigma / \sqrt{n} \), where \( \sigma^2 \) is the variance of the distribution being sampled; we examine the distribution of \( \sqrt{n}(\bar{Y}_n - \mu) \). This distribution is recorded in Figure 6.7 for the Bernoulli and in Figure 6.8 for the uniform.

As \( n \) increases, we notice in each case that the distribution of the adjusted variable \( \sqrt{n}(\bar{Y}_n - \mu) \) is gradually coming toward the bell-shaped form of the normal distribution. This phenomenon occurs generally if the mean \( \mu \) and variance \( \sigma^2 \) exist for the

---

**FIGURE 6.7**

Distribution of \( \sqrt{n}(\bar{Y}_n - \mu) \) for a sample of \( n \) from the Bernoulli(1/2).
distribution being sampled; it is covered by the Central Limit Theorem discussed in Section C.

B THE WEAK LAW OF LARGE NUMBERS

Consider a distribution for $y$ on the real line $R$ with probability measure $P$ and with mean $\mu$, say the Bernoulli distribution for recovery under a surgical treatment. Let $y = (y_1, \ldots, y_n)'$ be a sample: that is, with independent coordinates and with each coordinate having the same distribution $P$. The sample average $\bar{y}_n = (\Sigma^n_1 y_i)/n$ is a function from $R^n$ to $R$; it has a marginal distribution on $R$. In this section we see that the distribution of $\bar{y}_n$ concentrates or piles up at $\mu$ as the sample size $n \to \infty$.

THEOREM 1

Weak Law of Large Numbers: If $(y_1, \ldots, y_n)$ is a sample from a distribution with mean $\mu$, then

$$\lim_{n \to \infty} P(|\bar{y}_n - \mu| \leq \epsilon) = 1$$

for each $\epsilon > 0$; that is, $\bar{y}_n \to \mu$.

\[ n = 1 \]
\[ n = 2 \]
\[ n = 3 \]
\[ n = 4 \]
\[ n = 5 \]

FIGURE 6.8
Distribution of $\sqrt{n}(\bar{y}_n - \mu)$ for a sample of $n$ from the uniform(0, 1).
Proof. We give the proof under the simplifying assumption that the variance $\sigma^2$ exists for the distribution being sampled; the more general proof is given in Section F. From formula (5.4.5) we have

$$E(\overline{y}_n) = \mu, \quad \text{var} (\overline{y}_n) = \frac{\sigma^2}{n}.$$ 

The Chebyshev inequality (5.3.9) then gives

$$P(|\overline{y}_n - \mu| > \epsilon) \leq \frac{\sigma^2/n}{\epsilon^2},$$

$$P(|\overline{y}_n - \mu| \leq \epsilon) \geq 1 - \frac{\sigma^2/n}{\epsilon^2}.$$ 

Now consider the limit as $n \to \infty$:

$$\lim_{n \to \infty} P(|\overline{y}_n - \mu| \leq \epsilon) = 1;$$

and we have that $\overline{y}_n$ converges to $\mu$ in distribution, $\overline{y}_n \xrightarrow{\text{d}} \mu$.

**EXAMPLE 1**

Sample from the Bernoulli($\rho$): Consider a sample $(x_1, \ldots, x_n)$ from the Bernoulli($\rho$) distribution. The function $y = x_1 + \ldots + x_n$ gives the frequency for $n$ Bernoulli repetitions, and the function $\hat{\rho}_n = y/n = \overline{x}_n$ gives the corresponding proportion or estimated probability. For an application, consider some random system and a particular event of interest having probability $\rho$. For $n$ independent repetitions the function $\hat{\rho}_n = \overline{x}_n$ is the proportion of occurrences for the event. We now examine a property of the distribution that describes this proportion. From the Weak Law of Large Numbers we obtain

$$(2) \quad P(|\hat{\rho}_n - \rho| \leq \epsilon) \xrightarrow{n \to \infty} 1;$$

that is, the probability for $\hat{\rho}_n$ piles up at the true probability $\rho$. Our theory has produced, in fact, a model for the frequency interpretation of probability and the model is consistent with that frequency interpretation. This is an indication of the internal consistency of our theory.

**EXAMPLE 2**

Sample from the Cauchy distribution: Consider a sample $(y_1, \ldots, y_n)$ from the Cauchy distribution. From Example 5.2.3 we know that the mean does not exist. Thus it is not surprising to note from Problem 5.7.34 that the distribution for the sample average from the Cauchy distribution is just the same Cauchy distribution; thus
\[ P(|\bar{Y}| \leq \epsilon) = \int_{-\epsilon}^{\epsilon} \frac{1}{\pi(1 + y^2)} \, dy. \]

which is in sharp contrast to the Weak Law.

C The Central Limit Theorem

Consider a distribution for \( y \) on the real line \( \mathbb{R} \) with probability measure \( P \), with mean \( \mu \) and with variance \( \sigma^2 \). Let \( y = (y_1, \ldots, y_n)' \) be a sample from the distribution. We have seen that the marginal distribution of the sample average \( \bar{y}_n = (\Sigma_i^n y_i)/n \) concentrates at the mean \( \mu \). We now show that the shape of the marginal distribution approaches the normal; see for example, Figure 6.9.

From formulas (5.4.4) and (5.4.5) we have the location and scale characteristics of the sample sum \( \sum_1^n y_i \) and the sample average \( \bar{y}_n \):

\[
\begin{align*}
E\left(\sum_1^n y_i\right) &= n\mu, \\
\text{var}\left(\sum_1^n y_i\right) &= n\sigma^2, \\
E(\bar{y}_n) &= \mu, \\
\text{var}(\bar{y}_n) &= \frac{\sigma^2}{n}.
\end{align*}
\]

We now form a new variable by standardizing and measuring departure from the mean in units of the standard deviation:

\[
z_n = \frac{\sum_1^n y_i - n\mu}{\sqrt{n}\sigma} = \frac{\bar{y}_n - \mu}{\sigma/\sqrt{n}}.
\]

The standardized variable has mean zero and variance 1.

**FIGURE 6.9**

Distribution of \( \sqrt{n}(\bar{y} - \mu)/\sigma \) for a sample of \( n \) from the uniform(0, 1):
\( n = 1, 2, 5, 10, 30. \)
THEOREM 2

Central Limit Theorem: If \((y_1, \ldots, y_n)\) is a sample from a distribution with mean \(\mu\) and variance \(\sigma^2\), then the distribution of

\[
(4) \quad z_n = \frac{\sum_{i=1}^{n} y_i - n\mu}{\sqrt{n}\sigma} = \frac{Y_n - \mu}{\sigma/\sqrt{n}}
\]

converges to the standard normal distribution.

The theorem asserts that

\[
\lim_{n \to \infty} P(z_n \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \, e^{-z^2/2} \, dz;
\]

and thus that

\[
(5) \quad \lim_{n \to \infty} P(a < z_n \leq b) = \int_{a}^{b} \frac{1}{\sqrt{2\pi}} \, e^{-z^2/2} \, dz.
\]

The proof of the theorem is recorded in Section D.

EXAMPLE 3

Sample from the Bernoulli(p): Consider a sample \((x_1, \ldots, x_n)\) from the Bernoulli(p) distribution; the location and scaling of the Bernoulli are

\[
E(x) = p, \quad \text{SD}(x) = \sqrt{pq}.
\]

The marginal distribution of the sample sum \(Y = \sum x_i\) is binomial(n, p) and the distribution of \(\bar{Y} = \hat{p}\) is a rescaled binomial.

In Example 1 we saw that the distribution of \(\bar{Y} = \hat{p}\) piled up at the value p. The present Theorem 2 says that the distribution of

\[
(6) \quad z_n = \frac{Y - np}{(npq)^{1/2}} = \frac{\hat{p} - p}{(pq/n)^{1/2}}
\]

converges to the standard normal distribution.

Now suppose that we use this limiting distribution to calculate some approximate probabilities for the binomial(100, 1/2) distribution; we might, for example, be interested in the frequency \(y\) of heads for 100 tosses of an unbiased coin or in the number of 1’s in a sample of 100 random bits; we have

\[
E(y) = 50, \quad \text{SD}(y) = 5.
\]

Before using the normal approximation we should note that the binomial differs from the normal in an obvious way: the binomial has probability at the integers and the normal has probability continuously on the line. Toward compensating for this we can proceed as in Example 6.1.4 and view the binomial as having come from a continuous distribution by lumping probability to the nearest integer. Thus to approximate the probability \(P(45 \leq y \leq 55)\) we would use \(P(44\frac{1}{2} \leq y \leq 55\frac{1}{2})\) with the limiting distribution. Thus we have
\[ P(45 \leq y \leq 55) = P\left(\frac{45 - 50}{5} \leq \frac{y - 50}{5} \leq \frac{55 - 50}{5}\right) \]
\[ \approx P\left(\frac{44.1 - 50}{5} \leq z \leq \frac{55.1 - 50}{5}\right) = P(-1.1 \leq z < 1.1) \]
\[ = 0.8643 - 0.1357 = 72.9\%. \]

The normal approximation gives reasonable accuracy for \( p \) near 0.5 and \( n \geq 10 \), for \( p \) in the range 0.4 to 0.6 and \( n \geq 25 \), and for \( p \) in the range 0.3 to 0.7 and \( n \geq 100 \). For \( p \) near the extreme values 0, 1, the Poisson approximation in Section 6.1A can be used.

---

**D PROOF OF THE CENTRAL LIMIT THEOREM 2**

The ordinary exponential limit
\[ \lim_{n \to \infty} \left(1 + \frac{\delta}{n}\right)^n = e^\delta \]

can be used to show that
\[ \lim_{n \to \infty} \left(1 + \frac{\delta + \delta_n}{n}\right)^n = e^\delta, \]

where \( \{\delta_n\} \) is a null sequence (i.e., \( \lim \delta_n = 0 \)).

The existence of the mean \( \mu = E(y) \) and the inequality
\[ |iy e^{iyt}| \leq |y| \]

used with Supplement Proposition 5.8.9 give the convergence of the integral
\[ c'(t) = \frac{dc(t)}{dt} = \int_{-\infty}^{\infty} iye^{iyt} dP(y). \]

The existence of the second moment \( E(y^2) \) and the inequality
\[ |i^2 y^2 e^{iyt}| \leq y^2 \]

used with Supplement Proposition 5.8.9 give the convergence of the integral
\[ c''(t) = \frac{d^2c(t)}{dt^2} = \int_{-\infty}^{\infty} i^2y^2 e^{iyt} dP(y). \]

The continuity of the preceding integrand (as a function of \( t \)) used with Proposition 5.8.8 gives the continuity of \( c''(t) \). Thus the Taylor expansion of \( c(t) \) about the origin gives
\[ c(t) = 1 + c'(0)t + c''(\theta t) \frac{t^2}{2!} \quad |\theta| < 1, \]
\[ = 1 + i\mu t - \mu_2 \frac{t^2}{2!} + o(t^2), \]

where \( o(t^2) \) is a remainder that goes to zero as a proportion of \( t^2 \).

Now for the main part of the proof. We are concerned with the distribution of
\[ z_n = \frac{\sum y_i - n\mu}{\sqrt{n\sigma}} = \frac{y_1 - \mu}{\sqrt{n\sigma}} + \cdots + \frac{y_n - \mu}{\sqrt{n\sigma}}, \]

which is a sum of independent and identically distributed components. Note that only the deviations \( y_i - \mu \) with \( E(y_i - \mu) = 0 \) are involved; it thus suffices to prove the theorem for the case \( \mu = 0 \). We use the expression (10) and the properties of the characteristic function (Table 6.3). We take the limit of the characteristic function of \( z_n \):

**TABLE 6.3**  
Distributions and corresponding characteristic functions

<table>
<thead>
<tr>
<th>Variable</th>
<th>Characteristic function(t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y_i )</td>
<td>( 1 - \sigma^2 \frac{t^2}{2!} + o(t^2) )</td>
</tr>
<tr>
<td>( \frac{y_i}{\sqrt{n\sigma}} )</td>
<td>( 1 - \frac{t^2}{2n} + o \left( \frac{t^2}{n\sigma^2} \right) )</td>
</tr>
<tr>
<td>( z_n )</td>
<td>( \left[ 1 - \frac{t^2}{2n} + o \left( \frac{t^2}{n\sigma^2} \right) \right]^n )</td>
</tr>
</tbody>
</table>

\[ \lim_{n \to \infty} \left( 1 - \frac{t^2}{2n} + o \left( \frac{t^2}{n\sigma^2} \right) \right)^n = e^{-t^2/2}, \]

which is the characteristic function of the standard normal. It follows then by the Continuity Theorem 6.1.3 that the limiting distribution of \( z_n \) is standard normal.

**MORE GENERAL FORMS**  
**OF THE CENTRAL LIMIT THEOREM**

The Central Limit Theorem can be extended to cover independent components with different distributions. The following version is due to Lindeberg.

**THEOREM 3**

If \( y_1, \ldots, y_n \) are statistically independent, if \( y_i \) has a distribution with probability measure \( P_i \), mean \( \mu_i \), and variance \( \sigma_i^2 \), and if

\[ \sum_{i=1}^{n} \int_{|t| > k\tau_n^2} (t - \mu_i)^2 dP_i(t) \tau_n^2 \to 0 \]

for each \( k > 0 \), where \( \tau_n^2 = \sigma_1^2 + \cdots + \sigma_n^2 \), then the distribution of

\[ z_n = \frac{\sum_{i=1}^{n}(y_i - \mu_i)}{\tau_n} \]

converges to the standard normal distribution.
A stronger condition due to Liapounov gives the following:

**THEOREM 4**

If \( y_1, \ldots, y_n \) are statistically independent, if \( y_i \) has a distribution with mean \( \mu_i \), variance \( \sigma_i^2 \), and third absolute central moment \( \rho_i^3 = E(|y_i - \mu_i|^3) \), and if

\[
\frac{\rho_1^3 + \cdots + \rho_n^3}{\tau_n^3} \to 0,
\]

where \( \tau_n^2 = \sigma_1^2 + \cdots + \sigma_n^2 \), then the distribution of

\[
z_n = \frac{1}{\tau_n} \sum_{i=1}^{n} (y_i - \mu_i)
\]

has a limiting standard normal distribution.

Central Limit Theorems are also available for distributions on \( R^k \). The following is the analog of Theorem 2.

**THEOREM 5**

If \( (y_1, \ldots, y_n) \) is a sample from a distribution on \( R^k \) with mean \( \mu \) and variance matrix \( \Sigma \), then

\[
w_n = n^{-1/2} \left( \sum_{i=1}^{n} y_i - n\mu \right) = n^{1/2}(\overline{y} - \mu)
\]

has a limiting multivariate normal(0; \( \Sigma \)) distribution [convergence \( F_n \to F \): \( F_n(x) \to F(x) \) at points of continuity of \( F \)].

**F GENERAL PROOF OF THE WEAK LAW OF LARGE NUMBERS**

We now prove the Weak Law of Large Numbers assuming only that \( \mu = E(y) \) exists. From the preliminaries in Section D we have

\[
c(t) = 1 + c'(\theta)t = 1 + i\mu t + o(t), \quad |\theta| < 1,
\]

where \( o(t) \) is a remainder that goes to zero as a proportion of \( t \).

We are concerned with the distribution of

\[
\overline{y}_n = \frac{y_1}{n} + \cdots + \frac{y_n}{n},
\]

which is a sum of independent identically distributed components. The properties of the characteristic function then give

<table>
<thead>
<tr>
<th>Variable</th>
<th>( y_i )</th>
<th>( y_i/n )</th>
<th>( \overline{y}_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characteristic function(( t ))</td>
<td>( 1 + i\mu t + o(t) )</td>
<td>( 1 + i\mu \frac{t}{n} + o\left(\frac{t}{n}\right) )</td>
<td>( 1 + i\mu \frac{t}{n} + o\left(\frac{t}{n}\right)^n )</td>
</tr>
</tbody>
</table>
For any given \( z \) the characteristic function of \( Y_n \) has by (7) the limit \( e^{izt} \). But this is the characteristic function of the distribution with all probability at the point \( \mu \). Thus by the Continuity Theorem 6.1.3, the distribution of \( Y_n \) converges to \( \mu \); that is, \( Y \xrightarrow{D} \mu \).

### G EXERCISES

1. How many times should a possibly biased coin be tossed to be 95 percent sure that the proportion for heads is within 0.01 of the probability for heads? Use the Central Limit Theorem and compare with Problem 5.3.30.

2. Suppose that "1-pound" packages from a packaging machine have a weight distribution with mean 1.01 and standard deviation 0.01. Use reasonable assumptions to find the probability that a "100-pound" lot weighs less than 100 pounds.

3. Let \( (y_1, \ldots, y_n) \) be a sample from a distribution with moments \( \mu_1, \ldots, \mu_r \). Show that \( m_s \xrightarrow{D} \mu_s \) (s \( \leq r \)), where \( m_s \) is the sample s-th moment defined in Problem 5.6.7.

4. (Continuation) Let \( (y_1, \ldots, y_n) \) be a sample from a distribution with moments \( \mu_1, \ldots, \mu_{2r} \). Show that the standardized distribution of \( m_s \) converges to the standard normal.

5. Show that if \( Y_n \) converges in distribution to the constant \( c \), then \( g(Y_n) \) converges in distribution to the constant \( g(c) \), where \( g \) is a continuous function; recall the definition in Theorem 6.2.5.

6. (Continuation of Exercise 3) Let \( g \) be a continuous function of \( r \) arguments: Show that \( g(m_1, \ldots, m_r) \xrightarrow{D} g(\mu_1, \ldots, \mu_r) \).

### H PROBLEMS

7. Let \( (y_1, \ldots, y_n) \) be a sample from a distribution with mean \( \mu \) and variance \( \sigma^2 \). Show that \( s^2_n \xrightarrow{D} \sigma^2 \), where \( s^2_n \) is the sample variance; use Problem 6.

8. Let \( \{(y_{11}, y_{12}), \ldots, (y_{1n}, y_{2n})\} \) be a sample from a distribution with means \( \mu_1, \mu_2 \), variances \( \sigma_1^2, \sigma_2^2 \), and covariance \( \sigma_{12} \). Show that \( s_{12} \xrightarrow{D} \sigma_{12} \), where \( s_{12} \) is the sample covariance; use Problem 6.

9. Show that \( \bar{m}_s \xrightarrow{D} \mu_s \) (s \( \leq r \)), where \( \bar{m}_s \) is the sample s-th central moment defined in Problem 5.6.8; use Problem 6.

10. If \( \mu \) exists, show that the limiting distribution of \( \sqrt{n}(s^2_n - \sigma^2) \) is a central normal; see Problem 7 and recall Theorem 6.1.4.

11. If \( \mu_{2r} \) exists, show that the limiting distribution of \( \sqrt{n}(\bar{m}_r - \mu_r) \) is a central normal; see Problem 9 and recall Theorem 6.1.4.

12. Under the conditions of the Central Limit Theorem 2, show that \( \sqrt{n}(g(\bar{y}) - g(\mu)) \) has a limiting normal distribution \( N(0, g'(\mu)\sigma) \), where \( g \) is continuously differentiable in the neighborhood of \( \mu \).

13. Under the conditions of the Central Limit Theorem 5, show that the distribution of \( \sqrt{n}(g(\bar{y}_1, \ldots, \bar{y}_k) - g(\mu_1, \ldots, \mu_k)) \) converges to the normal \( N(0, \beta'\Sigma\beta) \) distribution, where \( g \) is continuously differentiable near \( \mu \) and where \( \beta' = (a_1, \ldots, a_k) = (\partial g/\partial \mu_1, \ldots, \partial g/\partial \mu_k) \) gives the derivatives taken at the mean \( (\mu_1, \ldots, \mu_k) \); use an obvious generalization of Problem 6.1.15.

14. Let \( (y_1, \ldots, y_k) \) have the multinomial \( (n, p_1, \ldots, p_k) \) distribution. Use the Central Limit
Theorem 5 to show that the distribution of
\[
(w_1, \ldots, w_k) = n^{-1/2}(y_1 - np_1, \ldots, y_k - np_k)
\]
converges to the multivariate normal distribution \((0, \Sigma)\), where \(a_{ij} = \rho_i(1 - \rho_i)\) and \(a_{ij} = -\rho_i\rho_j(i \neq j)\).

15 *Alternative proof of the Weak Law of Large Numbers:* Show that it suffices to prove the law for the case \(\mu = 0\).

16 *continuation* Assume that \(\mu = 0\) and for \(j = 1, \ldots, n\), define \(w_j = y_j^*\) if \(|y_j^*| \leq \delta n\) and zero otherwise; note that \(E(w_j) = \mu^*\) depends on \(n\). Show that \(\mu^* \to 0\) as \(n \to \infty\).

17 *continuation* Show that
\[
P(|\bar{y}_n - \mu^*| > t) \leq P(|\tilde{w}_n - \mu^*| > t) + P(\bar{y}_n \neq \tilde{w}_n) \leq \frac{\sigma_{n^2}}{nt^2} + nP(|y_n| > \delta n).
\]

18 *continuation* Show that
\[
\sigma_{n^2} \leq \delta n \int |x| \, dP(x)
\]
\[
P(|y_n| > \delta n) \leq \frac{1}{\delta n} \int_{|y_n| > \delta n} |x| \, dP(x).
\]

19 *continuation* Prove that
\[
P(|\bar{y}_n - \mu^*| > t) < \epsilon
\]
for \(n\) sufficiently large. Then prove the Weak Law.

**SUPPLEMENTARY MATERIAL**

Supplement Section 6.3 is on pages 554–563.

Sequences of functions on a probability space are examined and several forms of convergence are defined: convergence in probability, convergence in mean(\(r\)), and convergence almost surely. Criteria are given for convergence almost surely. The Continuity Theorem from Section 6.1 is proved. The Strong Law of Large Numbers is discussed and proved.

A concluding section shows how we have come full circle. We have been able to build up the model so that it can describe large-sample proportions and large-sample averages, which are in fact the pragmatic basis for the initial definition of probabilities and means. Reassuringly, we find that the model describes the proportions and averages in a consistent and coherent way.

**NOTES AND REFERENCES**

The Weak Law of Large Numbers and the Central Limit Theorem in Section 6.2 are concerned primarily with sequences of distributions. The Strong Law and convergence in probability in Supplement Section 6.3 are concerned with sequences of functions. The material in this chapter has been separated in accordance with this difference.

For further reading, see the following:


In Chapters 1–6 we developed introductory probability theory in a form suitable for further study of probability theory or for detailed study of statistics and its applications.

In Chapters 7–12 we shall develop statistical theory with applications. For this we assume a general background of introductory probability theory, such as that presented in the main sections of Chapters 1–6.

In Chapter 7 we shall introduce the basic ideas of statistical inference: estimation, tests of significance, and confidence intervals. We introduce these concepts in a general context, without restrictive assumptions such as the familiar one of normal distribution form. And we use results immediately available to us from the preceding chapters: specifically, symmetric probability distributions, and limiting or large-sample distributions.

By introducing the basic inference concepts in a general context we shall be able to see how inference can be applied to data without elaborate or specialized models or theory; we shall also have the freedom in later chapters to see statistical inference in a framework larger than the usual one involving the normal and related distributions. Certainly the normal distribution is important, but at best it is an approximation to a true distribution in an application. Our development acknowledges the need for methods suitable for other distribution forms and the need for methods suitable for a range of distribution forms.
7

STATISTICAL INFERENCE

7.1

THE STATISTICAL MODEL AND THE DATA

In this section we examine some simple examples of statistical models and then briefly consider some data from a biochemical investigation. Then in Section 7.2 we use large-sample theory to develop some basic inference methods; these methods are applied to the biochemical data. In Section 7.3 we examine finite population sampling, the elementary theory for sample surveys. In Section 7.4 we use symmetry to develop some standard nonparametric inference procedures. Supplement Section 7.5 looks briefly at tests of randomness, tests for the validity of a sample.

A PROBABILITY AND STATISTICS

In the preceding chapters we have been concerned with random systems as found throughout science and industry. We have seen how probability distributions can describe the variation that occurs with random systems. And we have developed in detail the theory for such probability distributions.

Probability theory is descriptive: it describes phenomena in the real world; specifically, the stable variation that occurs with random systems throughout science and industry.

A major question for any application of probability theory is—How do we determine the appropriate distribution? In a sense this is what statistics is about and is the subject matter for these next six chapters.

Statistical theory is prescriptive: it prescribes methods of determining the appropriate distribution for a random system. Specifically, it is concerned with the design of observational and experimental systems so that the things of interest are, in fact, those being investigated; it is concerned with performances of the system and the collection of the relevant data; and it is concerned with the analysis of the data and the drawing of conclusions concerning the things of interest.

How do we determine the appropriate distribution for a particular random system? In preceding chapters we have mentioned two methods. First, take a very large number of performances of the system, and use the observed proportions as estimates or approximations for the true probabilities. Second, examine the random system for symmetries, and use the available symmetries to calculate some or all of the true probabilities. In later chapters we shall develop a third basic method: check the background information for general properties that limit the range of possibilities for the
distribution, and use this limited range of possibilities together with data from a small or moderate number of performances of the system.

In this chapter we examine the first two methods, involving large samples and symmetry. We use the theory that is available to us from the preceding chapters. And we develop the process of statistical inference from observed data to conclusions concerning particular unknowns of interest.

In practice, however, large samples and symmetry are not always available. For example, a large number of repetitions can be unrealistic or uneconomical. And enough symmetry occurs rather infrequently. In Chapters 8, 9, 10, and 11 we will see how certain general properties of the system being investigated can make statistical inference feasible for a small to moderate number of repetitions of the system. As an example, a general property of the system could be that the variation in the response is approximately normally distributed, or that the variation in the response is approximately Student(\(\lambda\)) with the degrees of freedom in a moderate range, say from 2 to 15.

Then in the concluding chapter, we examine the most important part of statistics, the design of experimental investigations. In the design of experiments, we consider the choice of values for input variables of a system and the randomization of such choice against any nonrandom sources of variation. The purpose of the design of experiments is to determine cause–effect relationships.

Thus from Chapters 7 through 12 we progress from the simple use of probability concepts for statistical inference through to the central concern of scientific investigation, the determination of cause–effect relationships.

**B  STATISTICAL INFERENCE FROM LARGE SAMPLES AND SYMMETRY**

In this chapter we examine statistical inference using the theory available to us from preceding chapters, specifically the large-sample results from Chapter 6 and the symmetry results from Chapter 1.

This general approach does not prescribe any tight or strict model for an application. We are not, for example, saying that certain response variables are (exactly) normally distributed. In some ways this makes our initial approach to inference very open indeed. At times it may seem, in fact, to lack direction.

But this is as it should be. We do not wish to restrict our initial approach to inference to any narrow direction predicated by a particular model or method. Rather we wish the approach to inference to be broad and general. We will thus obtain an initial broad contact with the basic concepts of statistics; and we will see that a wide range of methods is available to us—just on the basis of large samples and symmetry.

In subsequent chapters, then, we will see how additional information and the corresponding specification for the model can lead to more directed and more specific inferences.

Our purpose now is to open as many directions for inference as possible; to have it as flexible and data- and application-based as possible; and to avoid conceptual commitment to any very specific form of modeling.

**C  THE UNKNOWNNS OF A RANDOM SYSTEM**

Consider a random system. If the distribution is known, we have a full description of the system as discussed in Chapters 1 to 6, and there is no statistical problem.
If the distribution is not known, a range of possibilities will exist for the system. Certain of the unknowns of the system may be of particular interest; consider some examples.

An investigator may be interested directly in certain probabilities of the system: for example, the probability of recovery from a certain disease using a particular drug therapy; or the probability of a lifetime greater than 5000 hours for a certain lighting unit.

Alternatively, the investigator may be interested in a characteristic of the system that is closely related to a probability: for example, the amount of the drug that gives a 50 percent probability of recovery from the particular disease; or the amount of gaseous additive that maximizes the probability of a lifetime greater than 5000 hours. These alternatives involve questions of experimental design and are examined in Chapter 12.

An investigator may be interested in characteristics that are derived from the probabilities for the system. He may, for example, be interested in the general level of a certain real-valued response variable $y$. The general level can often be measured by the mean $\mu = E(y)$. In other cases, if long tails are permitted for the response distribution, the general level may be measured more realistically by the median $\bar{y}$, the value that divides the distribution 50–50. Or he may be interested in the magnitude of the variation of a certain real-valued response variable. Often the variation can be measured by the variance $\sigma^2 = \text{var}(y)$. In other cases, if long tails are permitted for the response distribution, the variation can be measured by the range that contains the central 50 percent of the distribution—the interquartile range; or it can be measured by the distance plus or minus from the median that contains 68.26 percent of the distribution (this gives 1 standard deviation in the case of the normal).

Alternatively, the investigator may be interested in a characteristic that is closely related to the preceding characteristics: for example, in the change in the general response level caused by a change in an input variable to the system. This alternative involves questions of cause–effect relationships and is examined in Chapter 12.

The statistical model for a stable system with unknowns is the set of possible descriptions for the system. In some cases the statistical model can be the set of possible distributions for the response of the system. If there is very little information concerning the system, the set may be very large and embrace, say, all the distributions on the particular space. Or if there is a lot of information, the set may be quite small and embrace, say, the normal distributions on the real line or the normal distributions with specified variance. With such a model, some one of the distributions is the true distribution and provides a reasonable approximation for the system. The problem of statistical inference, then, is to use data from a small or moderate number of repetitions to draw conclusions concerning which distribution in the model is the true distribution, or to draw conclusions about a particular unknown of interest such as the mean or variance of the true distribution.

In other cases the investigator can describe the variation in the system by means of a probability model. The statistical model then involves the set of possibilities for the response as a function of the underlying variation; with such a model some one of the functions is the true function that provides a reasonable approximation for the system. The problem of statistical inference, then, is to use data from a small or moderate number of repetitions to draw conclusions about which function in the model is the true response function or to draw conclusions about some particular unknown of interest, say the mean or variance of the true response function.
EXAMPLES OF A STATISTICAL MODEL

Consider several examples of statistical models. The examples present a gradation in the amount of information available concerning the system being investigated.

EXAMPLE 1

Consider a biochemical system and let $y$ designate a certain real-valued response of the system as operated under specified conditions. Suppose that $y$ can be measured to a moderate number of figures so that the continuous range $R$ of possible values provides a reasonable approximation for the sample space. We would certainly be interested in intervals, so the Borel sets provide a reasonable class of events. If there are no constraints on the possible distributions, then the model would be the set $\mathcal{D} = \{P\}$ of all probability distributions $P$ on the real line. If there is information concerning continuity of the response, then the model could be the set $\mathcal{D} = \{f\}$ of all probability densities on the real line, or perhaps the set of all continuous probability densities on the real line. Note that we have shifted to a more convenient way of presenting the distribution, by density rather than measure. Also note that if we use, say, three figures to record the response, then the distribution is technically discrete, but a continuous distribution with a density function may be a very reasonable approximation. In some contexts there may be additional constraints on the distribution, say that the mean of the distribution exists, or that the mean and variance exist.

To obtain information on the system we would want observations on a sequence of independent repetitions of the system. In order to assess such observations we need the distribution for such a sequence of repetitions. Let $(y_1, \ldots, y_n)$ designate the vector response for a sequence of repetitions. A distribution for this sequence is the distribution of a sample from a distribution in $\mathcal{D}$; see Section 3.3F. Our model for the sequence is

$$\{f(y_1) \cdots f(y_n) : f \in \mathcal{D}\},$$

and we would have appropriate constraints on the possible densities $f$ as discussed in the preceding paragraph. In this chapter we see how an observed sequence can be used to make inferences concerning the true $f$ or concerning some relevant characteristic of the true $f$. Note in the display that we are concerned with the functions but that arguments have been inserted for clarity of presentation.

EXAMPLE 2

A new treatment is applied to animals of a certain breed—is applied to experimental units. Let $y$ designate the response. The variation affecting the response has been investigated under various treatments and conditions and is known to be approximately normal. Under present conditions let $\mu$ designate the general level of the response and $\sigma$ the scaling of the response. The statistical model for a treatment applied to a unit is

$$\left\{ \frac{1}{\sqrt{2\pi\sigma}} \exp \left\{ -\frac{1}{2\sigma^2}(y - \mu)^2 \right\} : \mu \in R, \sigma \in R^+ \right\},$$

using densities to present distributions on the line $R$. 
Now consider a sequence of \( n \) independent repetitions of the system, that is, the treatment applied to \( n \) experimental units. Let \((y_1, \ldots, y_n)\) designate the response sequence; the model for such a sequence is
\[
\frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum \left( y_i - \mu \right)^2 : \mu \in \mathbb{R} \right\},
\]
using densities to present distributions on \( \mathbb{R}^n \). In Chapters 8, 9, and 10 we shall examine how an observed sequence can be used to make inferences concerning the true density \( f \), that is, concerning the true values for the parameters \( \mu \) and \( \sigma \).

**EXAMPLE 3**

Consider the preceding example in somewhat greater detail. The distribution of the variation affecting the response has been identified. Let \( z \) designate this variation. In standard units the distribution for \( z \) is given by
\[
\frac{1}{\sqrt{2\pi}} \exp \left\{ -(1/2)z^2 \right\}.
\]
The model for the response is
\[
\left\{ y = \mu + \sigma z : \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+ \right\},
\]
where \( \mu \) and \( \sigma \) designate the general level and the scaling of the response \( y \).

Now consider a sequence of independent repetitions of the system. Let \((z_1, \ldots, z_n)\) designate the sequence for the variation and \((y_1, \ldots, y_n)\) designate the sequence for the response. The distribution for the variation is given by
\[
\frac{1}{(2\pi)^{n/2}} \exp \left\{ -(1/2) \sum z_i^2 \right\}.
\]
The model for the response is
\[
\left\{ y = \mu 1 + \sigma z : \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+ \right\},
\]
where \( 1 = (1, \ldots, 1)' \) designates the 1-vector. In Chapter 11 we will examine how an observed response sequence \( y \) can be used to make inferences concerning the realized sequence \( z \) and concerning the true values for the parameters \( \mu \) and \( \sigma \). These inference methods work with arbitrary distributions for the variation.

---

**E  EXAMPLE OF DATA**

In the biochemical investigation of a protein, the following \( n = 9 \) observations were obtained on a response \( y \):

\[
\begin{align*}
17.1 & \quad 14.9 & \quad 15.2 & \quad 18.1 & \quad 16.4 \\
19.1 & \quad 18.5 & \quad 14.6 & \quad 17.3.
\end{align*}
\]

We shall use these data for illustrations later in this chapter. Accordingly, it is convenient to record some simple calculations. The sample average is a simple analog
of the mean of a distribution; the sample average is calculated on the right in Table 7.1. The average vector is recorded in the second column, and the residual vector recording deviations from the average is recorded in the third column; see Figure 7.1 and recall Figure 5.11. Note that the vector \( \bar{y} \) is equal to the average vector \( \bar{y} \) plus the residual vector \( y - \bar{y} \). Also note that the squared length \( SS \) of the vector \( y \) is equal to the squared length \( SS_A \) of the average vector plus the squared length \( SS_V \) of the residual vector:

\[
\sum y_i^2 = n\bar{y}^2 + \sum (y_i - \bar{y})^2 = \frac{(\sum y_i)^2}{n} + \sum (y_i - \bar{y})^2;
\]

this is a consequence of the orthogonality of the average vector \( \bar{y} \) and the residual

![Figure 7.1](image)

Orthogonal components \( \bar{y} \) and \( y - \bar{y} \) add to give the vector \( y \). The corresponding squared lengths \( SS_A \) and \( SS_V \) add to give the squared length \( SS \).
vector \( \mathbf{y} = \bar{y} \mathbf{1} \). The sample variance and the sample standard deviation are calculated also.

Statistical inference is concerned with how to draw conclusions concerning unknowns: from (1) data such as just presented, and (2) a statistical model such as described in Section C, the model presenting the background information concerning the properties of the system being investigated.

**F EXPERIMENTAL DESIGN**

Usually an investigator must do much more than just collect data and analyze it with a statistical model. Factors that can affect the response must be controlled: some factors would have values chosen as part of seeing whether they affect the response; other factors would be kept constant. Other influences that can affect the response should be controlled as much as possible and randomization applied so that corresponding variation becomes random. Consider this further in the context of Example 2.

In Example 2 a treatment was applied to animals—to experimental units. The animals that receive the treatment should not be those "that did not get away." Rather, they should be representative of the collection or population of animals under study. Without information on all aspects of the animals, only probability theory seems able to give the appropriate balance and representativeness over the population. In its simplest form a symmetric distribution is used over the set of possible samples of a given size. This random sampling is discussed in Section 7.3.

A simple extension of Example 2 would have several different treatments. We would then have the problem of how to assign the various treatments to the experimental units that have been sampled as just described. An apparent difference in the effects of the treatments could be due to differences in the animals. Without information on all aspects of the animals, only probability theory seems able to give an appropriate balance among the treatments. In its simplest form a symmetric probability distribution is used over the set of possible assignments of treatments to experimental units, a specified number of experimental units for each treatment. This randomization is a fundamental ingredient of experimental design, of the process of determining cause–effect relationships. Other ingredients involve the control of the factors that could affect the response and the replication of a basic investigation so that treatment effects can be assessed over and above the inevitable variation from experimental unit to unit. Various aspects of experimental design are discussed in Chapter 12.

**G EXERCISES**

1. The resistance was determined for five pieces of wire sampled from production:
   
   \[ 0.138 \quad 0.140 \quad 0.136 \quad 0.142 \quad 0.140. \]
   
   Calculate the sample average, the sample average vector, the sample residual vector, and the sample variance.

2. The breaking load was determined for a sample of six from the production run of a certain plastic:
   
   \[ 18.3 \quad 19.3 \quad 21.2 \quad 22.4 \quad 17.1 \quad 27.7. \]
   
   Calculate the sample average, the sample average vector, the sample residual vector, and the sample variance.
The yield of a certain variety of corn was determined on seven plots of land:

87.7  87.2  92.3  94.2  93.6  85.6  93.5.

Calculate the sample average, the sample average vector, the sample residual vector, and the sample variance.

7.2

INFERENCE FROM LARGE SAMPLES

Consider a random system with a real-valued response $y$, for example, the biochemical system in the preceding section. Suppose that the investigator has had earlier experience with the system under various conditions and has grounds to assume that the response distribution is of reasonable form with not too much probability in the tails and with finite mean and variance. Also suppose that continuity properties of the system give grounds to assume that a continuous density function can provide a reasonable approximation to the true response distribution.

Consider an appropriate model. The sample space is clearly the real line, and the obvious class of events is the Borel class derived from the intervals. The statistical model can then be given in terms of density functions as $\mathcal{M} = \{ f \}$, where we restrict $\mathcal{M}$ in a general way to continuous densities that do not have too much probability in the tails and have a finite mean and a finite variance.

Now suppose, as is typical, that the investigator is primarily interested in the general level of the response, say, as given by the mean $\mu = E(y)$. The variance of the response $\sigma^2 = \text{var}(y)$ can also be of interest—for what it says of course about variation in the response, but also for what it says about the accuracy of an estimate of the general level $\mu$. Note that $\mu$ and $\sigma^2$ can be viewed as functions defined on $\mathcal{M}$: Each $f$ has a value for $\mu$ and a value for $\sigma^2$; such functions can be called parameters of the model. One of the distributions in $\mathcal{M}$ is the true distribution; the corresponding values of $\mu$ and $\sigma^2$ are then the true values for these parameters.

Suppose that the investigator plans a sequence of $n$ independent repetitions of the system in order to obtain information concerning the general level $\mu$. Let $(y_1, \ldots, y_n)$ designate the compound response for the sequence of repetitions. The sample space then is $\mathbb{R}^n$ and the obvious class of events is the Borel class on $\mathbb{R}^n$; the statistical model for the sequence of repetitions can be given in terms of density functions as

$$\{ f(y_1) \cdots f(y_n) : f \in \mathcal{M} \},$$

where $\mathcal{M}$ is restricted as before.

EXAMPLE 1

Biochemical example: We shall illustrate various inference procedures with the biochemical data in Section 7.1E:

17.1  14.9  15.2  18.1  16.4

19.1  18.5  14.6  17.3
with sample average $\bar{y} = 16.8$, sample variance $s_{\bar{y}}^2 = 2.672$, and sample standard deviation $s_{\bar{y}} = 1.635$.

A ESTIMATE FOR THE MEAN

We have been considering the sample average $\bar{y}$ as a natural analog of the mean $\mu$ of the distribution being sampled. The sample average $\bar{y}$ is a function defined on the sample space $R^n$; such a function can be called a statistic. The sample average $\bar{y}$ has an attractive property [formula (5.4.5)] in relation to the mean $\mu$:

(1) Unbiasedness: $E(\bar{y}) = \mu$.

Thus the distribution of $\bar{y}$ is located at the mean $\mu$ of the distribution being sampled; sometimes the values of $\bar{y}$ are larger than the true $\mu$ and sometimes smaller than the true $\mu$, but the mean or long-run average value of $\bar{y}$ is equal to the mean $\mu$, whatever value it may have.

For the biochemical illustration recorded in the introduction, the investigator can present the observed value $\bar{y} = 16.8$ as an estimate of the true mean response $\mu$. He knows that the value of $\bar{y}$ may be smaller or larger than the true mean $\mu$, but he gains some satisfaction from the property that the mean of the values he might have obtained is equal to the true mean response that he is interested in.

We record the general definition for unbiasedness:

DEFINITION 1

A function on the sample space is an unbiased estimator of a parameter of the model if the mean value of the function is equal to the value of the parameter, for each distribution in the model.

In Chapter 9 we shall discuss unbiased estimation in some detail, including how to choose a function on the sample space to obtain a good estimate.

A single value as an estimate of a parameter is not too useful unless we have some indication of its reliability. For this we think naturally of the scale of the distribution of $\bar{y}$, as given by, say, the variance or the standard deviation. For the sample average $\bar{y}$ we have [formula (5.4.5)]

(2) $\text{var}(\bar{y}) = \frac{s_{\bar{y}}^2}{n}, \quad \text{SD}(\bar{y}) = \frac{s_{\bar{y}}}{\sqrt{n}}$.

These formulas are useful if we should happen to know the standard deviation $\sigma$; we will see later that they can be useful otherwise.

Sometimes earlier experience with the system under various conditions can give a reasonably accurate value for the standard deviation $\sigma$; designate this known value by $\sigma_0$. The various conditions may, however, allow the general level $\mu$ to be different and thus not give any reasonable indication as to the current value for $\mu$. Suppose we know that the variance $s_{\bar{y}}^2$ of the distribution has the value $\sigma_0^2$. The investigator can then record the observed value of $\bar{y}$ together with the known standard deviation $\sigma_0/\sqrt{n}$ for the distribution of $\bar{y}$. 

EXAMPLE 1 (continuation)

For the biochemical example, suppose that the standard deviation \( \sigma \) is known to be approximately equal to \( \sigma_0 = 2.1 \). The investigator can then record the observed \( \bar{y} = 16.8 \) together with the standard deviation

\[
SD(\bar{y}) = \frac{2.1}{\sqrt{9}} = 0.7.
\]

This standard deviation indicates how far the estimate can be away from the true mean \( \mu \). It does this in a weak way by the Chebyshev inequality in Section 5.3C. And it does so in a somewhat stronger way if a normal approximation is appropriate; we shall examine this in Section C.

B TEST FOR THE MEAN: VARIANCE KNOWN

Consider the random system as discussed in the introduction and suppose that earlier experience with the system has given the value \( \sigma_0^2 \) for the variance of the distribution. Sometimes in applications a value \( \mu_0 \) for the general level \( \mu \) may be suggested by some theoretical calculations or by some related empirical investigations with the same or a similar system. The investigator may then be interested in whether the value \( \mu_0 \) applies to the system under current conditions. More specifically he may be interested in whether observed data are consistent with such a suggested value.

Note that we have used a subscript on \( \sigma \) to denote a known value and the same subscript on \( \mu_0 \) to denote a suggested or contemplated value. For such usages the context will provide the distinction.

The Central Limit Theorem 6.2.2 says that the sample average \( \bar{y} \) appropriately standardized has a limiting standard normal distribution. And empirical investigations show that the approach to normality is quite fast, provided there is not too much probability in the tails of the distribution. Thus for the present model we assume that \( \bar{y} \) is approximately normal(\( \mu, \sigma_0/\sqrt{n} \)) or that

\[
z = \frac{\bar{y} - \mu}{\sigma_0/\sqrt{n}}
\]

is approximately standard normal.

For our illustration a sample of size \( n = 9 \) is somewhat small and a size of \( n = 20, 30, \ldots \) would be preferable. For a simple illustration, however, we assume that the normal distribution is applicable; that is,

\[
z = \frac{\bar{y} - \mu}{0.7}
\]

is approximately standard normal.

We now determine the kind of sample average \( \bar{y} \) to expect if the mean \( \mu = \mu_0 \). If \( \mu = \mu_0 \), then the distribution of \( \bar{y} \) is located at \( \mu_0 \), is scaled by \( \sigma_0/\sqrt{n} \), and has approximately normal form. Or, equivalently, if \( \mu = \mu_0 \), then the distribution of

\[
(3) \quad z = \frac{\bar{y} - \mu_0}{\sigma_0/\sqrt{n}}
\]
is approximately standard normal. We can be very conservative and use the Chebyshev inequality, obtaining

$$P(|z| \geq 2) \leq 1/4,$$

$$P(|z| \geq 3) \leq 1/9.$$  \hspace{1cm} (4)

Or, more realistically, we can accept the normal approximation, obtaining

$$P(|z| \geq 1.96) \approx 5\%,$$

$$P(|z| \geq 2.58) \approx 1\%;$$  \hspace{1cm} (5)

the values are obtained from Appendix II.

We can now check whether observed data are consistent with the suggested or hypothesized value \( \mu_0 \) for the mean. We calculate the observed value of

$$z = \frac{\bar{y} - \mu_0}{\sigma_0 / \sqrt{n}}.$$

We compare it with the standard normal distribution, and we assess the hypothesis \( \mu = \mu_0 \) accordingly. If the observed value of \( z \) is reasonable for the distribution, then our data are in general conformity with the hypothesis \( \mu = \mu_0 \). If the observed value is an extreme, an unusual, or a virtually impossible value for the distribution, then we would have evidence against the hypothesis \( \mu = \mu_0 \) with corresponding strength.

EXAMPLE 1 (continuation)

For the biochemical example, some theoretical calculations have indicated a value 14.8 for the mean \( \mu \). We test this indicated value, and thus by implication test the theory that was used for the calculations. The observed value of \( \bar{y} \) is 16.8, and the corresponding value for \( z \) is

$$z = \frac{16.8 - 14.8}{0.7} = 2.86;$$

![Standard normal distribution and observed value z = 2.86 under the hypothesis \( \mu = 14.8 \).](image)
this value is compared with the standard normal distribution in Figure 7.2. The observed value $z$ is not in the center of the distribution—in fact, it is well out on the right tail of the distribution. Thus the observed value of $z$ is not the sort of value we could reasonably expect under the hypothesis, and it provides moderately strong evidence against the hypothesis. The operating principle in this argument is: if $A$ implies $B$, then $B^c$ implies $A^c$. One way of assessing the observed value is to calculate the probability, under the hypothesis, of a value as far or farther from the center of the distribution:

$$P(|z| \geq 2.86) \simeq 0.42\%.$$  

This is called the observed level of significance and it measures the significance of the observed value in relation to the hypothesis $\mu = 14.8$.

In general, if the observed value of $z$ in formula (3) is in the central part of the distribution, then, as assessed, the data are in accord with the hypothesis $\mu = \mu_0$. Of course this does not say that the hypothesis $\mu = \mu_0$ is true; in fact, we would not expect $\mu$ to be exactly equal to $\mu_0$. Rather it is saying that the assessment does not give evidence against the value $\mu_0$ as a reasonable approximation to the true $\mu$.

On the other hand, if the observed value of $z$ in formula (3) is very far out on a tail of the distribution, the data are substantially contradicting the hypothesis $\mu = \mu_0$, and we have substantial evidence against the hypothesis $\mu = \mu_0$. In summary, we would say that the observed $\bar{y}$ and the corresponding $z$ are statistically significant in relation to the hypothesis $\mu = \mu_0$. However, in certain of these cases the value indicated for $\mu$ by the estimate $\bar{y}$ may, in fact, differ from $\mu_0$ by an amount that makes no practical difference; in that case the statistical significance is of no practical significance.

One way of obtaining a numerical assessment of an observed value in relation to a distribution is to calculate the probability of a value as far or farther from the center of the distribution as that observed; this is called the observed level of significance. Such a calculation is reasonable typically with distributions that are symmetrical about a central point and drop off toward the tails.

C INTERVAL ESTIMATE FOR THE MEAN: VARIANCE KNOWN

Consider the random system discussed earlier and suppose that earlier experience with the system has given the value $\sigma_x^2$ for the variance $\sigma^2$ of the distribution.

In Section A we examined $\bar{y}$ as an estimate of $\mu$ and we noted that the standard deviation $\sigma_0/\sqrt{n}$ gives some indication of its reliability. We are now in a position to be more precise. We shall test various values for $\mu$ and see which ones are reasonable by the assessment procedures in Section B. Suppose that for each value for $\mu$ we determine a 95 percent interval from the distribution of the criterion

$$\frac{\bar{y} - \mu}{\sigma_0/\sqrt{n}};$$

and then view this as an interval of acceptable values for our criterion. There is some moderate advantage in having the same interval for each hypothesized value; and the interval $(-1.96, +1.96)$ is a rather natural choice based on the approximate distribution.
EXAMPLE 1 (continuation)

Consider an interval estimate for $\mu$ in the biochemical example. A value $\mu$ is acceptable, as just discussed, if

$$-1.96 \leq \frac{16.8 - \mu}{2.1/\sqrt{9}} \leq 1.96$$

or, equivalently, if

$$16.8 - 1.96 \frac{2.1}{\sqrt{9}} \leq \mu \leq 16.8 + 1.96 \frac{2.1}{\sqrt{9}}.$$

Thus we see that the values of $\mu$ that are acceptable on the basis of the data form the interval

$$(8) \quad \left(16.8 \pm 1.96 \frac{2.1}{\sqrt{9}}\right) = (15.4, 18.2).$$

Note that the interval is $(\bar{y} \pm 1.96$ standard deviations of $\bar{y})$. We thus have an interval estimate of the mean $\mu$ as based on a 95 percent acceptance level.

Now consider the general random system. A value $\mu$ is acceptable as discussed if

$$-1.96 \leq \frac{\bar{y} - \mu}{\sigma_0/\sqrt{n}} \leq 1.96$$

or, equivalently, if

$$\bar{y} - 1.96 \frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \bar{y} + 1.96 \frac{\sigma_0}{\sqrt{n}}.$$

Thus we obtain the interval estimate

$$(9) \quad \left(\bar{y} \pm 1.96 \frac{\sigma_0}{\sqrt{n}}\right)$$

for the mean as based on a 95 percent acceptance level.

We can obtain the same interval estimate by a different route that provides an interesting interpretation. We follow the same preliminaries as before but consider the function $\bar{y}$ rather than the observed value from the data. We can make the probability statement

$$(10) \quad P\left(-1.96 \leq \frac{\bar{y} - \mu}{\sigma_0/\sqrt{n}} \leq 1.96\right) \approx 95\%;$$

this holds for each value for the mean $\mu$ of the underlying distribution. We then rearrange the inequalities as before, obtaining

$$(11) \quad P\left(\bar{y} - 1.96 \frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \bar{y} + 1.96 \frac{\sigma_0}{\sqrt{n}}\right) \approx 95\%.$$ 

Thus the probability is approximately 95 percent that the interval

$$(12) \quad \left(\bar{y} - 1.96 \frac{\sigma_0}{\sqrt{n}}, \bar{y} + 1.96 \frac{\sigma_0}{\sqrt{n}}\right)$$
brackets the true mean $\mu$. Note that the interval is a function of the sample $(y_1, \ldots, y_n)$ and thus has a distribution; the probability statement is made on the basis of this distribution. For observed data we can then calculate the observed value of the interval (12); it is called a 95 percent confidence interval for the parameter $\mu$.

EXAMPLE 1 (continuation)

For the biochemical data we obtain the interval

$$\left(16.8 \pm 1.96 \frac{2.1}{\sqrt{9}}\right) = (15.4, 18.2)$$

as before; it is a 95 percent confidence interval for the true mean $\mu$. Of course, either the interval brackets the true mean $\mu$ or it does not; the investigator does not know. But he gains confidence from the property that approximately 95 percent of the intervals he might have obtained would bracket the true mean $\mu$.

A long-run interpretation can be given for this confidence-interval procedure. Suppose that we contemplate repeated samples of 9 from the biochemical system under investigation, and in each case calculate the observed value of the confidence interval (12); then we would obtain results as presented in Figure 7.3. The probability statement (11) asserts that in the long run, approximately 95 percent of these intervals will intersect the line representing the true value for the parameter $\mu$. More on confidence intervals in Sections 8.3, 10.5, and 11.1.

D ESTIMATE FOR THE VARIANCE

Consider the random system as discussed in the introduction and suppose now that the variance $\sigma^2$ is unknown. The investigator may be interested in the variance $\sigma^2$ for what it says about the variation in the response or for what it says about the accuracy of an estimate such as $\bar{Y}$ for the general level $\mu$. We have considered the sample variance $s^2_Y$ as a natural analog of the variance $\sigma^2$ of the distribution being sampled. The sample variance $s^2_Y$ is a function defined on the sample space $\mathcal{R}^n$; it has the attractive property (5.4.9) of unbiasedness in relation to the variance $\sigma^2$:

$$\text{(13) Unbiasedness: } E(s^2_Y) = \sigma^2.$$
Thus the values for $s^2_y$ are sometimes larger and sometimes smaller than the true $\sigma^2$, but the mean value is equal to the variance $\sigma^2$ whatever value it may have.

**EXAMPLE 1 (continuation)**

For the biochemical example the investigator can present the observed value $s^2_y = 2.672$ as an estimate of the true variance $\sigma^2$. He knows that the value may be smaller or larger than the true variance $\sigma^2$, but he gains some satisfaction from the unbiasedness property concerning what he might have obtained.

For the reliability of the estimator $s^2_y$ we might naturally turn to the scale of its distribution as given by, say, the variance: we have

\[(14) \quad \text{var} (s^2_y) = \frac{1}{n} \left( \mu_4 - \frac{n - 3}{n - 1} \sigma^4 \right) = \frac{1}{n} \left( \kappa_4 + \frac{2n}{n - 1} \sigma^4 \right)\]

(Problem 11). But this then requires information concerning the fourth central moment or the fourth cumulant.

For the standard deviation $\sigma$ the investigator will probably use the sample standard deviation $s_y$. Unfortunately, however, it does not have the unbiasedness property; in fact,

\[(15) \quad \sigma > E(s_y),\]

as is easily seen from the following:

\[\text{var} (s_y) = E(s^2_y) - (E(s_y))^2 = \sigma^2 - (E(s_y))^2,\]

\[\sigma^2 - (E(s_y))^2 > 0.\]

What we are seeing here, in fact, is an unattractive property of unbiasedness—that unbiasedness is tied to a particular way of scaling or expressing a parameter of a distribution. In Chapters 9 and 10 we shall consider further some questions of estimation and reliability for the variance.

**E TEST FOR THE MEAN: VARIANCE UNKNOWN**

Consider the random system as discussed in the introduction to this section and suppose that the variance $\sigma^2$ is unknown. We now consider a test for the hypothesis $\mu = \mu_0$.

The Central Limit Theorem 6.2.2 says that

\[z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}\]

has a limiting standard normal distribution. Problems 6.2.7 and 6.3.5 show that $s_y$ converges to $\sigma$ in probability and almost surely as $n \to \infty$. It follows, then, by Problem 6.1.16 that

\[(16) \quad t = \frac{\bar{Y} - \mu}{s_y/\sqrt{n}}\]
has a limiting standard normal distribution as \( n \to \infty \). Empirical investigations, however, show that the approach to normality can be quite slow: the approach is strongly influenced by the slowness of convergence of \( s_y \) to \( \sigma \), particularly if the distribution is substantially different from normal form (see, for example, Section 8.3B). In the present context we assume that the sample size is sufficiently large that the normal distribution provides a reasonable approximation.

We now determine approximately the kind of sample average to expect under the hypothesis \( \mu = \mu_0 \). If \( \mu = \mu_0 \), then

\[
t = \frac{\bar{y} - \mu_0}{s_y/\sqrt{n}}
\]

is approximately standard normal, and we have

\[
P(|t| \geq 2) \approx 5%.
\]

\[
P(|t| \geq 3) \approx 1%.
\]

Note that we acknowledge the rough approximation by using the simple values 2 and 3 rather than the values 1.96 and 2.58 for the normal.

We can now check whether observed data are consistent with the hypothesized value for \( \mu_0 \). We calculate the observed value of \( t \) in formula (17), compare it with the standard normal, and assess the hypothesis accordingly.

---

**EXAMPLE 1 (continuation)**

For the biochemical example we have

\[
t = \frac{16.8 - 14.8}{1.64/\sqrt{9}} = 3.66.
\]

The observed value is well out on the right tail of the distribution, thus providing moderate evidence against the hypothesis: \( \mu = 14.8 \). The strength of such evidence, however, should be substantially qualified by the rough approximation involved with such a small sample size \( n = 9 \).

---

**F INTERVAL ESTIMATE FOR THE MEAN: VARIANCE UNKNOWN**

Consider further the random system as discussed in the introduction and suppose that the variance \( \sigma^2 \) is unknown. The distribution properties discussed in Section E give the following probability statement:

\[
P\left(-2 \leq \frac{\bar{y} - \mu}{s_y/\sqrt{n}} \leq +2 \right) \approx 95% 
\]

or, equivalently,

\[
P\left(\bar{y} - 2 \frac{s_y}{\sqrt{n}} \leq \mu \leq \bar{y} + 2 \frac{s_y}{\sqrt{n}} \right) \approx 95%.
\]
Thus
\[ \bar{y} \pm 2s_y/\sqrt{n} \]
is an approximate 95 percent confidence interval for the parameter \( \mu \).

**EXAMPLE 1 (continuation)**

For the biochemical example we have the observed interval
\[ 16.8 \pm 2 \frac{1.64}{\sqrt{9}} = (15.7, 17.9). \]

This is an approximate 95 percent confidence interval for the parameter \( \mu \); but we should be mindful of the rough approximation involved with such a small sample size. This interval can also be obtained by testing each possible parameter value \( \mu \) against the data as in Section C.

For these approximate confidence intervals it is reasonable to use the simple values 2 and 3, corresponding to 95 percent and 99 percent confidence. More on confidence intervals in Sections 8.3, 10.5, and 11.1.

**G INFERENCES OR DECISIONS**

Consider briefly the testing of a hypothesis \( \mu = \mu_0 \). Some developments of statistical theory suggest that on the basis of the data a decision should be made—to accept the hypothesis or to reject the hypothesis. The theoretical problem, then, is a matter of choosing a decision rule that associates one or other decision with each possible set of data. For example, in Section B a rule might be to accept the hypothesis if the value of \( z \) falls in \((-1.96, +1.96)\) and to reject the hypothesis otherwise. With this rule the decision "reject" would be obtained for the biochemical data with \( \mu_0 = 14.8 \).

There is certainly some appeal to a categorical statement. However, as we have seen, the only instances where there would be substantial evidence to support a categorical statement would be those where the deviation was "impossibly" large, as assessed under the hypothesis. The use of a decision rule to "decide" on truths of hypothesis is perhaps due to the attractiveness of a mathematically very elegant theory; we shall discuss some aspects of this theory in Chapter 10. In practice, however, the use of such rules is essentially unscientific (making categorical assertions without supporting evidence), and the hazards of having scientific results asserted on such a basis are substantial. For some observations on how decision rules can interact with editorial policies of scientific journals, see Sterling (1959).

**H EXERCISES**

It is important to make accurate numerical calculations.

1. Sixteen measurements are made on a chemical response, yielding \( \bar{y} = 285.3 \). The response distribution is known to be of reasonable form with \( s \) approximately 3.1. Test the hypothesis \( \mu = 280 \). Calculate a 95 percent confidence interval for \( \mu \). Calculate a 99 percent confidence interval for \( \mu \).

2. The resistances of five pieces of wire sampled from production are recorded in Exercise 7.1.1. The standard deviation is known to be approximately 0.002 and the distribution form
to be such that the limiting normal distribution is reasonable for this sample size. Test the hypothesis that \( \mu = 0.135 \).

3 The breaking loads for a sample of 6 are recorded in Exercise 7.1.2. The standard deviation is known to be approximately 2.5 and the distribution form to be such that the normal distribution is reasonable for this sample size. Calculate a 95 percent interval for the mean breaking load \( \mu \).

4 The yield for a certain variety of corn on 7 plots is recorded in Exercise 7.1.3. The distribution is known to be such that the normal approximation for \( t \) in formula (16) is roughly adequate. Test the hypothesis that \( \mu = 90 \). Calculate an approximate 95 percent confidence for the mean yield of the particular corn under the particular conditions.

5 A sample of 11 rats was obtained randomly from a population of rats. A blood viscosity test of each yielded 3.29, 3.91, 4.64, 3.55, 3.67, 4.18, 3.74, 4.67, 3.03, 4.61, and 3.84. On the assumption that the distribution has reasonable form, test the hypothesis: \( \mu = 4.6 \). Calculate an approximate 95 percent confidence interval for \( \mu \). Calculate an approximate 99 percent confidence interval for \( \mu \).

I TWO-SAMPLE PROBLEMS

Consider a sample \( \{y_{11}, \ldots, y_{ia}\} \) from a first response and an independent sample \( \{y_{12}, \ldots, y_{ib}\} \) from a second response. Let \( \mu_1 \) and \( \sigma_1^2 \) be the mean and variance for the first response and \( \mu_2 \) and \( \sigma_2^2 \) be the mean and variance for the second response. And let \( \bar{y}_1 \) and \( s_1^2 \) be the sample average and variance for the first sample and \( \bar{y}_2 \) and \( s_2^2 \) be the sample average and variance for the second sample.

The difference \( \bar{y}_2 - \bar{y}_1 \) is the natural sample analog of the difference \( \mu_2 - \mu_1 \) between the mean responses. From Chapter 5 we know that \( \bar{y}_2 - \bar{y}_1 \) has mean \( \mu_2 - \mu_1 \) and variance \( \sigma_1^2/m + \sigma_2^2/n \).

The results in Chapter 6 can be used to show that

\[
\left( \frac{\bar{y}_2 - \bar{y}_1 - (\mu_2 - \mu_1)}{\left(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)^{1/2}} \right)
\]

has a limiting standard normal distribution as \( m, n \to \infty \) in a given ratio.

With some additional argument, the results in Chapter 6 show that

\[
\left( \frac{\bar{y}_2 - \bar{y}_1 - (\mu_2 - \mu_1)}{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^{1/2}} \right)
\]

has a limiting standard normal distribution.

If it is known that \( \sigma_1 = \sigma_2 \), then the sample variances can be combined in a natural way (discussed in Chapter 9) to give the combined estimate of variance,

\[
s^2 = \frac{\sum_{i=1}^{m}(y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{n}(y_{2i} - \bar{y}_2)^2}{m + n - 2} = \frac{(m - 1)s_1^2 + (n - 1)s_2^2}{m + n - 2}.
\]

The results in Chapter 6 can then be used to show that

\[
\left( \frac{\bar{y}_2 - \bar{y}_1 - (\mu_2 - \mu_1)}{s(1/m + 1/n)^{1/2}} \right)
\]

has a limiting standard normal distribution.

6 (continuation of Exercise 1) In addition, 16 measurements with a treatment input were made on the response yielding \( \bar{y}_1 = 288.5 \). The treatment is known not to alter the response variance. Test the hypothesis: \( \mu_1 = \mu = 0 \). Calculate an approximate 95 percent confidence interval for \( \delta = \mu_1 - \mu \).

7 (continuation of Exercise 5) In fact, 22 rats were randomly obtained from the population and
11 were assigned randomly as controls, yielding the data in Exercise 5, and the remaining were given an ephedrin treatment with data as follows: 3.45, 4.26, 4.71, 3.14, 3.45, 5.01, 4.43, 4.91, 4.22, 4.83, and 3.55. On the assumption that the treatment distribution has reasonable form, test the hypothesis: \( \mu_t - \mu = 0 \). Calculate an approximate 99 percent confidence interval for \( \delta = \mu_t - \mu \).

\( \text{(continuation)} \) Assume that the variances, with and without treatment, are equal. Test the hypothesis that \( \mu_t - \mu = 0 \). Calculate an approximate 95 percent confidence interval for \( \delta = \mu_t - \mu \).

Record the general formula for the confidence interval in Problem 7 for the case of samples of size \( m \) and \( n \); with possibly different variances.

Record the general formula for the confidence interval in Problem 8 for the case of samples of size \( m \) and \( n \); with equal variances.

**J PROBLEMS**

11 Prove formula (14) for var (\( s^2 \)); use \( \bar{x}_d = x_4 + 3d^4 \).

12 **Jensen inequality:** A real-valued function \( c \) on an open interval is called convex if \( ac(s_1) + (1 - a)c(s_2) \geq c(as_1 + (1 - a)s_2) \) for \( 0 < a < 1 \); the function is called strictly convex if the inequality is strict. Some analysis shows that any point \( (s_0, c(s_0)) \) on the graph of a convex function has a line of support, a line \( l \) through it such that \( c(s) \geq l(s) \); and if \( c \) is strictly convex, then \( c(s) > l(s) \) for \( s \neq s_0 \). Given a distribution for \( s \) on the open interval, show that \( E(c(s)) \geq c(E(s)) \); mean convex function is greater than or equal to convex function of mean. If \( c \) is strictly convex, show that the inequality is strict unless all probability is at a point.

13 Show that \( c^2 \geq (E(s))^2 \) unless the distribution of \( s^2 \) has all probability at a point; use Problem 12.

14 Consider a distribution for \( y \) with all probability on the positive axis. Use Problem 12 to show that

\[
E\left( \frac{1}{y} \right) \geq \frac{1}{E(y)}
\]

with equality if and only if \( y \) has all probability at a point.

15 Consider a distribution for \( y \) on \( R \) with mean \( \mu \). Show that for \( t \neq 0 \),

\[
E(e^{yt}) \geq e^{yt}
\]

with equality if and only if \( y \) has all probability at a point.

### 7.3 INFERENCE FOR FINITE POPULATIONS

We now examine finite population sampling, an important area of inference that uses both large-sample and symmetry results from probability theory. Finite population sampling enters into almost all parts of statistics—scientific, industrial, commercial, and governmental. This ranges: From the scientist who samples experimental units for the application of treatments of interest; for example, people for a test of ascorbic acid effects, or mice for a test of cancer treatments, or fruit flies for a test of radiation effects. Through the industrial statistician who samples a day’s production, or an incoming lot of a manufactured material. To the statistician in market research, opinion polls, or government information bureaus who wants information about people or production or resources.
A population may consist of the people of a given geographical district, or the houses in a certain city, or the rabbits of a particular genetic strain in a certain laboratory, or the transistors received on a particular contract, or the psychology students at a certain university.

Consider a population under investigation. The population must be well defined so that the elements are in some way clearly specified. As part of this, the size of the population is typically known to reasonable accuracy; let \( N \) designate the size. An element in the population will usually have several characteristics or variables of possible interest; typically one of the variables will be of central interest—we consider here the case of a single real variable. Let \( c_1, \ldots, c_N \) designate the values of the variable for the \( N \) elements in the population. We then abstract and take the finite population to be the set

\[
\Pi = \{c_1, \ldots, c_N\}.
\]

In general an investigator will have some interest in knowing all the values in the population. In practice, however, he will be concerned with one or several derived characteristics. Almost certainly he will be interested in the location of the values in the population; for this we define the mean of \( \Pi \) to be the average value in the population:

\[
\mu = \frac{\sum_{i=1}^{N} c_i}{N}.
\]

And typically he will also be interested in the scaling or variability in the population—for its own sake or for the information it gives concerning the reliability of an estimate of \( \mu \); for this we define the variance of \( \Pi \):

\[
\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (c_i - \mu)^2 = \frac{1}{N-1} \left( \sum_{i=1}^{N} c_i^2 - \left( \sum_{i=1}^{N} c_i \right)^2 / N \right).
\]

The division by \( N - 1 \) rather than \( N \) has certain computational advantages; and with any large population it makes no significant numerical difference. Note that as yet we have not mentioned any probability distribution.

An obvious way of obtaining information concerning \( \mu \) and \( \sigma^2 \) is by means of a census, contacting and obtaining the value of each element in the population. A census of a large population can be very expensive and time consuming. And it can often be substantially inaccurate, owing to operational difficulties—such as that of obtaining skilled enumerators.

In some populations the elements are available in a serial order, such as in a directory or by geographical layout. Sometimes, then, systematic sampling is proposed in which every tenth or every fiftieth element is examined. With systematic sampling there are potential hazards, as, for example, if unexpected patterns exist between the variable of interest and the criterion underlying the serial ordering of the population.

In this section we consider simple random sampling from the population: we impose a symmetric probability distribution on the set of all possible samples of a given size.

**A SIMPLE RANDOM SAMPLING**

A random sample of size \( n \) from the population \( \Pi \) is a sequence \( (y_1, \ldots, y_n) \) with probability function
\( p(y_1, \ldots, y_n) = 1/N^n \)

for each of the \( N^n \) possible sequences of \( n \) elements from the population \( \Pi \); note for this that we are treating the elements in the population as distinguishable even though some \( c_\alpha \) values may be equal. This random sampling involves a symmetric distribution on the set \( S_\alpha \) as defined in Section 3.4B.

The symmetry in the distribution for \((y_1, \ldots, y_n)\) leads to a simple method for performing the sampling. The first element \( y_1 \) can be any element in the population, each with the same probability \( 1/N \). And for given \( y_1 \), the second element \( y_2 \) can be any of the remaining elements in the population, each with the same probability \( 1/(N-1) \), and so on. Now suppose that serial numbers have been applied to the elements in the population. Random numbers can then be used to randomly sample the population: Use a block size of random numbers big enough to cover all the serial numbers; take blocks of random numbers until an actual serial number is obtained and then examine the corresponding population elements; continue in this way until a sample of size \( n \) is obtained. For some examples, see Section 4.2I.

The randomness in random sampling can be expressed explicitly by means of a matrix of indicators. Let \( I_{i\alpha} \) be equal to 1 if the \( i \)th sample element is the \( \alpha \)th element \( c_\alpha \) in \( \Pi \) and zero otherwise; then

\[
y_i = \sum_{\alpha=1}^{N} I_{i\alpha} c_\alpha
\]

or in matrix form,

\[
\begin{pmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{pmatrix}
= \begin{pmatrix}
I_{11} & \cdots & I_{1N} \\
I_{21} & \cdots & I_{2N} \\
\vdots & \vdots & \vdots \\
I_{n1} & \cdots & I_{nN}
\end{pmatrix}
\begin{pmatrix}
c_1 \\
c_2 \\
\vdots \\
c_N
\end{pmatrix}
\]

Note that choosing the first sample element is equivalent to choosing the position for the single 1 in the first row; that choosing the second sample element is equivalent to choosing the position for the single 1 in the second row—any position except that corresponding to the 1 in the first row; and so on.

The indicators satisfy the obvious relations:

\[
I_{i\alpha} I_{i\beta} = 0, \quad i \neq j, \quad \alpha \neq \beta,
\]

\[
I_{i\alpha} I_{i\alpha} = I_{i\alpha}.
\]

The probability function for the indicator matrix has some simple marginal probabilities:

\[
P(I_{i\alpha} = 1) = \frac{1}{N},
\]

\[
P(I_{i\alpha} I_{i\beta} = 1) = \frac{1}{N(N-1)}, \quad i \neq j, \quad \alpha \neq \beta,
\]

and so on. For some complex calculations the indicators can make certain calculations mechanical. For the proofs here, however, we find it easier to argue directly.

**B THE BASIC RESULTS FOR RANDOM SAMPLING**

The basic results for random sampling can be assembled in the following theorem:
THEOREM 1

If \((y_1, \ldots, y_n)\) is a random sample from a population of size \(N\) with mean \(\mu\) and variance \(\sigma^2\), then

\[
E(\bar{y}) = \mu, \quad \text{var}(\bar{y}) = \frac{\sigma^2}{n} \left(1 - \frac{n}{N}\right),
\]

\[
E(s_y^2) = \sigma^2.
\]

And if \(n \to \infty\) and \(N/n \to \infty\) together with mild regularity conditions for \(\Pi\), then

\[
\frac{\bar{y} - \mu}{\frac{\sigma}{\sqrt{n}} \left(1 - \frac{n}{N}\right)^{1/2}} \quad \frac{\bar{y} - \mu}{\frac{s_y}{\sqrt{n}} \left(1 - \frac{n}{N}\right)^{1/2}}
\]

have a limiting standard normal distribution.

The proof of the first part of the theorem is given in Section C. The proof of the second part is more difficult and is omitted; it can be derived from the Wald–Wolfoowitz–Hoefding theorem for permutation distributions; see, for example, Fraser (1957, p. 239). The approximate normality is often adequate for medium to large samples.

EXAMPLE 1

A random sample of 25 from a population of 500 medical students gives the following reduced data:

\[
\bar{y} = 126.0, \quad s_y = 7.5.
\]

Some related investigations in other professional groups suggest the value \(\mu = 120\). We can test the hypothesis \(\mu = 120\) by calculating the standardized deviation

\[
z = \frac{126.0 - 120}{7.5 \left(1 - \frac{25}{500}\right)^{1/2}} = 4.10
\]

and comparing it with the standard normal distribution. This value is an extreme value for the standard normal and gives moderately strong evidence against the hypothesis.

Now consider a confidence interval for \(\mu\). At the 95 percent level we have the interval \((-1.96, 1.96)\) for the standard normal. Thus

\[
P \left(\bar{y} - \frac{s_y}{\sqrt{25}} \left(1 - \frac{25}{500}\right)^{1/2} \leq 1.96 \right) = 95%.
\]

which can be rearranged to give

\[
P \left(\bar{y} - 1.96 \frac{s_y}{\sqrt{25}} \left(1 - \frac{25}{500}\right)^{1/2} \leq \mu \leq \bar{y} + 1.96 \frac{s_y}{\sqrt{25}} \left(1 - \frac{25}{500}\right)^{1/2} \right) = 95%.
\]

Thus a 95 percent confidence interval for \(\mu\) is
\[
(126.0 \pm 1.96 \frac{7.5}{\sqrt{25}} \left(1 - \frac{25}{500}\right)^{1/2}) = (123.1, 128.9).
\]

\section*{C PROOFS}

We now prove the first part of Theorem 1. For the sample average \( \bar{y} \) we have
\[
E(\bar{y}) = \frac{1}{n} \sum_{i=1}^{n} E(y_i) = \frac{1}{n} \cdot n \cdot \sum_{i=1}^{N} c_i = \mu;
\]
this uses the linearity of \( E \) and the fact that a \( y_i \) can take each value \( c_i \) with equal probability \( 1/N \).

For the next two formulas it is convenient to make a preliminary simplification. Suppose that we subtract \( \mu \) from each element in \( \Pi \) and obtain a modified population \( \Pi_0 = \{c_i - \mu\} = \{c_i - \mu\} \). Then any sample \( (y_1, \ldots, y_n) \) from \( \Pi \) has a corresponding sample \( (w_1, \ldots, w_n) \) = \( (y_1 - \mu, \ldots, y_n - \mu) \) from \( \Pi_0 \), and clearly
\[
\bar{w} = \bar{y} - \mu, \quad s_{\text{w}}^2 = s_{\text{y}}^2.
\]
Thus \( \text{var} (\bar{w}) = \text{var} (\bar{y}) \) and \( E(s_{\text{w}}^2) = E(s_{\text{y}}^2) \). It follows then that it suffices to prove the formulas for the case \( \mu = 0 \).

We assume that \( \mu = 0 \). As an immediate consequence, we obtain
\[
\sum_{i=1}^{N} c_i = 0, \quad \sum_{i=1}^{N} c_i^2 + \sum_{i \neq j} c_i c_j = 0.
\]

Then for the variance of \( \bar{y} \) we have
\[
\text{var} (\bar{y}) = E(\bar{y}^2) = \frac{1}{n^2} E \left( \sum_{i,j=1}^{n} y_i y_j \right) - \frac{1}{n^2} E \left( \sum_{i \neq j}^{n} y_i y_j \right)
\]
\[
= \frac{1}{n^2} \left( \sum_{i=1}^{n} c_i^2 \frac{1}{N} + n(n-1) \frac{\sum_{i \neq j} c_i c_j}{N(N-1)} \right) - \frac{1}{n^2} \left( n \frac{(N-1)s^2}{N} + n(n-1) \frac{\sigma^2 (N-1)}{N(N-1)} \right)
\]
\[
= \frac{\sigma^2}{n} \left( 1 - \frac{n}{N} \right).
\]

The factor \( 1 - n/N \) is called the \underline{finite population correction factor}; for example, it forces \( \text{var} (\bar{y}) \) to zero if the sample becomes a census \( (n = N) \).

And for the mean of \( s_{\text{y}}^2 \), we have
\[
E(s_{\text{y}}^2) = \frac{1}{n-1} E \left( \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \right)
\]
\[
= \frac{1}{n-1} \left( n \frac{(N-1)s^2}{N} - n \frac{\sigma^2}{n} \left( 1 - \frac{n}{N} \right) \right) = \sigma^2.
\]

\section*{D EXERCISES}

1 The registration records of a university were used to obtain a random sample of 100 undergraduates from the 12,352 full-time undergraduate students. Each was interviewed
and the value of a numerical quantity determined. The sample mean and standard deviation are $\bar{y} = 1452$, $s_y = 485$. Determine a 95 percent confidence interval for the population mean $\mu$.

2 Consider a population of size $N$ that consists of 0's and 1's only; the population mean $\mu$ then is $P$, where $P$ is the proportion of 1's. Let $(y_1, \ldots, y_n)$ be a random sample; the sample mean $\bar{y}$ is the proportion of 1's in the sample. Specialize the general formulas and show that

$$E(\bar{y}) = P, \quad \text{var}(\bar{y}) = \frac{PQ}{n} \frac{N - n}{N - 1}.$$ 

Check these formulas by specializing the results for the hypergeometric distribution in Problem 5.3.14.

3 (continuation) A random sample of 150 was chosen from a population of 4085 invoices. Detailed examination of each invoice in the sample showed that 32 of them contained errors. Estimate the number of invoices in the population that contain errors; also record a 90 percent confidence interval.

4 A random sample of 30 households in a city having 14,848 households yielded the following numbers of persons per household:

4, 2, 4, 3, 4, 2, 1, 3, 3, 4, 3, 3, 4, 4, 5,

3, 4, 7, 2, 3, 4, 4, 3, 3, 3, 2, 3, 3, 6, 5.

Obtain a 90 percent confidence interval for the total number of people in the city.

5 A random sample of 20 pages from a 530-page text yielded the following number of typographical errors on an intensive examination:

3, 1, 0, 5, 3, 2, 4, 1, 0, 5,

3, 3, 2, 3, 0, 3, 1, 5, 3, 3.

Obtain a 90 percent confidence interval for the total number of typographical errors in the text.

6 Consider the population \( \{1, 2, \ldots, N\} \). For a random sample of $n$, determine expressions for $E(\bar{y})$ and $\text{var}(\bar{y})$ as functions of $N$.

---

**PROBLEMS ON STRATIFIED SAMPLING**

7 Sometimes it is possible to stratify a population into moderately homogeneous strata. Suppose that a population of $N$ elements is stratified into $k$ strata containing $N_1, \ldots, N_k$ elements, respectively. Let $\mu_j$ and $\sigma_j^2$ be the mean and variance in the $j$th stratum and $\mu$ and $\sigma^2$ be the mean and variance in the population. Show that

$$(N - 1)\sigma^2 = \sum (N_j - 1)\sigma_j^2 + \sum N_j(\mu_j - \mu)^2.$$  

8 (continuation) Now suppose that random samples are obtained in each stratum. Let $y_j$ be the mean of a sample of $n_j$ from the $j$th stratum. If $\Sigma_j n_j\bar{y}_j$ is necessarily an unbiased estimate of $\mu$, then show that $a_j = N_j/N$. Let

$$\bar{y}_s = \frac{\sum \frac{N_j \bar{y}_j}{N}}{N}$$

be the unbiased estimate of $\mu$; then show that the variance of $\bar{y}_s$ is

$$\text{var}(\bar{y}_s) = \sum \frac{(N_j/N)^2}{n_j} \frac{\sigma_j^2}{n_j} \left(1 - \frac{n_j}{N_j}\right).$$
9 \text{(continuation)} If the sampling is proportional to stratum size, that is, \( n_j/N_j = n/N \), where \( n = \sum n_j \), and if all the strata have the same variance \( \sigma^2 \), then show that

\[
\text{var} (\overline{y}_j) = \frac{\sigma^2}{n} \left( 1 - \frac{n}{N} \right).
\]

Thus if the variation within strata is substantially smaller than the overall variation, stratified sampling can give improvement in the precision of the estimate.

10 \text{(continuation)} Consider the minimization of \( \text{var} (\overline{y}_j) \) subject to the constraint \( \sum n_j = n \), where \( n \) is held fixed. Use partial differentiation with respect to \( n_j \) (taken to be continuous) and a Lagrange multiplier to accommodate the constraint, and show that minimum variance is attained with \( n_j \) proportional to \( N_j \rho_j \); that is, \( n_j = n N_j \rho_j / \sum N_i \rho_i \). To optimize thus requires information concerning the relative sizes of the strata variances \( \sigma^2_j \).

11 \text{(continuation)} Let \( V_1 \) be the variance of the estimate from a random sample of \( n \); let \( V_2 \) be the variance of the estimate from a proportional sample of \( n \); let \( V_3 \) be the variance from an optimum allocation (Problem 10) of a sample of \( n \).

(a) If the stratum population sizes are large, so that the finite population correction factors can be ignored, show that

\[
V_1 = \frac{\sigma^2}{n}, \quad V_2 = \sum \frac{N_j \sigma^2_j}{N} / n, \quad V_3 = \left( \sum \frac{N_j \sigma^2_j}{N} \right)^2 / n.
\]

(b) Under the same assumption, show that

\[
V_1 = V_2 + \sum \frac{N_j}{N} (\mu_j - \mu)^2 / n
\]

\[
V_2 = V_3 + \sum \frac{N_j}{N} (\sigma_j - \sigma)^2 / n
\]

where \( \sigma = \sum N_j \rho_j / N \).

7.4

INFERENCE FROM SYMMETRY

We now examine some distribution-free or nonparametric methods of statistical inference, an important theoretical area of statistics that directly uses symmetry results from probability theory. Consider a random system with a real-valued response \( y \), for example the biochemical system examined earlier in the chapter. And suppose there is continuity in the system and sufficient figures are used for the recorded response that a distribution without discrete probabilities is appropriate.

Consider an appropriate statistical model. The lack of discrete probabilities can conveniently be expressed in terms of continuity of the distribution function. Accordingly, we take the model to be \( \mathcal{M} = \{ F \} \), where \( \mathcal{M} \) is restricted to continuous distribution functions. The theory we discuss in this section is sometimes called distribution-free—it does not depend on having a particular functional form for the distribution in the model, functional form such as the normal. The theory is also called nonparametric—the model is not given completely in terms of one, two, or several real parameters.

Suppose, as is typical, that the investigator is primarily interested in the general level of the response. For the model \( \mathcal{M} \) the mean \( \mu \) does not exist for some of the
distributions $F$. Accordingly, it is convenient to use the median $\xi$ (or $\mu$) as a measure of the general level of the response. The median of a distribution given by the distribution function $F$ is defined by

$$(1) \quad F(\xi) = 1/2 = 1 - F(\xi)$$

provided the distribution does not have a central interval containing no probability; if there is such an interval, then there is an interval of medians, any one of which can be used for the analysis here.

Suppose that the investigator plans a sequence of $n$ independent repetitions. Let $(y_1, \ldots, y_n)$ designate the compound response for the sequence. The sample space then is $R^n$ and the statistical model can be given as

$$(2) \quad \{F(y_1) \cdots F(y_n) : F \in \mathcal{M}\}.$$  

**EXAMPLE 1**

**Biochemical example:** We shall illustrate several inference procedures using the biochemical data from Section 7.1E. The data, with some simple calculations to be used later, are given in Table 7.2.

<table>
<thead>
<tr>
<th>$y_i$</th>
<th>$y_i - 14.8$</th>
<th>sgn $(y_i - 14.8)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>17.1</td>
<td>2.3</td>
<td>+1</td>
</tr>
<tr>
<td>14.9</td>
<td>0.1</td>
<td>+1</td>
</tr>
<tr>
<td>15.2</td>
<td>0.4</td>
<td>+1</td>
</tr>
<tr>
<td>18.1</td>
<td>3.3</td>
<td>+1</td>
</tr>
<tr>
<td>16.4</td>
<td>1.6</td>
<td>+1</td>
</tr>
<tr>
<td>19.1</td>
<td>4.3</td>
<td>+1</td>
</tr>
<tr>
<td>18.5</td>
<td>3.7</td>
<td>+1</td>
</tr>
<tr>
<td>14.6</td>
<td>-0.2</td>
<td>-1</td>
</tr>
<tr>
<td>17.3</td>
<td>2.5</td>
<td>+1</td>
</tr>
</tbody>
</table>

| $S = 7$ |

---

**A SIGN SYMMETRY**

The sign function sgn is defined by

$$\text{sgn} (y) = +1, 0, -1$$

according as $y > 0$, $= 0$, $< 0$.

Now consider $y$ with a continuous distribution $F$ and median $\xi$. The marginal distribution of $\text{sgn} (y - \xi)$ has probability $1/2$ at $+1$ and probability $1/2$ at $-1$. This is a relocated, rescaled Bernoulli distribution. In fact, $(1/2)(\text{sgn} (y - \xi) + 1)$ is Bernoulli$(1/2)$; note that $(1/2)(\text{sgn} (y - \xi) + 1)$ is the indicator function for a positive value for the deviation $y - \xi$.

Now consider a sample $(y_1, \ldots, y_n)$ from the distribution $F$ with median $\xi$. Then

$$(3) \quad S = \sum_{i=1}^{n} \text{sgn} (y_i - \xi)$$
has a distribution with probability at the points \(-n, -n + 2, \ldots, n - 2, n\). This is a relocated, rescaled binomial\((n, 1/2)\) distribution. In fact,

\[ K = (1/2)(S + n) = \sum_{i=1}^{n} (1/2)(\text{sgn}(y_i - \xi) + 1) \]

is binomial\((n, 1/2)\); note that \(K\) is just the number of positive deviations \(y_1 - \xi, \ldots, y_n - \xi\).

For a sample size \(n = 9\), the distribution of \(S\) and \(K\) is as follows:

<table>
<thead>
<tr>
<th>(S)</th>
<th>-9</th>
<th>-7</th>
<th>-5</th>
<th>(\ldots)</th>
<th>5</th>
<th>7</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>(K)</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>(\ldots)</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>(p)</td>
<td>(1/512)</td>
<td>(9/512)</td>
<td>(36/512)</td>
<td>(\ldots)</td>
<td>(36/512)</td>
<td>(9/512)</td>
<td>(1/512)</td>
</tr>
</tbody>
</table>

This is binomial\((9, 1/2)\) for \(K\).

For large samples the distribution of \(S\) and \(K\) can be approximated by treating

\[ \frac{K - n/2}{\sqrt{n/2}} = \frac{S}{\sqrt{n}} \]

as a standard normal variable; see Example 6.2.3.

**B SIGN TEST**

Consider the random system discussed in the introduction to this section and suppose that a value \(\xi_0\) has been suggested for the median of the distribution. We consider a simple test for the hypothesis \(\xi = \xi_0\).

The function

\[ S = \sum_{i=1}^{n} \text{sgn}(y_i - \xi_0) \]

(6)

gives the sum of the signs of the deviations from \(\xi_0\), and the adjusted function

\[ K = (1/2)(S + n) = \sum_{i=1}^{n} (1/2)(\text{sgn}(y_i - \xi_0) + 1) \]

(7)

gives the number of positive deviations from \(\xi_0\). This function indicates whether sample values \(y_1, \ldots, y_n\) are distributed in a reasonable way about the supposedly central value \(\xi_0\).

If \(\xi = \xi_0\), then \(K\) is binomial\((n, 1/2)\). On the other hand, if the true median \(\xi\) is different from \(\xi_0\), then \(K\) is binomial\((n, 1 - F(\xi_0))\) and we would expect extreme values for \(K\): a higher probability for large values of \(K\) if \(F(\xi_0)\) is small, and higher probability for small values of \(K\) if \(F(\xi_0)\) is large.

Consider the following test for the hypothesis \(\xi = \xi_0\). We calculate the value of \(K\) in formula (7) and compare it with the binomial\((n, 1/2)\) distribution. A central value for \(K\) is in conformity with the hypothesis. An extreme value on one or other tail of the distribution is an indication against the hypothesis.
EXAMPLE 1 (continuation)

Consider the biochemical data and suppose that we want to test the value $\xi = 14.8$ as suggested by some theoretical calculations. The observed value of $S$ is 7 and the corresponding value of $K$ is 8. Note that this is the next-to-last point on the right tail of the distribution as tabulated at the end in Section A. We can, in part, assess the significance of this observed value by calculating the observed level of significance, the probability under the hypothesis of a value as far or farther from the center of the distribution as that observed:

$$\frac{1}{512} + \frac{9}{512} + \frac{9}{512} + \frac{1}{512} = \frac{20}{512} \approx 4\%.$$

This is a moderately small probability, and it gives a moderate indication against the hypothesis.

C SIGN CONFIDENCE INTERVAL

Now suppose that we want a confidence interval for the general level $\xi$. The distribution of

$$K = \sum_{i=1}^{n} \frac{1}{2} \left( \text{sgn}(\gamma_i - \xi) + 1 \right)$$

is binomial($n$, 1/2). From this binomial distribution we can determine a central 95 percent interval $[k_1, k_2]$ for $K$, central in the sense of approximately equal probability on each tail. We can then write

$$(8) \quad P(k_1 \leq K \leq k_2) \approx 95\%$$

$$= P(k_1 \leq \text{number of } \gamma_i > \xi \leq k_2)$$

$$= P(\gamma_{(n-k_2)} \leq \xi < \gamma_{(n-k_1+1)}),$$

where $(\gamma_{(1)}, \ldots, \gamma_{(n)})$ is the order statistic that records the values $\gamma_1, \ldots, \gamma_n$ arranged in order from the smallest to the largest. Thus $[\gamma_{(n-k_2)}, \gamma_{(n-k_1+1)}]$ is a 95 percent confidence interval for the median $\xi$. If $n$ is large, we can use the normal approximation giving $[k_1, k_2] \approx (n/2 \pm z_{2.5\%} \sqrt{n}/2)$, where $z_\gamma$ is the value exceeded with probability $\gamma$ according to the standard normal.

EXAMPLE 1 (continuation)

Consider the biochemical data. The distribution for $K$ in Section A gives

$$(9) \quad P(2 \leq K \leq 7) = 0.961$$

$$= P(\gamma_{(3)} \leq \xi < \gamma_{(8)}).$$

Thus $[\gamma_{(2)}, \gamma_{(8)}] = [14.9, 18.5]$ is a 96.1 percent confidence interval for the median $\xi$. 
D MODEL WITH ADDITIONAL SYMMETRIES

For the random system discussed in the introduction, suppose there is the additional information that the response distribution is symmetrical about its median. As an appropriate statistical model based on continuity and the symmetry just mentioned, consider $\Omega = \{ F \}$, where the elements $F$ are restricted to continuous distribution functions with the symmetry property

$$F(\xi - d) = 1 - F(\xi + d), \quad d \geq 0,$$

where $\xi$ is the median of $F$. The model then for a sequence of repetitions, $(y_1, \ldots, y_n)$, is

$$\{ F(y_1) \cdots F(y_n) : F \in \Omega \}.$$

EXAMPLE 1 (continuation)

We shall illustrate some additional inference procedures using the biochemical data. The data with some calculations of interest later are given in Table 7.3.

| $y_i$ | $|y_i - 14.8|$ | $\text{sgn} (y_i - 14.8)$ | $|y_i - 14.8|$ $\text{sgn} (y_i - 14.8)$ |
|-------|--------------|-----------------|------------------|
| 17.1  | 2.3          | +1              | 2.3              |
| 14.9  | 0.1          | +1              | 0.1              |
| 15.2  | 0.4          | +1              | 0.4              |
| 18.1  | 3.3          | +1              | 3.3              |
| 16.4  | 1.6          | +1              | 1.6              |
| 19.1  | 4.3          | +1              | 4.3              |
| 18.5  | 3.7          | +1              | 3.7              |
| 14.6  | 0.2          | -1              | -0.2             |
| 17.3  | 2.5          | +1              | 2.5              |
| 18.4  |              |                 | 18.0             |

E CONDITIONAL SIGN SYMMETRY

Now consider $y$ with a continuous distribution $F$ that is symmetric about the median. The symmetry suggests that the conditional distribution of $y - \xi$ given $|y - \xi|$ has probability 1/2 at $+ |y - \xi|$ and probability 1/2 at $- |y - \xi|$; see Figure 7.4 expressed in terms of a density function. For a formal statement of the distribution theory it is convenient to relocate at the median and in effect consider the case $\xi = 0$.

LEMMA 1

*If $y$ has a continuous distribution that is symmetric about the origin, then $|y|$ and $\text{sgn} (y)$ are statistically independent and $\text{sgn} y$ has probability 1/2 at $-1$ and at $+1$. 

Sec. 7.4: Inference from Symmetry

FIGURE 7.4
Conditional distribution of \( y - \xi \) given \(|y - \xi|\) has probability 1/2 at each of \(+|y - \xi|, -|y - \xi|\).

Proof: We show independence by showing that the distribution function for \((|y|, \text{sgn}(y))\) factors, so the variables separate; see Proposition 3.3.6. The function \(\text{sgn}(y)\) takes essentially two values, \(-1\) and \(+1\); accordingly, it suffices to consider the distribution function for \((\text{sgn}(y), |y|)\) with \(\text{sgn}(y)\) at an intermediate point, say \(0\):

\[
P(\text{sgn}(y) \leq 0, |y| \leq u) = P(-u \leq y \leq 0) = \frac{1}{2} - F(-u),
\]

\[
P(\text{sgn}(y) \leq 0) = P(y \leq 0) = \frac{1}{2},
\]

\[
P(|y| \leq u) = P(-u \leq y \leq u) = F(u) - F(-u).
\]

And then, by symmetry, we have

\[
(1/2) \cdot (F(u) - F(-u)) = (1/2) \cdot 2\left((1/2) - F(-u)\right) = (1/2) - F(-u).
\]

Thus the product of the marginal distribution functions is equal to the joint distribution function, and it follows that \(|y|\) and \(\text{sgn}(y)\) are statistically independent.

Now, in general, consider \(y\) with symmetric distribution function \(F\) and median \(\xi\). Then by the lemma we have that \(|y - \xi|\) and \(\text{sgn}(y - \xi)\) are statistically independent and \(\text{sgn}(y - \xi)\) has probability 1/2 at \(-1\) and at \(+1\). Or, equivalently, we have that the conditional distribution of \(\text{sgn}(y - \xi)\) given \(|y - \xi|\) has probability 1/2 at \(-1\) and at \(+1\).

Now consider a sample \(\{y_1, \ldots, y_n\}\) from the distribution \(F\) with median \(\xi\). We investigate the conditional distribution of

\[
(10) \quad T = \sum_{i=1}^{n} |y_i - \xi| \text{sgn}(y_i - \xi) = \sum_{i=1}^{n} (y_i - \xi)
\]

given the absolute deviations \(|y_1 - \xi|, \ldots, |y_n - \xi|\). This distribution can be expressed by

\[
(11) \quad \sum_{i=1}^{n} |y_i - \xi| s_i,
\]

where \(\{s_1, \ldots, s_n\}\) is a sample from a distribution with probability 1/2 at \(-1\) and at \(+1\). For small and moderate samples this distribution can be calculated by hand or
machine. And for large samples there is approximate normality provided the set of
absolute deviations forms a reasonable array.

As an illustration consider a sample of size \( n = 9 \) and suppose the absolute
deviations from the true median are

\[
0.1, \ 0.2, \ 0.4, \ 1.6, \ 2.3, \ 2.5, \ 3.3, \ 3.7, \ 4.3.
\]

The conditional distribution of \( T \) can be obtained by enumeration, easily done for the
tails, but tedious for the center of the distribution.

<table>
<thead>
<tr>
<th>( T )</th>
<th>-18.4</th>
<th>-18.2</th>
<th>-18.0</th>
<th>\ldots</th>
<th>18.0</th>
<th>18.2</th>
<th>18.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P )</td>
<td>1/512</td>
<td>1/512</td>
<td>1/512</td>
<td>1/512</td>
<td>1/512</td>
<td>1/512</td>
<td></td>
</tr>
</tbody>
</table>

The distribution is symmetric about 0. Some of the probabilities differ from 1/512, for example

\[ P(T = 13.8) = 2/512, \]

as can be seen from

\[
18.4 - 2(2.3) = 18.4 - 2(0.1 + 0.2 + 0.4 + 1.6) = 13.8;
\]

thus the distribution is not flat or symmetric in the sense of Section 1.3E.

\[ F \] SIGNED-DEVIATION OR FISHER TEST

Consider the random system discussed in Section D and suppose that \( \xi_0 \) has been
suggested for the median of the distribution. We now consider the Fisher test for the
hypothesis \( \xi = \xi_0 \).

The function

\[
(12) \quad T = \sum_{i=1}^{n} |y_i - \xi_0| \text{sgn}(y_i - \xi_0) = \sum_{i=1}^{n} (y_i - \xi_0)
\]
gives the sum of deviations from \( \xi_0 \). Given the absolute deviations \(|y_1 - \xi_0|, \ldots, |y_n - \xi_0|\), the value of the function \( T \) provides an indication of how sample values \( y_1, \ldots, y_n \)
are distributed in relation to the supposedly theoretically central value \( \xi_0 \): the
signs of the deviations from \( \xi_0 \) are weighted by the distances from \( \xi_0 \).

If \( \xi = \xi_0 \), then the distribution of \( T \) given \(|y_1 - \xi_0|, \ldots, |y_n - \xi_0|\) is distributed as

\[
(13) \quad \sum_{i=1}^{n} |y_i - \xi_0| s_i,
\]

where \((s_1, \ldots, s_n)\) is a sample from a distribution with probability 1/2 at -1 and
at +1.

We can test the hypothesis \( \xi = \xi_0 \) by calculating the value of \( T \) in formula (12) and
comparing it with the distribution expressed by (13). An extreme value on one or other
tail of the distribution would be an indication against the hypothesis.

EXAMPLE 1 (continuation)

Consider the biochemical example as given in Section D and suppose that we want
to test the hypothesis \( \xi = 14.8 \). The observed value of \( T \) is 18.0. The distribution
for $T$ under the hypothesis $\xi = 14.8$ is recorded in Section E. Note that the observed value 18.0 is the third point in on the right tail of the distribution. The observed level of significance, the probability of a value as far or farther from the center of the distribution as that observed, is

$$(1/512) + (1/512) + (1/512) + (1/512) + (1/512) + (1/512) = 6/512,$$

which is just slightly larger than 1 percent. This is not the kind of value that we would reasonably expect under the hypothesis, and in turn it provides moderate evidence against the hypothesis.

In conclusion, it is worth noting some equivalent functions that can be used for this conditional test. We can write

$$T = \sum_{y_i > \xi_o} (y_i - \xi_o) + \sum_{y_i \leq \xi_o} (y_i - \xi_o)$$

$$= 2 \sum_{y_i > \xi_o} |y_i - \xi_o| - \sum_{y_i \leq \xi_o} |y_i - \xi_o|$$

$$= 2 T^+ - k,$$

where $k = \Sigma |y_i - \xi_o|$ is a constant; thus $T$ and $T^+$ are monotone functions of each other (just as $S'$ and $K$ were in Section B). In another direction we can write

$$\frac{T^2/n}{\sum |y_i - \xi_o|^2} = \frac{n(\bar{y} - \xi_o)^2}{\sum (y_i - \xi_o)^2} = \frac{n(\bar{y} - \xi_o)^2}{n(\bar{y} - \xi_o)^2 + \sum (y_i - \bar{y})^2} = \frac{1}{1 + (n - 1)/t^2},$$

where

$$t = \frac{\bar{y} - \xi_o}{s_y/\sqrt{n}}$$

is the $t$-function given by (7.2.17) in Section 7.2E. Thus $|T|$ can be put in one-to-one correspondence with $|t|$, and, in addition, the signs of $T$ and $t$ correspond; thus we could equivalently use $t$ for the test discussed in this section.

**G SIGNED-DEVIATION CONFIDENCE INTERVAL**

Now suppose that we want a confidence interval for the general level $\xi$. The function

$$T = \sum_{1}^{n} |y_i - \xi| \text{sgn} (y_i - \xi) = \sum_{1}^{n} (y_i - \xi)$$
can be used together with the conditional distribution described by
\[
\sum_{i=1}^{n} |y_i - \zeta| s_i,
\]
where \((s_1, \ldots, s_n)\) has probability \(1/2^n\) at each of the \(2^n\) points formed by \(s_i = \pm 1\). Consider the following equivalent statements:
\[
\sum_{i=1}^{n} (y_i - \zeta) \leq m\text{th smallest of the } \sum_{i=1}^{n} (y_i - \zeta)s_i,
\]
\[
\sum_{i=1}^{n} (\zeta - y_i) \geq m\text{th largest of the } \sum_{i=1}^{n} (\zeta - y_i)s_i,
\]
\[
0 \geq m\text{th largest of the } \sum_{i=1}^{n} (s_i - 1)(\zeta - y_i),
\]
\[
0 \geq m\text{th largest of the } 2 \sum_{s_i = -1} (y_i - \zeta),
\]
\[
0 \geq m\text{th largest of the } \left( \sum_{s_i = -1} y_i \right) / \left( \sum_{s_i = -1} 1 \right) - \zeta,
\]
\[
\zeta \geq m\text{th largest of the } \left( \sum_{s_i = -1} y_i \right) / \left( \sum_{s_i = -1} 1 \right),
\]
where, for example, \(\sum_{s_i = -1} 1\) gives the number of coordinates of \((s_1, \ldots, s_n)\) that are equal to \(-1\). Thus to form a confidence interval, we consider the \(2^n - 1\) possible subsets of \(\{y_1, \ldots, y_n\}\); we form the \(2^n - 1\) subset averages
\[
\left( \sum_{s_i = -1} y_i \right) / \left( \sum_{s_i = -1} 1 \right) = \sum_{s_i = -1} y_i / \text{(number of } s_i = -1);\]
we take the \(m_{\text{st}}\) smallest and \(m_{\text{nd}}\) largest of these subset averages; and we obtain a confidence interval for \(\zeta\) having confidence level \(1 - (m_1 + m_2)/2^n\).

**EXAMPLE 1 (continued)**

Consider further the biochemical example as presented in Section D. The arithmetic is reasonable if we choose a 98 percent confidence interval: we exclude 1 percent on each tail; 1 percent of 512 is approximately 5; thus we form an interval from the 5th smallest to the 5th largest of the 511 subset averages. The data rearranged in accord with the order statistic are given in Table 7.4. For the subset averages on the left tail it is convenient to use 14.6 as a temporary origin:

<table>
<thead>
<tr>
<th>Subset</th>
<th>({0})</th>
<th>({0.0.3})</th>
<th>({0.3})</th>
<th>({0.0.6})</th>
<th>({0.0.3.0.6})</th>
<th>(\cdots)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>0</td>
<td>0.15</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td></td>
</tr>
</tbody>
</table>
TABLE 7.4
Data for example 1

<table>
<thead>
<tr>
<th>( \nu_i )</th>
<th>( \nu_i - 14.6 )</th>
<th>19.1 - ( \nu_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.6</td>
<td>0</td>
<td>4.5</td>
</tr>
<tr>
<td>14.9</td>
<td>0.3</td>
<td>4.2</td>
</tr>
<tr>
<td>15.2</td>
<td>0.6</td>
<td>3.9</td>
</tr>
<tr>
<td>16.4</td>
<td>1.8</td>
<td>2.7</td>
</tr>
<tr>
<td>17.1</td>
<td>2.5</td>
<td>2.0</td>
</tr>
<tr>
<td>17.3</td>
<td>2.7</td>
<td>1.8</td>
</tr>
<tr>
<td>18.1</td>
<td>3.5</td>
<td>1.0</td>
</tr>
<tr>
<td>18.5</td>
<td>3.9</td>
<td>0.6</td>
</tr>
<tr>
<td>19.1</td>
<td>4.5</td>
<td>0</td>
</tr>
</tbody>
</table>

For the subset averages on the right tail it is convenient to measure back from 19.1:

<table>
<thead>
<tr>
<th>Subset</th>
<th>( {0} )</th>
<th>( {0, 0.6} )</th>
<th>( {0, 1.0} )</th>
<th>( {0, 0.6, 1.0} )</th>
<th>( {0.6} )</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>0</td>
<td>0.3</td>
<td>0.5</td>
<td>0.533</td>
<td>0.6</td>
<td></td>
</tr>
</tbody>
</table>

Thus an approximate 98 percent confidence interval for \( \xi \) is

\( (14.6 + 0.3, 19.1 - 0.6) = (14.9, 18.5) \).

Note that this happens to be the same interval as obtained in Section C, yet the confidence levels are different. The confidence level in each case, however, is determined in relation to intervals that "might have been obtained."

H EXERCISES

1. The resistances of five pieces of wire sampled from production are recorded in Exercise 7.1.1; assume continuity of the distribution function. Test the hypothesis that the median \( \xi = 0.135 \). Form a 93.75 percent confidence interval for \( \xi \).

2. The breaking loads for six samples of plastic are recorded in Exercise 7.1.2; assume continuity of the distribution function. Test the hypothesis that the median breaking strength is 27.0. Form a 96.875 percent confidence interval for the median \( \xi \).

3. For Exercise 1, assume in addition that the distribution is symmetric; answer Exercise 1 using the signed-deviation function.

4. For Exercise 2, assume in addition that the distribution is symmetric; answer Exercise 2 using the signed-deviation function.

5. A sample of 11 rats was obtained randomly from a population of rats. A blood viscosity test of each yielded 3.29, 3.91, 4.64, 3.65, 3.67, 4.18, 3.74, 4.67, 3.03, 4.61, 3.84. (a) Use the sign test to test the hypothesis: median \( \xi = 4.6 \). (b) Form a 98.8 percent confidence interval for the median \( \xi \).

6. (continuation) Suppose there are grounds for the assumption that the distribution is symmetrical about the median \( \xi \). (a) Use the signed-deviation test to test the hypothesis: median \( \xi = 4.6 \). (b) Form a 99 percent confidence interval for the median \( \xi \).
7. **On the sign test:** The sign test is based on \( K = \sum^n c(y_i - \xi) \), where \( c \) is the indicator function for the positive axis \((0, \infty)\). Some statistical theory suggests that decisions be made in a statistical investigation. From this point of view the conventional 5 percent (or \( \alpha \)) sign test is to "reject" if \( |K - n/2| \geq k \) and to "accept" otherwise; the constant \( k \) is chosen so that \( P(|K - n/2| \geq k) \leq 5 \text{ percent (or } \leq \alpha \text{) if } \xi = \xi_0 \). The investigator then knows that the test will reject a true hypothesis with probability at most 5 percent. But what does the test do if \( \xi \neq \xi_0 \)? Let \( p \) designate the probability \( P(y > \xi) \) for such a case. The investigator would then be interested in the **power** of the test: the probability \( \beta(p) = P(|K - n/2| \geq k; p) \) of rejecting the hypothesis when the distribution has probability \( p \) to the right of \( \xi_0 \). For \( p \neq 1/2 \), the investigator would hope for high power, that is, high probability of rejecting the hypothesis. For a sample of \( n = 7 \), determine \( k \) for a 2 percent test. Calculate the power function \( \beta \) and plot against \( p \).

8. Let \( r_1, \ldots, r_n \) be the ranks of \( y_1, \ldots, y_n \); recall Exercise 3.2.15. Show that
\[
    r_i = \sum^n_{j=1} c(y_i - y_j) + 1,
\]
where \( c \) is the indicator function for the positive axis \((0, \infty)\). Assume that the \( y_i \)'s are all different.

9. **Wilcoxon signed-rank test:** Consider the absolute deviations \(|y_1 - \xi|, \ldots, |y_n - \xi|\) and let \( r(|y_1 - \xi|), \ldots, r(|y_n - \xi|)\) be the corresponding ranks; thus \( r(|y_i - \xi|) \) is equal to \( r \) if \(|y_i - \xi|\) is the \( r \)th smallest of the absolute deviations. Under the assumptions in Section D we consider the conditional distribution of the function
\[
    T^+ = \sum_{|y_i - \xi| \geq \xi} r(|y_i - \xi|) = \sum_{|y_i - \xi| \geq \xi} r(|y_i - \xi|)c(y_i - \xi)
\]
given the absolute deviations. The function \( c \) is defined in Problem 8.

(a) Tabulate the conditional distribution of \( T^+ \) for the case \( n = 3 \).

(b) Show that
\[
    E(T^+) = \frac{n(n+1)}{4}, \quad \text{var}(T^+) = \frac{n(n+1)(2n+1)}{24}.
\]

10. **(continuation)** For the example in Exercises 5 and 6, use the preceding signed-rank test to test the hypothesis \( \xi = 4.6 \) by:

(a) Calculating the necessary part of the distribution of \( T^+ \) directly; note that the distribution of \( T^+ \) is symmetrical.

(b) Using the limiting normal distribution: Theorem 6.2.4 shows that \( T^+ \), standardized with respect to its mean and standard deviation, has a limiting standard normal distribution.

11. **(continuation)** The continuity assumptions in Section D assure that there is zero probability for a tie in a sample \((y_1, \ldots, y_n)\) of size \( n \). Thus assume that the ranks in Problem 8 are well defined.

(a) Use Problem 8 to show that the signed-rank test function can be reexpressed as
\[
    T^+ = \sum^n_{i=1} c(y_i - \xi) + \sum^n_{i,j=1, i \neq j} c(y_i - \xi)c(|y_i - \xi| - |y_j - \xi|).
\]

(b) Let \( S = \{(y_i + y_j)/2 : i \leq j = 1, \ldots, n\} \) be the set of \( n(n+1)/2 \) subset averages for subsets of sizes 1 and 2. Show that \( T^+ \) is the number of points in \( S \) that are larger than \( \xi \).

12. **(continuation)** The **signed-rank confidence interval:** A value of \( T^+ \) can be associated with having \( \xi \) in one of the \( n(n+1)/2 + 1 \) intervals formed by the \( n(n+1)/2 \) points of \( S \). Let \( P(t_1 \leq T^+ \leq t_2) = 1 - \alpha \). Then show that a \( 1 - \alpha \) confidence interval is obtained by combining the intervals for which \( T^+ \) lies in the range \( t_1 \leq T^+ \leq t_2 \).

13. **(continuation)** For the example in Exercises 5 and 6, use the signed-rank test to form a central 99 percent confidence interval for \( \xi \).
J TWO-SAMPLE PROBLEMS

14 The median-test or Mood–Westenberg test: Let \((y_{11}, \ldots, y_{1m})\) be a sample from a first distribution and \((y_{21}, \ldots, y_{2n})\) be a sample from a second distribution. Consider the case where the distributions have the same form; and suppose that we are interested in the difference in the locations of the two distributions. Specifically, let the first distribution be given by \(F(y_1)\) and the second distribution be given by \(F(y_2 - \theta)\), where \(F\) can be any continuous distribution function; and suppose that we are interested in testing the hypothesis \(\theta = 0\). The median test is a partial analog of the sign test. It is performed in the following manner: The two samples are combined to form a larger "sample" of \(m + n\) values; the \(m + n\) values are arranged in order of magnitude from the smallest to the largest; the sequence is divided into two equal or nearly equal parts so that \([(m + n)/2]\) values are larger than the dividing point and \(m + n - [(m + n)/2]\) values are smaller; each value in the sequence is replaced by an \(A\) if it originated from the first sample and by a \(B\) if it originated from the second sample; let \(S\) be the number of \(A\)'s to the right of the dividing point. In other words, \(S\) is the number of first-sample values that are greater than the median of the combined samples. Under the hypothesis, show that the probability function of \(S\) is

\[ p_S(S) = \binom{m}{S} \binom{n}{n - S} / \binom{N}{n} \]

where \(N = m + n\) and \(h = \lceil N/2 \rceil\).

15 (continuation) Under the hypothesis \(\theta = 0\), obtain an expression for \(E(S^0)\) and show that

\[ E(S) = \frac{mh}{N}, \quad \text{var}(S) = \frac{mn(N - h)}{N^2(N - 1)}. \]

If the two distributions have the same continuous distribution function, it can be shown (Problem 24) that \(z = (S - E(S))/\sqrt{\text{var}(S)}\) has a limiting standard normal distribution.

16 (continuation) For the data in Exercise 7.2.5 and Problem 7.2.7, test the hypothesis \(\theta = 0\).

17 (continuation) The median-test confidence interval: Let \(s_1\) and \(s_2\) be chosen so that \(P(s_1 \leq S \leq s_2) = 95\%\) (or \(1 - \alpha\)) using the hypothesis distribution of \(S\). A 95 percent (or \(1 - \alpha\)) confidence "interval" consists of all those \(\theta\) values for which the median test applied to \((y_{11}, \ldots, y_{1m}), (y_{21} - \theta, \ldots, y_{2n} - \theta)\) yields a value of \(S\) in the interval \([s_1, s_2]\). Let \((y_{1(1)}, \ldots, y_{1(m)})\) be the order statistic for the first sample and \((y_{2(1)}, \ldots, y_{2(n)})\) the order statistic for the second sample.

(a) Show that the value \(S\) implies \(y_{2(m-s_2+1)} > y_{2(m-s_2)+1} - \theta\) and \(y_{2(m-s_2)+1} - \theta > y_{1(m-s_1)}\).

(b) Obtain the corresponding interval for \(\theta\).

(c) Show that the 95 percent (or \(1 - \alpha\)) confidence interval is \((y_{2(m-h_s+1)} - y_{1(m-h_s+1)})\), \(y_{2(m-h_s+1)} - y_{1(m-h_s+1)}\).

18 The Mann–Whitney test: Assume the conditions given in Problem 14. The Mann–Whitney test is also a partial analog of the sign test. Let \(U\) be the number of pairs \((y_{1i}, y_{2j})\) having \(y_{2j} > y_{1i}:

\[ U = \sum_{i=1}^{m} \sum_{j=1}^{n} c(y_{2j} - y_{1i}). \]

Under the hypothesis \(\theta = 0\), show that

\[ E(U) = \frac{mn}{2}, \quad \text{var}(U) = \frac{mn(N + 1)}{12}. \]

where \(N = m + n\). A generalized central limit theorem for what are called \(U\) statistics can be used to show that \(U\) standardized with respect to its mean and standard deviation has a limiting standard normal distribution; see, for example, Fraser (1957, p. 225).
For the data in Exercise 7.2.5 and Problem 7.2.7, test the hypothesis \( \theta = 0 \).

The Mann–Whitney confidence interval: Suppose that \( P(u_1 \leq U \leq u_2) = 95\% \) percent (or \( 1 - \alpha \)) under the hypothesis distribution. The confidence interval is obtained by calculating \( U \) for \( (y_{11}, \ldots, y_{1n}) \) and \( (y_{21} = \theta, \ldots, y_{2n} = \theta) \). Show that the 95 percent (or \( 1 - \alpha \)) confidence interval is \((d_{\overline{y}_11}, d_{\overline{y}_n})\), where \( d_{\overline{y}_m} \) are the \( mn \) differences \( y_{2j} - y_{1j} \) arranged in order of magnitude from smallest to largest.

The Wilcoxon test: The Wilcoxon test is performed in the following manner: the two samples are combined to form a larger "sample" of \( m + n \) values; the \( m + n \) values are arranged in order of magnitude from the smallest to the largest; the values in the sequence are replaced by their ranks from 1 to \( m + n \); the sum \( W \) of the ranks corresponding to those values that come from the second sample is calculated. Use Problem 8 to show that

\[
W = U + \frac{n(m + 1)}{2}.
\]

Thus the Wilcoxon test is equivalent to the Mann–Whitney test. The Mann–Whitney test was proposed several years after the Wilcoxon test and has some advantages for computation.

The Pitman test: The Siegel–Tukey test is a two-sample analog of the Fisher test; assume the conditions given in Problem 14: a sample of \( m \) from a distribution \( F(y) \) and a sample of \( n \) from a distribution \( F(y_2 - \theta) \), where \( F \) can be any continuous distribution function. Having the same distribution form is the property that replaces the symmetry used in the Fisher test. The test is a conditional test given the \( m + n \) values in the combined sample; the conditional distribution under the hypothesis gives equal probability \( \frac{1}{\binom{m+n}{n}} \) to each assignment of \( m \) values to the first sample set and the remaining \( n \) to the second sample set. The conditional test of the hypothesis: \( \theta = 0 \) given the set \( \{y_{11}, \ldots, y_{1m}, y_{21}, \ldots, y_{2n}\} \) can be based on \( \overline{y}_2 \), or on \( d = \overline{y}_2 - \overline{y}_1 \), or on

\[
t = \left(\frac{1}{m} + \frac{1}{n}\right)^{1/2} \left(\frac{\sum (y_{2j} - \overline{y}_2)^2 + \sum (y_{1j} - \overline{y}_1)^2}{m + n - 2}\right)^{1/2}
\]

as defined in Section 7.2.1. Show that the three functions are equivalent in the order they place on the \( \binom{m+n}{n} \) segregations into possible first and second samples. The test is performed by calculating the observed value and comparing it with the conditional distribution based on the \( \binom{m+n}{n} \) different segregations of the combined sample. For small samples the test can be performed easily using the function \( \sum \gamma_{2j} \). It involves calculating the possible values of \( \sum \gamma_{2j} \) under the various allocations of \( n \) values as possible second samples; often these can be calculated for just the extreme cases necessary for determining the level of significance.

The following samples were obtained with two treatments applied to experimental units: (27.1, 25.3, 28.4); (28.3, 29.2, 28.5). Test the hypothesis \( \theta = 0 \).

Proof of the limiting normality: Let \( m = \alpha N \), \( n = \beta N \) and let

\[
f_n(x) = p_n(x) + z\sigma(x)\sigma_n.
\]

By substituting the expressions for the mean and standard deviation and using Stirling's formula for the factorials, show that \( \lim_{\alpha \to \infty} f_n(x) = \exp\left(-x^2/2\right)\sqrt{2\pi} \). Then use Proposition 6.1.2 to prove the limiting normality; note the method in Problems 6.1.1 and 6.1.12.

SUPPLEMENTARY MATERIAL

Supplement Section 7.5 is on pages 563–572.

The inference methods in this chapter assume that there is randomness. The supplementary material is concerned with tests of randomness. A very powerful industrial tool, quality-control
charts, is examined in some detail. And a simple run test for randomness is presented, partly for its quick convenient scientific uses and partly for the way it illustrates the application of some combinatorial formulas.

NOTES AND REFERENCES

The following are textbooks that develop introductory statistics:


This text introduces statistics in an open-ended frame presenting tests of significance, estimation, and confidence methods as integrated components of statistical inference. These methods are introduced directly from large-sample and symmetry results of probability theory. Thus the methods are not conceptualized in the context of the common normal distribution model. Rather, inference is seen as a process from data to conclusions, a process that depends on the particular model accepted for an application.

In later chapters this text progressively examines more specialized statistical models.

For some additional tests of randomness (Supplement Section 7.5), see Knuth (1969).


In this chapter we have introduced the basic concepts of statistical inference. Many of the fruitful ideas of contemporary inference are due to R. A. Fisher; the following three fundamental books of his are recorded for supplemental reading.

In Chapter 7 we introduced estimation, tests of significance, and confidence intervals, and showed how these closely related components of inference can be applied easily to data. The three methods of inference were examined in a context of applications—unlinked to particular models that would specify precise details of distribution form. The background theory used was the symmetry and large-sample distribution theory from the probability chapters.

In Chapter 8 we shall examine statistical models that do specify precise details of distribution form. In Section 8.1 we show that such a model together with data leads to an observed likelihood function, a function that makes inferences immediately available concerning an unknown parameter value. For illustrations we then examine models that use the normal distribution: the distribution theory in Section 8.2, and the inference methods in Section 8.3. The process of going from a sample space to a likelihood function is investigated in Section 8.4. Certain easy results are then found to be restricted to the exponential models in Section 8.5 or to the large-sample models in Section 8.6.
THE LIKELIHOOD FUNCTION IN STATISTICAL INFERENCE

8.1

THE OBSERVED LIKELIHOOD FUNCTION

We now investigate statistical inference based on a parametric statistical model. Such a model is a more structured model in which the form of the distribution is known except for one, two, or several real parameters. As possible applications, consider the following. The lifetime of aluminum crucibles is approximately normally distributed with unknown location and scale parameters $\mu$ and $\sigma$. The determination of potassium content in a solution is approximately Student($\lambda$) with unknown location and scale parameters $\mu$ and $\sigma$ and unknown parameter $\lambda$ determining the shape of the distribution. The lifetime of an electronic component is approximately exponential($\theta$) with unknown scale parameter $\theta$.

In this chapter we investigate what a parametric statistical model says about observed data. This leads us to the likelihood function in Section 8.1 and then to a detailed examination of inference for normal models in Sections 8.2 and 8.3. A theoretically important statistic based on likelihood analysis is examined in Section 8.4; this leads us to the exponential models in Section 8.5 and large-sample likelihood in Section 8.6.

A THE PARAMETRIC STATISTICAL MODEL

The term parametric statistical model is used for models in which the form of the distribution is known except for one, two, or several real parameters. For example, a determination of potassium content is known to be approximately of Student($\lambda$) form with location $\mu$ and scale $\sigma$; for a single determination, the density function is

\[
f(y|\mu, \sigma, \lambda) = \frac{1}{\sigma} \frac{\Gamma\left(\frac{\lambda + 1}{2}\right)}{\Gamma(1/2)\Gamma(\lambda/2)} \left(1 + \frac{(y - \mu)^2}{\sigma^2\lambda}\right)^{-(\lambda+1)/2} \frac{1}{\sqrt{\lambda}}.
\]

The full parameter is $\theta = (\mu, \sigma, \lambda)$ with values in the space $\Omega = \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}^+$. One of the values of the parameter $\theta$ is the true value, and to a reasonable approximation it gives the distribution for a potassium determination. Note that $\theta$ for this example has dimension 3. Also note that the shape parameter $\lambda$ determines the thickness of the tails of the distribution.
Now in general let \( f(y|\theta) \) be the density function for the response \( y \) on a sample space \( S \). The parameter \( \theta \) with values in a parameter space \( \Omega \) indexes the various possible density functions for the response. Accordingly, the statistical model is

\[
\mathcal{M} = \{ f(\cdot|\theta): \theta \in \Omega \},
\]

as expressed in terms of density functions.

The sample space \( S \) will usually be an open set on \( R^1, R^2, \) or \( R^k \), and the function \( f \) will be density with respect to length, area, or volume measure. Alternatively, we see from Problem 5.8.15 that the binomial and the Poisson can be presented in terms of density functions with respect to the cardinality measure (Example 1.3.2) on the integers. Accordingly, we use the notation \( f(y|\theta) \) both for a density function in the absolutely continuous case and for the probability function in the discrete case.

The various possible distributions for the response are obtained as the parameter \( \theta \) ranges over the parameter space \( \Omega \). If the model is a correct model for an application, then some value of \( \theta \) in \( \Omega \) is the true value that gives the appropriate distribution for the response.

In most applications we will need the model for a sequence of independent repetitions of the system. Let \( y = (y_1, \ldots, y_n)^T \) designate the vector response. Then the density function for the compound \( y \) is

\[
f(y|\theta) = f(y_1|\theta) \cdots f(y_n|\theta).
\]

Note that it is convenient to use the letter \( f \) to designate both the single and the compound density function; the size of the response argument provides the distinction.

B THE LIKELIHOOD FUNCTION FROM OBSERVED DATA

Consider a response \( y \) that is known to be approximately normal(\( \theta, 1 \)). For a sample of \( n \) independent repetitions, the model is

\[
\frac{1}{(2\pi)^{n/2}} \exp \left\{ -(1/2) \sum_{i=1}^{n} (y_i - \theta)^2 \right\}: \theta \in R.
\]

As an example, consider the lifetime in months of crucibles used in the electrolytic production of aluminum. For this application we should perhaps formally exclude negative values; for typical data, however, there is no difficulty in having the extended ranges.

For \( n = 3 \) repetitions, the following data were obtained:

\[
(y_1, y_2, y_3) = (10.5, 10.7, 9.4).
\]

This is plotted in Figure 8.1.

Now consider what the statistical model (4) says about the observed data (5). The probability at the mathematical point (10.5, 10.7, 9.4) is of course 0. Hence to examine probabilities we must consider a small neighborhood \( N \) of the point (10.5, 10.7, 9.4); let \( \nu \) be the volume of \( N \). The probability at the observed point (10.5, 10.7, 9.4) is then given approximately by

\[
P(N|\theta) = (2\pi)^{-3/2} \exp \left\{ -(1/2)[(10.5 - \theta)^2 + (10.7 - \theta)^2 + (9.4 - \theta)^2] \right\} \nu,
\]
where $\theta$ takes values in $\Omega = \mathbb{R}$. Of course, $\nu$ could be arbitrarily small and certainly should be small for the approximation using the density function at the center of the neighborhood. Note that we write $P(N|\theta)$ for the probability in the set $N$ as calculated from the distribution determined by the parameter value $\theta$.

For any neighborhood of the observed data point with volume $\nu$, the probability (6) is a function of $\theta$; it records the probability for what has been observed as a function of the possible values $\theta$ for the physical quantity being measured. The function is called the likelihood function (of $\theta$) from the observed response $y = (10.5, 10.7, 9.4)$ and designated $L(y|\cdot)$:

$$L(y|\theta) = (2\pi)^{-3/2} \exp \left\{ -\frac{1}{2} [(10.5 - \theta)^2 + (10.7 - \theta)^2 + (9.4 - \theta)^2] \right\} \nu$$

$$= c \exp \left\{ -\frac{1}{2} \sum_{i=1}^{3} (y_i - 10.2)^2 + 3(10.2 - \theta)^2 \right\}$$

$$= c \exp \left\{ -\frac{1}{2(1/3)} (10.2 - \theta)^2 \right\};$$

for simplification we have used the standard identity (7.1.1) for sums of squares and we have incorporated some multiplicative constants into the arbitrary constant $c$ derived from the arbitrary volume $\nu$.

The likelihood function (7) is plotted in Figure 8.2. The likelihood function is, in fact, not a single function but an equivalence class of similarly shaped curves; the class is generated by taking arbitrary values for $\nu$ in (6) or for the constant $c$ in (7); three typical curves in the class are plotted in Figure 8.2. Note that the likelihood function in this example has the shape of the normal density function as located at $\theta = 10.2$ and scaled by the factor $1/\sqrt{3} = 0.577$.

The likelihood function provides the basic probability assessment of an observed response; it is what the model says about the observed response.

**C  \ ASSESSMENT OF THE OBSERVED LIKELIHOOD FUNCTION**

The likelihood function in Figure 8.2 records the probability for the observed response under various possible values for the parameter; it records what the model says about the observed response. For some values of $\theta$ close to 10.2, the probability for the observed response is relatively high; and for other values of $\theta$, three or four
scaling units from 10.2, the probability for the observed response is relatively very low.

Note that the likelihood function provides only a relative or comparative assessment of the various possible values for the parameter $\theta$. This follows from the presence of the arbitrary multiplicative constant $c$ or arbitrary volume $v$. Of course, each of the similarly shaped curves gives the same ratio of probabilities as determined for one $\theta$ value relative to another $\theta$ value; such a ratio is called the likelihood ratio for those $\theta$ values.

Our development makes clear the need for the arbitrary constant $c$: no particular value for the volume at the observed response has any special significance. Some texts neglect the arbitrary constant; this is a mistake both from the point of view of interpretation and of the theoretical results that depend on it (for example, the results in Sections 8.4 and 8.5).

Consider the observed likelihood function as recorded in Figure 8.2. The parameter value $\theta = \hat{\theta} = 10.2$ is the value for which the observed sample has maximum probability; it is called the maximum likelihood estimate (MLE) of the true parameter value. The likelihood function drops off from the maximum value and reaches the proportion

$$e^{-2/2} = e^{-2} = 0.135$$

of maximum value at two scaling units from the maximizing point. Accordingly, we form the interval (9.05, 11.35) for which the likelihood is greater than 0.135 of the maximum value; with nonnormal likelihood shapes, the fraction 0.135 gives a comparison in terms of the familiar normal shape.

Some developments of statistical theory have suggested a likelihood principle. The likelihood principle says that any inferences in a statistical investigation should use only the observed likelihood function; and that any other information about the data or about the statistical model should be suppressed. More specifically, it says that if two different statistical applications produce the same observed likelihood
function, then inferences concerning the true parameter value should be the same. This is a severe principle in the amount that it excludes—particularly in the exclusion of all information concerning the statistical model beyond that implicitly contained in the observed likelihood function. The development of theory in this text does not endorse the likelihood principle. In fact, much of the remaining developments run counter to the likelihood principle.

For the present context we note that the likelihood function presents all that the model has to say about an observed response value. We are, of course, free to consider other possible response values and to calculate the corresponding likelihood functions. Indeed, we can even consider the distribution of possible likelihood functions that would be obtained under any particular value for the parameter.

### D  General Definition of the Likelihood Function

Consider a statistical model

\[(8) \quad \{ f(\cdot | \theta) : \theta \in \Omega \}, \]

and let \(y\) be a possible value for the response. The probability for the response value \(y\) as calculated for a small neighborhood \(N\) with volume \(v\) is

\[(9) \quad L(y | \theta) = f(y | \theta) v.\]

This gives us the likelihood function \(L(y | \cdot)\) from the observed response value \(Y\), where

\[(10) \quad L(y | \theta) = cf(y | \theta).\]

and \(c\) is arbitrary in \((0, \infty)\); see Figure 8.3. The constant \(c\) generates an equivalence class of similarly shaped functions; several of the similarly shaped curves are plotted in the figure. Sometimes constants from \(f(y | \theta)\) can be combined with \(v\); accordingly, we have used \(c\) for the arbitrary positive multiplicative constant.

Often if \(L\) does not take zero values, there is convenience in using the log-likelihood function \(l(y | \cdot)\) from the observed response \(y\), where

\[(11) \quad l(y | \theta) = \ln f(y | \theta) + a,\]

and \(a\) is arbitrary in \((-\infty, \infty)\); see Figure 8.3.

And occasionally if \(L\) is differentiable with respect to a real parameter \(\theta\), there is convenience in using the score function \(S(y | \cdot)\) from the observed response \(y\), where

\[(12) \quad S(y | \theta) = \frac{\partial}{\partial \theta} \ln f(y | \theta) = \frac{\partial}{\partial \theta} l(y | \theta);\]

this records the slope of the log-likelihood function. See Figure 8.3 and note that \(S(y | \cdot)\) does not have an arbitrary constant.

If we have any one of the three functions \(L(y | \cdot)\), \(l(y | \cdot)\), \(S(y | \cdot)\), then we can calculate the others; for example, from \(S(y | \theta)\) we can integrate to obtain \(l(y | \cdot)\) and the additive constant \(a\) emerges as the arbitrary constant of integration. Thus the three functions convey the same information.

The arbitrary constant that generates the equivalence class of similarly shaped functions can occasionally be a nuisance. One way of avoiding the nuisance is to
choose a representative from the equivalence class. For example, if there is a value \( \theta_{\Delta} \) for which \( f(y | \theta_{\Delta}) > 0 \) for all possible sample points, then we could use the representative

\[
L_1(y | \theta) = \frac{f(y | \theta)}{f(y | \theta_{\Delta})}, \quad l_1(y | \theta) = \ln f(y | \theta) - \ln f(y | \theta_{\Delta});
\]

this amounts to choosing the \( L \) curve that has height 1 at \( \theta_{\Delta} \); see Figure 8.4. Or if each likelihood curve has a finite maximum, we could use

\[
L_2(y | \theta) = \frac{f(y | \theta)}{f(y | \hat{\theta})}, \quad l_2(y | \theta) = \ln f(y | \theta) - \ln f(y | \hat{\theta}),
\]

where \( \hat{\theta} = \hat{\theta}(y) \) is the parameter value at which the likelihood is a maximum; see Figure 8.4. Or if there are only two parameter values, that is, if \( \Omega = \{\theta_1, \theta_2\} \), then we could use

\[
L_3(y) = \frac{f(y | \theta_2)}{f(y | \theta_1)}, \quad l_3(y) = \ln f(y | \theta_2) - \ln f(y | \theta_1).
\]
FIGURE 8.4
Likelihood and log-likelihood functions standardized (a) with respect to a value \( \theta_\Delta \), (b) with respect to the maximum likelihood value \( \hat{\theta} \).

which is called the observed likelihood ratio (for \( \theta_2 \) versus \( \theta_1 \)). Note that \( L_3(y) \) is just a real number and hardly needs to be presented graphically.

E EXAMPLES

Consider three examples, two involving the normal distribution and one involving the binomial distribution.

EXAMPLE 1

The location normal: For example, the electrolytic crucibles in Section B. Let \( y = (y_1, \ldots, y_n)' \) be a sample from a normal(\( \theta, \sigma_0^2 \)) distribution with \( \theta \) in \( \Omega = R \); note that \( \sigma_0 \) designates a known value for the scaling \( \sigma \). The likelihood function \( L(y|\cdot) \) from an observed response \( y \) is given by

\[
L(y|\theta) = c(2\pi\sigma_0^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} (y_i - \theta)^2 \right\}
\]

\[
= c \exp \left\{ -\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} (y_i - \bar{y})^2 - \frac{n}{2\sigma_0^2} (\bar{y} - \theta)^2 \right\}
\]

\[
= c \exp \left\{ -\frac{n}{2\sigma_0^2} (\bar{y} - \theta)^2 \right\};
\]

and, correspondingly, the log-likelihood \( l(y|\cdot) \) and score \( S(y|\cdot) \) functions are given by

\[
l(y|\theta) = -\frac{n}{2\sigma_0^2} (\bar{y} - \theta)^2 + a,
\]
See Figure 8.5 and note that the likelihood function has normal density shape located at \( \bar{y} \) and scaled by \( \sigma_0/\sqrt{n} \). Also note that the log-likelihood is quadratic with maximum at \( \bar{y} \). And note that the score is linear with slope \(-n/\sigma_0^2\). The value of \( \theta \) that maximizes the likelihood can be obtained by solving

\[
S(y|\theta) = \frac{n}{\sigma_0^2}(\bar{y} - \theta).
\]

the resulting maximum likelihood estimate MLE for \( \theta \) is

\[
\hat{\theta} = \bar{y}.
\]

As a point of particular interest, note that the same likelihood function is obtained from all the sample points \((y_1, \ldots, y_n)\) that have a particular value for the function \( \bar{y} \). Thus the model says the same thing about each of the sample points that have a particular value for the function \( \bar{y} \). Accordingly, for any analysis of observed data with this normal model, it suffices to have just the value of \( \bar{y} \) together with the model. In more general situations, both location and scaling of the likelihood function may depend on the response.

**EXAMPLE 2**

The Bernoulli: Let \( x = (x_1, \ldots, x_n) \) be a sample of \( n \) from the Bernoulli(\( \rho \)) distribution with \( \rho \) in \([0, 1] \). The likelihood function \( L(x|\cdot) \) from an observed response \( x \) is given by
\( \mathcal{L}(x|\rho) = c\rho^{n\rho}q^{n-2n}. \)

Now suppose that \( n = 100 \) and that from observed data we obtain \( \sum x_i = 55. \) Then \( \mathcal{L}(x|\cdot), \log\mathcal{L}(x|\cdot), \) and \( S(x|\cdot) \) are given by

\[
\mathcal{L}(x|\rho) = c\rho^{55}(1 - \rho)^{45},
\]

\[
\log\mathcal{L}(x|\rho) = 55 \ln \rho + 45 \ln q + a,
\]

\[
S(x|\rho) = \frac{55}{\rho} - \frac{45}{q}.
\]

Note that the likelihood function has beta(56, 46) shape. The maximum likelihood estimate for \( \rho \) can be obtained by solving

\[
S(x|\rho) = \frac{55}{\rho} - \frac{45}{q} = 0,
\]

which gives \( \hat{\rho} = 55/100; \) this is the proportion or estimated probability which we have discussed earlier.

As a point of particular interest, note that the same likelihood function is obtained from all the sample points \( (x_1, \ldots, x_n) \) that have a particular value of the function \( \sum x_i; \) thus the model says the same thing about each of the sample points with a particular value of the function \( \sum x_i. \) Accordingly, for any analysis of observed data with this Bernoulli model, it suffices to have just the value of \( \sum x_i, \) together with the model.

The arguments used for the definition of the likelihood function do not immediately apply to the case of a discrete distribution such as the Bernoulli. We can, however, always contemplate observing and modeling some additional response characteristic with a continuous distribution—for example, the orientation of a coin in addition to, say, heads or tails. The arguments would then apply and we would obtain the likelihood function as all that the model says about an observed response.

**EXAMPLE 3**

The location-scale normal: Let \( (y_1, \ldots, y_n) \) be a sample from a normal(\( \mu, \sigma^2 \)) distribution with \( \theta = (\mu, \sigma^2) \) in \( R \times R^+ \), for example the measurement of current flow in a conductor. The likelihood \( \mathcal{L}(y|\cdot), \) log-likelihood \( \log\mathcal{L}(y|\cdot), \) and score \( S(y|\cdot) \) functions are given by

\[
\mathcal{L}(y|\mu, \sigma^2) = c(2\pi\sigma^2)^{-n/2} \left( \exp - \frac{1}{2\sigma^2} \sum (y_i - \mu)^2 \right)
\]

\[
= c\sigma^{-n/2} \exp \left( - \frac{1}{2\sigma^2} \sum (y_i - \bar{y})^2 \right) \exp \left( - \frac{n}{2\sigma^2}(\bar{y} - \mu)^2 \right);
\]

\[
\log\mathcal{L}(y|\mu, \sigma^2) = -\frac{n}{2} \ln \sigma^2 - \frac{\sum (y_i - \bar{y})^2}{2\sigma^2} - \frac{(\bar{y} - \mu)^2}{2\sigma^2/n} + a,
\]

\[
S(y|\mu, \sigma^2) = \frac{\partial}{\partial \mu} \log\mathcal{L}(y|\mu, \sigma^2)
\]
\[
S_2(y \mid \mu, \sigma^2) = \frac{\partial}{\partial \sigma^2} \log \text{Likelihood} \\
= - \frac{n}{2\sigma^2} + \frac{\sum (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2}{2\sigma^4}.
\]

See Figure 8.6 and note that the likelihood function is defined over the upper half-plane \( \Omega = R \times \mathbb{R}^+ \). It is a function that drops to zero at the boundary and rises to a hump at the maximum-likelihood estimate. Also note that the slope can be calculated in the direction \( \mu \), giving \( S_1 \), and in the direction \( \sigma^2 \), giving \( S_2 \). The maximum-likelihood estimate can be obtained by solving

\[
S_1(y \mid \mu, \sigma^2) = \frac{n}{\sigma^2} (\bar{y} - \mu) = 0,
\]

\[
S_2(y \mid \mu, \sigma^2) = - \frac{n}{2\sigma^2} + \frac{\sum (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2}{2\sigma^4} = 0;
\]

we obtain

\[
\hat{\mu} = \bar{y}, \quad \hat{\sigma}^2 = \frac{\sum (y_i - \bar{y})^2}{n} = \frac{n-1}{n} s_y^2.
\]

As a point of particular interest, note that the same likelihood function is obtained from all the sample points \((y_1, \ldots, y_n)\) that have a particular value for the function \((\bar{y}, s_y^2)\); thus the model says the same things about each of these
sample points. Accordingly, for any analysis of observed data with this normal model it suffices to have just the value of \((\bar{Y}, s_y^2)\) together with the model.

**F Exercises**

1. A new drug is administered to 100 infected animals. Assume that the responses of the animals are independent; let \(\rho\) be the probability of recovery. For the data in Example 2, sketch the likelihood, log-likelihood, and score functions.

2. The lifetime of a certain transducer is known to be exponential(\(\theta\)) with \(\theta\) in \((0, \infty)\). A sample of three yielded \(Y = (2.1, 1.3, 2.9)\).
   (a) Determine and sketch the functions \(L, l, \) and \(S.\)
   (b) Calculate the MLE estimate of \(\theta\).

3. The number of defects in a 1000-ft spool of cable is Poisson(\(\theta\)) with \(\theta\) in \((0, \infty)\). A sample of three yielded \(Y = (3, 7, 5)\).
   (a) Determine and sketch the functions \(L, I,\) and \(S.\)
   (b) Calculate the MLE estimate of \(\theta\).

4. Let \(Y = (2.1, 3.9, 3.4)\) be a sample of three from the location Laplace distribution with density \(f(y|\theta) = \exp \left(-|y - \theta|\right)/2.\)
   (a) Determine and sketch the functions \(L\) and \(I.\)
   (b) Calculate the MLE estimate of \(\theta.\)
   (c) Record a different sample with a different likelihood function and yet the same MLE estimate of \(\theta.\)

5. Let \(Y = (1.2, 4.7, 0.8, 3.5)\) be a sample from the uniform(0, \(\theta\)) distribution with \(\theta\) in \((0, \infty)\).
   (a) Determine and sketch the functions \(L\) and \(I;\) allow the value \(-\infty\) for \(I.\)
   (b) Calculate the MLE estimate of \(\theta.\)

6. Let \(Y = (y_1, \ldots, y_n)\) be a sample from the normal(\(\mu_0, \sigma\)) distribution with \(\sigma\) in \((0, \infty)\).
   (a) Determine and sketch the functions \(L, I,\) and \(S\) for the parameter \(\sigma^2.\)
   (b) Calculate the MLE estimate of \(\sigma^2.\)
   (c) Show that the same likelihood function is obtained from all sample points that have a particular value for \(\sum (y_i - \mu_0)^2.\)

**G Problems**

7. Let \(Y = (1.2, 4.7, 0.8, 3.5)\) be a sample from the uniform(\(\theta_1, \theta_2\)) distribution with \(-\infty < \theta_1 < \theta_2 < \infty.\)
   (a) Determine the likelihood function and sketch it.
   (b) Find the MLE estimate of \((\theta_1, \theta_2).\)

8. A certain electrical component is known to have an exponential(\(\theta\)) lifetime. A sample of 10 was tested for 50 weeks: 3 survived the period; 7 failed during the test period at lifetimes 4, 27, 14, 24, 49, 4, and 33.
   (a) Determine the likelihood function and sketch it.
   (b) Find the MLE of \(\theta.\)

9. An offspring in a breeding investigation can be \(A, B,\) or \(C\) with respective probabilities \((2 + \rho)/4, (1 - \rho)/2,\) or \(\rho/4.\) For a sample of \(n\), let \(a, b,\) and \(c\) be the number of progeny of types \(A, B,\) and \(C.\) The following data were obtained: \(a = 58, b = 33,\) and \(c = 9.\)
   (a) Determine the likelihood function.
   (b) In terms of \(a, b,\) and \(c,\) determine a quadratic equation for the MLE estimate of \(\rho.\)
8.2

DISTRIBUTION THEORY FOR THE NORMAL MODEL

The normal distribution has a prominent place in statistics. This is due partially to its success in providing a first approximation to the variation found in many applications, and due partially to some exceptional mathematical properties that make many theoretical calculations almost trivial for the normal—calculations that are virtually impossible for other forms for the variation. Indeed, the theory in Chapters 9 and 10 is essentially limited to the normal and certain related distribution forms. We will examine statistical inference for other forms for the variation in Chapter 11.

The exceptional property of the normal distribution is the rotational symmetry of the distribution for a sample \((y_1, \ldots, y_n)\) from a central normal \((0, \sigma)\) distribution. The density function is

\[
f(y) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum y_i^2 \right\};
\]

note that it depends only on the squared distance \(r^2 = \sum y_i^2\) from the origin; see Figure 8.7. Thus if we should use orthogonally rotated axes, the normal distribution would look exactly the same.

In Section 8.1 we considered an observed sample in relation to a normal model with unknown location or unknown location and scale. And we saw that we needed from the data only the value of \(\overline{y}\) in the first case and the value of \((\overline{y}, s^2_y)\) in the second case.

In this section we establish two theorems that formally express the rotational symmetry of the normal, and we use these theorems to obtain the marginal distributions for the functions \(\overline{y}\) and \(s^2_y\) based on a normal model.

![Figure 8.7](image)

**FIGURE 8.7**

Distribution of a sample \((y_1, \ldots, y_n)\) from the normal\((0, \sigma)\) is rotationally symmetrical about the origin.
A SYMMETRY: SAMPLE
FROM THE NORMAL(0, σ)

A sample \((y_1, \ldots, y_n)\) from the normal(0, \(σ\)) distribution has a rotationally symmetric distribution on \(R^n\). Some aspects of this are given by the following theorem.

**THEOREM 1**

If \(y = (y_1, \ldots, y_n)'\) is a sample from the normal(0, \(σ\)) distribution and if \(C\) is an \(n \times n\) orthogonal matrix, then \(u = C'y\) also has the distribution of a sample from the normal(0, \(σ\)) distribution.

A special case of this theorem was proved as Example 4.2.4. The proof of the theorem is omitted, as it is a special case of Theorem 2 in Section B. We complete this section with a brief discussion of orthogonal transformations.

Consider an \(n \times n\) matrix formed by \(n\) column vectors in \(R^n\):

\[ C = (c_1, \ldots, c_n) = \begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nn} \end{pmatrix} \]

And suppose that the vectors \(c_1, \ldots, c_n\) are each of length 1 and mutually orthogonal:

\[ (c_i, c_i) = c_i^2 = \sum_{a=1}^{n} c_{ai}^2 = 1, \]

\[ (c_i, c_j) = c_i^2 = \sum_{a=1}^{n} c_{ai}c_{aj} = 0, \quad i \neq j. \]

We call such a set of \(n\) vectors an orthonormal set. The relations (2) can be expressed compactly by the following matrix equation:

\[ C'C = I, \]

where \(I\) is the \(n \times n\) identity matrix. Note that \(C'C\) is the inner-product matrix for the \(n\) vectors \(c_1, \ldots, c_n\); see Supplement Section 2.5B. A matrix \(C\) satisfying property (2) or (3) is called an orthogonal matrix. Note from (3) that an orthogonal matrix has a very simple inverse—just the transpose

\[ C^{-1} = C'. \]

Consider an orthogonal matrix \(C\) and suppose that we use \(C'\) as a transformation matrix carrying \(y\) in \(R^n\) into \(u\) in \(R^n\):

\[ u = C'y, \]

or, equivalently,

\[ u_i = c'_i'y = c_{i1}y_1 + \cdots + c_{in}y_n. \]

Premultiplication by the matrix \(C\) gives

\[ Cu = CC'y = y; \]

thus
FIGURE 8.8
Coordinates $u_1$ and $u_2$ with respect to the new axes $c_1 = (\cos a, \sin a)'$, $c_2 = (-\sin a, \cos a)'$.

\[ y = Cu = u_1c_1 + \ldots + u_nc_n. \]

Note that equation (7) expresses $y$ as a sum of components in the various orthogonal directions $c_1, \ldots, c_n$; and that the coordinate $u_i$ in the direction $c_i$ is obtained as the inner product (6) of $y$ with the unit vector $c_i$; see Figure 8.8.

Consider the orthogonal transformation (5). The squared length of $y$ can be expressed in terms of the new coordinates $u_1, \ldots, u_n$ by using relation (3):

\[ (y, y) = \sum_{i=1}^{n} y_i^2 = y'y = u'Cu = u'u = \sum u_i^2 = (u, u). \]

We thus have a decomposition of the vector $y$ and of its squared length $\sum y_i^2$ according to the $n$ orthogonal directions given by $c_1, \ldots, c_n$ (Table 8.1).

**TABLE 8.1**
Components and lengths of vector $y$

<table>
<thead>
<tr>
<th>Subspace $L(c_i)$</th>
<th>Dimension $1$</th>
<th>Component $u_ic_i$</th>
<th>Squared Length $u_i^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L(c_1)$</td>
<td>$1$</td>
<td>$u_1c_1$</td>
<td>$u_1^2$</td>
</tr>
<tr>
<td>$L(c_2)$</td>
<td>$1$</td>
<td>$u_2c_2$</td>
<td>$u_2^2$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$L(c_n)$</td>
<td>$1$</td>
<td>$u_nc_n$</td>
<td>$u_n^2$</td>
</tr>
<tr>
<td>\textbf{n}</td>
<td>\textbf{}</td>
<td>\textbf{y}</td>
<td>$\sum y_i^2$</td>
</tr>
</tbody>
</table>

**B SYMMETRY: VARIATION FROM THE NORMAL($0, \sigma$)**

Consider independent coordinates $y_1, \ldots, y_n$ and suppose that each coordinate has normal($0, \sigma$) variation. An orthogonal transformation leaves these distribution properties unchanged; for an example, see Figure 8.9.

**THEOREM 2**

If $y_1, \ldots, y_n$ are statistically independent and $y_1$ is normal($\mu_1, \sigma$) and if $u = C'y$, where $C$ is an $n \times n$ orthogonal matrix, then $u_1, \ldots, u_n$ are statistically in-
dependent and \( u_i \) is normal(\( \nu_i, \sigma \)), where the \( \nu_i \) are given by the same orthogonal transformation, \( \nu = C'\mu \).

Proof A change of variable for a probability density is best expressed in terms of a probability differential, the formal component of an integral expression for a probability. The probability differential for \( \mathbf{y} \) is

\[
(2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{\sum (\nu_i - \mu_i)^2}{2\sigma^2} \right\} \prod d\nu_i = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{(y - \mu)'(y - \mu)}{2\sigma^2} \right\} dy.
\]

The effect of the transformation on the volume differential is

(8) \[ d\mathbf{y} = |C| d\mathbf{u} = d\mathbf{u}; \]

for clearly the Jacobian determinant of an orthogonal transformation is unity: \( |C|^2 = |C'C| = |I| = 1 \). The effect of the transformation on the squared distance in the exponent is obtained by using \( y = Cu, \mu = C\nu \); and thus \( y - \mu = C(u - \nu) \):

(9) \[ (y - \mu)'(y - \mu) = (u - \nu)'C'C(u - \nu) = (u - \nu)'(u - \nu). \]

The probability differential for \( u \) is then obtained by substituting (8) and (9) in the differential for \( \mathbf{y} \):

\[
(2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{(u - \nu)'(u - \nu)}{2\sigma^2} \right\} d\mathbf{u} = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{\sum (u_i - \nu_i)^2}{2\sigma^2} \right\} d\mathbf{u};
\]

this is the probability differential for the distribution cited in the theorem.

---

**FIGURE 8.9**

Distribution of a sample from the normal(\( \mu, \sigma \)) distribution is rotationally symmetric about the central point \( \mu1 \) on the line \( \mathcal{L}(1) \).
Consider the interpretation of this theorem in terms of the orthogonal components discussed in Section A. The component of \( y \) in the direction \( c_i \) is \( u_i c_i \), and the component of \( \mu \) in the same direction is \( v_i c_i \):

\[
u_i = c_i \mu.
\]

The theorem says that \( u_i \) is normally distributed about \( v_i \) with standard deviation \( \sigma \).

**C** THE DISTRIBUTION OF \( \bar{y} \) AND \( s^2_u \) FOR A NORMAL MODEL

Consider the case of a sample \((y_1, \ldots, y_n)\) from the normal \((\mu, \sigma)\) distribution. From Section 8.1 we know that with unknown location we need only the value of \( \bar{y} \) and with unknown location and scale we need only \((\bar{y}, s^2_u)\). Thus \( \bar{y} \) and \( s^2_u \) are key functions for sampling from the normal.

In Section 7.1 we saw that the vector \( y \) can be expressed in terms of two components: the projection \( \bar{y} 1 \) onto the line \( \mathcal{L}(1) \), and the projection \( y - \bar{y} 1 \) onto the plane \( \mathcal{L}^\perp(1) \) [the orthogonal complement of \( \mathcal{L}(1) \)]; and also that the squared length of \( y \) is the sum of the squared lengths of the two components:

\[
\sum y_i^2 = n\bar{y}^2 + \sum (y_i - \bar{y})^2 = n\bar{y}^2 + (n - 1)s^2_u.
\]

Recall Figure 7.1. This gives Table 8.2. We now obtain the distribution theory for \( \bar{y} \) and \( s^2_u \) by using new axes that align with the subspaces \( \mathcal{L}(1) \) and \( \mathcal{L}^\perp(1) \).

**TABLE 8.2**
Components of vector \( y \); corresponding squared lengths

<table>
<thead>
<tr>
<th>Subspace</th>
<th>Dimension</th>
<th>Component</th>
<th>Squared Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathcal{L}(1) )</td>
<td>1</td>
<td>( \bar{y} 1 )</td>
<td>( n\bar{y}^2 )</td>
</tr>
<tr>
<td>( \mathcal{L}^\perp(1) )</td>
<td>( n - 1 )</td>
<td>( y - \bar{y} 1 )</td>
<td>( \frac{\Sigma (y_i - \bar{y})^2}{n} )</td>
</tr>
</tbody>
</table>

The sample average \( \bar{y} \) measures location relative to the 1-vector; accordingly, as a first vector \( c_i \) for an orthogonal matrix \( C \), we choose the unit vector in the direction of the 1-vector:

\[
c_i' = \frac{1}{\sqrt{n}} (1, \ldots, 1) = \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right).
\]

An orthogonal transformation \( u = C' y \) then has the form

\[
\begin{pmatrix}
\frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & \cdots & 0 \\
\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\frac{1}{\sqrt{n(n-1)}} & \cdots & \frac{1}{\sqrt{n(n-1)}} & \cdots & \frac{-(n-1)}{\sqrt{n(n-1)}}
\end{pmatrix}
\begin{pmatrix}
y_1 \\
y_2 \\
y_3 \\
\vdots \\
y_n
\end{pmatrix}
\]
the remaining rows of $C'$ are a convenient set of orthonormal vectors that span $\mathcal{L}_k(1)$.

Now consider the effect of the transformation $C'$ on the components of $\mathbf{y} = \bar{y}\mathbf{1} + (\mathbf{y} - \bar{y}\mathbf{1})$:

$$
C' \begin{pmatrix} \bar{y} \\ y_1 \\ \vdots \\ y_n \end{pmatrix} = C' \begin{pmatrix} \sqrt{n}\bar{y} \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} u_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix},
$$

$$
C' \begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix} = \begin{pmatrix} 0 \\ u_2 \\ \vdots \\ u_n \end{pmatrix};
$$

this uses the orthogonality between the first row (representing the 1-vector) of $C'$ and the remaining rows. Correspondingly, for the squared lengths we have

$$
n\bar{y}^2 = u_1^2, \quad \sum (y_i - \bar{y})^2 = u_2^2 + \cdots + u_n^2.
$$

We can now give a simple proof of the following theorem.

**Theorem 3**

If $\mathbf{y} = (y_1, \ldots, y_n)$ is a sample from the normal($\mu, \sigma$) distribution, then $\bar{y}$ and $\Sigma (y_i - \bar{y})^2$ are statistically independent, $\bar{y}$ is normal($\mu$, $\sigma/\sqrt{n}$), and $\Sigma (y_i - \bar{y})^2$ has the distribution of $\sigma^2 \chi^2$, where $\chi^2$ has the chi-square($n - 1$) distribution.

**Proof** Consider the orthogonal transformation $\mathbf{u} = C'\mathbf{y}$ as given by formula (10). From Theorem 2 we have that $u_1, \ldots, u_n$ are statistically independent and $u_1$ is normal($\nu_1$, $\sigma$), where

$$
\nu = C' \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix} = \begin{pmatrix} \frac{\mu}{\sqrt{n}} \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} \sqrt{n}\mu \\ 0 \\ \vdots \\ 0 \end{pmatrix}.
$$

the original mean is at $\mu\mathbf{1}$ on the 1-vector; the first new axis is in the direction of the 1-vector; accordingly, in terms of new coordinates, the mean is on the first new axis.

First, we have that $\sqrt{n}\bar{y} = u_1$, which is normal($\sqrt{n}\mu$, $\sigma$); thus $\bar{y}$ is

normal($\mu$, $\sigma/\sqrt{n}$).

Second, we have that $\Sigma (y_i - \bar{y})^2 = u_2^2 + \cdots + u_n^2$, which is the sum of squares of $n - 1$ normal($0$, $\sigma$) variables. Hence by Problem 3.3.6, or by Problems 5.7.31 and 5.7.33, or by the detailed development in Supplement Section 3.6B, it follows that $\Sigma (y_i - \bar{y})^2/\sigma^2$ has the chi-square($n - 1$) distribution or, equivalently, that $s^2_y$ is distributed as $\sigma^2 \chi^2/(n - 1)$, where $\chi^2$ is chi-square($n - 1$).

Third, the independence of $u_1$ and $(u_2, \ldots, u_n)$ implies the independence of $\bar{y}$ and $s^2_y$ (use Proposition 3.3.5).

We shall use these results for statistical inference in the next section.

**D Exercises**

1. Show that the inner product of two vectors is invariant under an orthogonal transformation.
Let $u_1$ and $u_2$ be the vectors obtained by applying the orthogonal transformation $\mathbf{u} = \mathbf{C}^*\mathbf{y}$ to $y_1$ and $y_2$; show that

$$(y_1, y_2) = (u_1, u_2).$$

2. For a sample of size $n = 2$, express $s_1^2$ in terms of $(y_1 - y_2)$; use the transformation (10). Let $(y_1, y_2) = (10.5, 10.9)$; calculate $(u_1, u_2)$ as given by (10); use $1/\sqrt{2} = 0.707$.

3. For a sample of size $n = 3$, express $s_2^2$ in terms of $(y_1 - y_2)$ and $(y_1 + y_2 - 2y_3)$; use the transformation (10). Let $(y_1, y_2, y_3) = (5.50, 5.78, 5.31)$; calculate $(u_1, u_2, u_3)$ as given by (10); use $1/\sqrt{2} = 0.707, 1/\sqrt{3} = 0.577, 1/\sqrt{6} = 0.408$.

4. Verify that the matrix $\mathbf{C}$ is orthogonal:

$$
\mathbf{C}^* = \begin{pmatrix}
\frac{1}{\sqrt{5}} & 1/\sqrt{5} & \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\
\frac{2}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{2}{\sqrt{30}} & -3/\sqrt{30} \\
\frac{1}{\sqrt{2}} & -1/\sqrt{2} & 0 & 0 \\
\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -2/\sqrt{6} & 0 \\
0 & 0 & 0 & 1/\sqrt{2}
\end{pmatrix}.
$$

5. Let $\mathbf{y} = (y_1, y_2, y_3, y_21, y_22)^T$, $\mathbf{y}_1 = (y_{11}, y_{12}, y_{13})^T$, $\mathbf{y}_2 = (y_{21}, y_{22})^T$, and $\mathbf{u} = \mathbf{C}^*\mathbf{y}$. Use the matrix $\mathbf{C}$ in Exercise 4 to show that

$$u_1 = \sqrt{(1/5)}/2, \quad u_2 = (\sqrt{\mathbf{y}_1} - \mathbf{y}_2)/[(1/3) + (1/2)]^{1/2}.$$ 

$$u_3^2 + u_4^2 + u_5^2 = \sum_{i=1}^{3} (y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{2} (y_{2i} - \bar{y}_2)^2.$$ 

E. TWO-SAMPLE PROBLEMS

6. Let $(y_{11}, \ldots, y_{1n})$ be a sample from the normal$(\mu_1, \sigma)$ and $(y_{21}, \ldots, y_{2n})$ be a sample from the normal$(\mu_2, \sigma)$ in $R^2 \times R^+$. Use Theorem 3 to show that $\bar{y}_1, \bar{y}_2,$ and $S^2 = \sum (y_{1i} - \bar{y}_1)^2 + \sum (y_{2i} - \bar{y}_2)^2$ are statistically independent; that $\bar{y}_1$ is normal$(\mu_1, \sigma/\sqrt{m})$, $\bar{y}_2$ is normal$(\mu_2, \sigma/\sqrt{n})$, and $S^2$ is $\sigma^2 \chi^2$, where $\chi^2$ has the chi-square$(m + n - 2)$ distribution; of course, assume independent samples.

7. Verify that the matrix $\mathbf{C}$ is orthogonal:

$$
\mathbf{C}^* = \begin{pmatrix}
1 & \ldots & 1 & \ldots & 1 \\
\sqrt{m+n} & \ldots & \sqrt{m+n} & \ldots & \sqrt{m+n} \\
\sqrt{n}/n & \ldots & \sqrt{n}/n & \ldots & \sqrt{n}/n \\
\sqrt{m+n} & \ldots & \sqrt{m+n} & \ldots & \sqrt{m+n} \\
A_m & \ldots & A_m & \ldots & A_m \\
0 & \ldots & 0 & \ldots & 0
\end{pmatrix},
$$

where $A_m$ is the $(n - 1) \times n$ matrix obtained by deleting the first row of the matrix $\mathbf{C}^*$ in formula (10).

8. Use Theorem 2 and the matrix in Problem 7 to prove the distribution results quoted in Problem 6.

9. Let $(y_{11}, \ldots, y_{1n})$ be a sample from the normal$(\mu_1, \sigma_1)$ and $(y_{21}, \ldots, y_{2n})$ be a sample from the normal$(\mu_2, \sigma_2)$ in $R^2 \times (0, \infty)$. Let $S_1^2$ and $S_2^2$ be the respective sample variances. Use Problem 3.2.33 or the distribution results in Supplement Section 3.6B to show that if $\sigma_1^2 = \sigma_2^2$, then

$$F = \frac{s_1^2}{s_2^2}$$

has an $F(m - 1, n - 1)$ distribution and that in general $\sigma_2^2F/\sigma_1^2$ has an $F(m - 1, n - 1)$ distribution.
8.3

STATISTICAL INFERENCE FOR
THE NORMAL MODEL

We have noted that the normal distribution provides an approximation to the variation found in many applications and that the normal has some exceptional mathematical properties. In the preceding section we used these exceptional properties to derive the marginal distributions for the sample average $\bar{y}$ and the sample variance $s_y^2$.

In this section we use these marginal distribution results to examine inference methods for applications where the normal pattern for variation is appropriate. We then see that simple adjustments to the large-sample methods of Section 7.2 give methods that are exact for small samples with the normal model.

In practice, however, there is often more probability in the tails than is allowed by the normal distribution. The Student family, as indicated in Section 8.1A, can frequently provide a reasonable model for such cases. Inference methods for alternative distributions for variation will be examined in Chapter 11.

\section{Inference for the Mean: Variance Known}

Consider a random system with a response $y$ known to be normal with known variance $\sigma_y^2$; for example, the treatment applied to animals in Example 7.1.2 but now with variance known from a related investigation. As a model for $n$ repetitions, we have a sample $(y_1, \ldots, y_n)$ from the normal($\theta$, $\sigma^2$) distribution with $\theta$ in $\mathbb{R}$. We now consider how an observed sample, together with this model, can produce inferences concerning the mean $\theta$.

From Example 8.1.1 we know that it suffices to have just the value of $\bar{y}$ together with the model. And from Theorem 8.2.3 we know that $\bar{y}$ is normal($\theta$, $\sigma_0/\sqrt{n}$) or that

$$z = \frac{\bar{y} - \theta}{\sigma_0/\sqrt{n}}$$

is standard normal.

First, we consider the problem of testing the hypothesis $\theta = \theta_0$. For an observed sample $(y_1, \ldots, y_n)$ we can calculate the value $\bar{y}$ and can compare the value with the hypothesis distribution which is normal($\theta_0$, $\sigma_0/\sqrt{n}$). Or, equivalently, we can calculate the standardized deviation from $\theta_0$.

$$(1) \quad z = \frac{\bar{y} - \theta_0}{\sigma_0/\sqrt{n}}$$

and can compare the value with the standard normal distribution; see Appendix II. We then assess the hypothesis accordingly—as we have discussed in Section 7.2B. Thus we see that the large-sample test discussed in Section 7.2B is, in fact, an exact test for any sample size, provided that we have the restricted normal model.
Second, we consider the problem of forming a confidence interval for the location parameter \( \theta \). The distribution of the function

\[
Z = \frac{\bar{y} - \theta}{s_0/\sqrt{n}}
\]

is standard normal for each value of \( \theta \). Let \((-z_{\alpha/2}, z_{\alpha/2})\) be the central interval containing \(1 - \alpha\) probability for the standard normal; then

\[
P(-z_{\alpha/2} < \frac{\bar{y} - \theta}{s_0/\sqrt{n}} < z_{\alpha/2}) = 1 - \alpha.
\]

This can be rearranged as in Section 7.2C, giving

\[(2) \quad P(\bar{y} - z_{\alpha/2}s_0/\sqrt{n} < \theta < \bar{y} + z_{\alpha/2}s_0/\sqrt{n}) = 1 - \alpha.
\]

Thus \((\bar{y} \pm z_{\alpha/2}s_0/\sqrt{n})\) is an exact \(1 - \alpha\) confidence interval for the location \(\theta\). Thus we see that the large-sample confidence interval discussed in Section 7.2C is, in fact, an exact confidence interval for any sample size, provided that we have the restricted normal model. Recall the repeated sampling-frequency interpretation discussed in Section 7.2C.

**EXAMPLE 1**

An instrument for measuring current is unbiased and the error is approximately normal with standard deviation \(s = 0.025\). The following measurements were obtained for a constant current \(\theta\):

\[0.672, \ 0.619, \ 0.656, \ 0.651.\]

Some theory suggests the value \(\theta = 0.700\); this can be tested by calculating

\[
Z = \frac{\bar{y} - 0.700}{0.025/\sqrt{4}} = \frac{0.6495 - 0.700}{0.0125} = -4.04
\]

and comparing it with the standard normal; this is an extreme value for the standard normal, and it gives strong evidence against the hypothesis. A 95 percent confidence interval for \(\theta\) can be obtained from formula (2):

\[
(\bar{y} \pm 1.96 \frac{0.025}{\sqrt{4}}) = (0.6495 \pm 0.0245) = (0.6250, 0.6740).
\]

**B  INFERENCE FOR THE VARIANCE**

Now consider a random system with a response \(y\) known only to have normal variation, for example, the treatment applied to animals in Example 7.1.2. As a model for \(n\) repetitions we have a sample \((y_1, \ldots, y_n)\) from the normal \((\mu, \sigma)\) with parameter \(\theta = (\mu, \sigma)\) in \(R \times R^+\). In this section we examine how an observed sample can be used with this model for inference concerning the variance \(\sigma^2\).

From Example 8.1.3 we know that it suffices to have just the value of \((\bar{y}, s^2_\nu)\). And from Theorem 8.2.3 we know that \(\bar{y}\) is normal \((\mu, \sigma/\sqrt{n})\), that \((n - 1)s^2_\nu\) is \(\sigma^2\chi^2\), where \(\chi^2\) is chi-square \((n - 1)\), and that \(\bar{y}\) and \(s^2_\nu\) are statistically independent. The distribution of \(\bar{y}\) has unknown location \(\mu\); accordingly, an observed value can give no informa-
tion concerning the variance \( \sigma^2 \). We thus restrict our attention to \( s^2_y \) for inferences concerning \( \sigma^2 \).

First, we consider the problem of testing the hypothesis \( \sigma^2 = \sigma_0^2 \). For an observed sample \((y_1, \ldots, y_n)\), we can calculate the value of \((n - 1)s^2_y = \sum (y_i - \bar{y})^2\) and compare it with the hypothesis distribution, which can be described by \( \sigma_0^2 \chi^2 \). Or, equivalently, we can calculate

\[
(3) \quad \chi^2 = \frac{\sum (y_i - \bar{y})^2}{\sigma^2_0} = \frac{(n - 1)s^2_y}{\sigma^2_0}
\]

and can compare it with the chi-square\((n - 1)\) distribution; see Appendix II. We then assess the hypothesis accordingly—in the pattern we have used before.

Second, we consider the problem of forming a confidence interval for the variance \( \sigma^2 \) or for the standard deviation \( \sigma \). The distribution of the function

\[
\chi^2 = \frac{\sum (y_i - \bar{y})^2}{\sigma^2} = \frac{(n - 1)s^2_y}{\sigma^2}
\]

is chi-square\((n - 1)\). Let \((\chi^2_{1-a/2}, \chi^2_{\alpha/2})\) be a central interval containing \(1 - \alpha\) probability for the chi-square\((n - 1)\) distribution; see Appendix II. The chi-square distribution is asymmetric; the interval just suggested is central in a probability sense, \(\alpha/2\) probability excluded on each tail of the distribution. We then have

\[
P\left(\frac{(n - 1)s^2_y}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n - 1)s^2_y}{\chi^2_{1-a/2}}\right) = 1 - \alpha
\]

for each value for \( \sigma^2 \). The inequalities can be rearranged [take reciprocals and multiply through by \((n - 1)s^2_y\)], giving

\[
P\left(\frac{\chi^2_{\alpha/2}}{\chi^2_{1-a/2}} < \sigma^2 < \frac{\chi^2_{1-a/2}}{\chi^2_{\alpha/2}}\right) = 1 - \alpha,
\]

and it follows that

\[
\left(\frac{n - 1}{\chi^2_{\alpha/2}} s^2_y, \frac{n - 1}{\chi^2_{1-a/2}} s^2_y\right)
\]

is a \(1 - \alpha\) confidence interval for the variance, and, of course,

\[
\left(\frac{(n - 1)^{1/2}}{\chi^2_{\alpha/2}} s_y, \frac{(n - 1)^{1/2}}{\chi^2_{1-a/2}} s_y\right)
\]

is a \(1 - \alpha\) confidence interval for the standard deviation.

**TABLE 8.3**
Factors for \( s^2_y \) and \( s_y \)

<table>
<thead>
<tr>
<th>Factors for ( s^2_y )</th>
<th>Factors for ( s_y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower</td>
<td>Upper</td>
</tr>
<tr>
<td>0.59</td>
<td>2.09</td>
</tr>
<tr>
<td>0.67</td>
<td>1.64</td>
</tr>
<tr>
<td>0.77</td>
<td>1.35</td>
</tr>
</tbody>
</table>
There is some interest in seeing the 95 percent factors that must be applied to \( s_y^2 \) and \( s_y \) for various sample sizes (Table 8.3). Thus, even with a sample of \( n = 101 \), the factors allow a range of 12 percent low to 16 percent high for the standard deviation \( \sigma \) relative to \( s_y \).

**EXAMPLE 2**

An instrument for measuring pH is unbiased and the error is approximately normal. A sample of five yielded \( \bar{y} = 7.921 \), \( s_y = 0.017 \). A 95 percent confidence interval for \( \sigma \) can be obtained from formula (6) with \( n = 5 \):

\[
\left( \frac{\sqrt{4}}{\sqrt{11.143}} \cdot 0.017, \quad \frac{\sqrt{4}}{\sqrt{0.4844}} \cdot 0.017 \right) = (0.010, 0.049);
\]

the percentage points are available from Appendix II.

**C INFERENCE FOR THE MEAN: VARIANCE UNKNOWN**

Consider further a random system with a response \( y \) known only to have normal variation. As a model for \( n \) repetitions we have a sample \( (y_1, \ldots, y_n) \) from the normal(\( \mu, \sigma \)) with parameter \( \theta = (\mu, \sigma) \) in \( R \times R^+ \). In this section we examine how an observed sample can be used with this model for inference concerning the mean \( \mu \).

From Example 8.1.3 we know that it suffices to have just the value of \( (\bar{y}, s_y^2) \). And from the distribution theory recorded in Section B, we know that the deviation \( \bar{y} - \mu \) of the average \( \bar{y} \) from the true mean \( \mu \) has a normal(0, \( \sigma / \sqrt{n} \)) distribution. The adjusted standard deviation \( s_y / \sqrt{n} \) provides an estimate of \( \sigma / \sqrt{n} \). Accordingly, we consider the deviation \( \bar{y} - \mu \) in units of its estimated standard deviation:

\[
t = \frac{\bar{y} - \mu}{s_y / \sqrt{n}} = \frac{\sqrt{n}(\bar{y} - \mu)/\sigma}{s_y / \sigma}.
\]

Consider the distribution of \( t \): the numerator of the right-hand expression is standard normal; the denominator is \( \chi^2 / \sqrt{n - 1} \), where \( \chi^2 \) is chi-square(\( n - 1 \)); the numerator and denominator are independent. By Problem 3.2.35 for the canonical Student or by the details in Supplement Section 3.6B it follows that \( t \) has the Student(n – 1) distribution. We thus restrict our attention to the function \( t \) for inference concerning \( \mu \).

First, we consider the problem of testing the hypothesis \( \mu = \mu_0 \). We can measure the deviation of \( \bar{y} \) from \( \mu_0 \) by calculating

\[
t = \frac{\bar{y} - \mu_0}{s_y / \sqrt{n}};
\]

note that if \( \mu = \mu_0 \), this \( t \) is Student(n – 1). For an observed sample \( (y_1, \ldots, y_n) \), we can calculate the value of \( t \) in formula (7) and can compare it with the Student(n – 1) distribution; see Appendix II. We then assess the hypothesis accordingly—as we have discussed in Section 7.2E. Thus we see that the large-sample test discussed in Section 7.2E is easily adjusted to give an exact test for any sample size, provided that we have the normal model.

Note that if \( \mu \) is different from \( \mu_0 \), then the distribution of \( t \) in formula (7) is
noncentral Student \((n - 1)\) with noncentrality \(\delta = \sqrt{n}(\mu - \mu_0)/\sigma\); see Problem 2 in Supplement Section 3.6.

Second, we consider the problem of forming a confidence interval for the location parameter \(\mu\). The distribution of the function

\[
(8) \quad t = \frac{\bar{y} - \mu}{s_{\bar{y}}/\sqrt{n}}
\]

is the Student \((n - 1)\) distribution for each value of \(\theta = (\mu, \sigma)\) in \(R \times R^+\); it is symmetric about \(0\). Let \((- t_{\alpha/2}, t_{\alpha/2})\) be the central interval containing \(1 - \alpha\) probability for the Student \((n - 1)\) distribution; see Appendix II. We then have

\[
P(- t_{\alpha/2} < \frac{\bar{y} - \mu}{s_{\bar{y}}/\sqrt{n}} < t_{\alpha/2}) = 1 - \alpha
\]

for each value of \((\mu, \sigma)\). The inequalities can be rearranged as in Section 7.2F, giving

\[
P(\bar{y} - t_{\alpha/2} \frac{s_{\bar{y}}}{\sqrt{n}} < \mu < \bar{y} + t_{\alpha/2} \frac{s_{\bar{y}}}{\sqrt{n}}) = 1 - \alpha.
\]

and it follows that \((\bar{y} \pm t_{a/2} s_{\bar{y}}/\sqrt{n})\) is a \(1 - \alpha\) confidence interval for \(\theta\).

**TABLE 8.4**
Comparison of 95 percent confidence intervals

<table>
<thead>
<tr>
<th>Variance Known</th>
<th>(n)</th>
<th>Variance Estimated</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\bar{y} \pm 1.96\sigma_0/\sqrt{10})</td>
<td>10</td>
<td>(\bar{y} \pm 2.26s_{\bar{y}}/\sqrt{10})</td>
</tr>
<tr>
<td>(\bar{y} \pm 1.96\sigma_0/\sqrt{20})</td>
<td>20</td>
<td>(\bar{y} \pm 2.09s_{\bar{y}}/\sqrt{20})</td>
</tr>
</tbody>
</table>

There is some interest in comparing the form of the confidence intervals for the two cases: variance known, and variance unknown. For \(n = 10\) and 20, the 95 percent confidence intervals for \(\mu\) are as shown in Table 8.4. In Section 7.2F we used the factor \(\pm 2\) with an estimated standard deviation to form an approximate large-sample confidence interval; and we recommended caution. We now see that even with normal variation and with a moderately large sample, there is need for a factor bigger than 2. From Example 6.1.2 we know that the Student \((n)\) distribution converges to the standard normal as \(n \to \infty\). Thus for large samples the Student limits for the confidence interval approximate the normal limits proposed in Section 7.2F.

**EXAMPLE 3**

Consider further the data and model in Example 2. A 95 percent confidence interval for the mean pH is

\[
(7.921 \pm 2.776 \frac{0.017}{\sqrt{5}}) = (7.900, 7.942).
\]

**D EXERCISES**

1. A strength measurement on steel castings has a normal(\(\theta, 2.5\)) distribution with \(\theta\) in \(R\). A random sample of 10 castings yields
Sec. 8.3: Statistical Inference for the Normal Model

59.5  61.5  63.5  63.0  64.5  
61.5  60.0  65.0  59.5  57.0.
(a) Test the hypothesis $\theta = 58.0$.
(b) Form 95 and 99 percent confidence intervals for $\theta$.

A method of measuring temperature remotely has a normal($\mu$, $\sigma$) distribution with ($\mu$, $\sigma$) in $\mathbb{R} \times \mathbb{R}^+$. The following seven determinations were made:

683, 688, 683, 687, 692, 687, 682.
(a) Test the hypothesis $\sigma = 2.0$.
(b) Form a 95 percent confidence interval for $\sigma$.

(continuation) (a) Test the hypothesis $\mu = 680$. (b) Form 95 and 99 percent confidence intervals for $\mu$.

4 The breaking strength for samples of a certain yarn is approximately normal. The following data were obtained:

(a) Calculate a 95 percent confidence interval for $\sigma$.
(b) Calculate a 95 percent confidence interval for $\mu$.

5 Determinations of viscosity for gasoline are approximately normal. For a particular type of gasoline the following results were obtained for 25 determinations: $\bar{y} = 36.48$, $s_y = 11.37$. Calculate a 99 percent confidence interval for the mean viscosity $\mu$.

6 The lifetime for certain light bulbs is approximately normal. A sample of 10 for a particular type yielded

$\bar{y} = 1495$,  
$s_y = 217$.
(a) Calculate a 95 percent confidence interval for $\sigma$.
(b) Calculate a 95 percent confidence interval for the mean lifetime $\mu$.

E TWO-SAMPLE PROBLEMS

7 For samples from the normal($\mu_1$, $\sigma$) and normal($\mu_2$, $\sigma$), the distribution of

$$
\frac{s^2}{\sigma^2} = \frac{\sum (y_{1i} - \bar{y}_1)^2 + \sum (y_{2i} - \bar{y}_2)^2}{\sigma^2} = \frac{(m + n - 2)s^2}{\sigma^2}
$$

is chi-square($m + n - 2$), as obtained in Problem 8.2.6.
(a) Describe the procedure for testing the hypothesis $\sigma^2 = \sigma_0^2$.
(b) Determine a $1 - \alpha$ confidence region for $\sigma^2$.

(continuation) The lifetime for certain light bulbs is approximately normal with variance independent of the type of such bulbs. Ten bulbs of type A and 10 of type B yielded the following data:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1028</td>
<td>1143</td>
</tr>
<tr>
<td>1</td>
<td>1270</td>
<td>1017</td>
</tr>
<tr>
<td>2</td>
<td>1094</td>
<td>1138</td>
</tr>
<tr>
<td>3</td>
<td>1486</td>
<td>1021</td>
</tr>
<tr>
<td>4</td>
<td>1643</td>
<td>1711</td>
</tr>
<tr>
<td>5</td>
<td>1340</td>
<td>1092</td>
</tr>
<tr>
<td>6</td>
<td>1497</td>
<td>1383</td>
</tr>
<tr>
<td>7</td>
<td>1614</td>
<td>1065</td>
</tr>
<tr>
<td>8</td>
<td>1380</td>
<td>1627</td>
</tr>
<tr>
<td>9</td>
<td>1293</td>
<td>1061</td>
</tr>
</tbody>
</table>

Calculate a 95 percent confidence interval for the common variance $\sigma^2$.

9 For samples from the normal($\mu_1$, $\sigma$) and normal($\mu_2$, $\sigma$), the distribution of

$$
t = \frac{\bar{y}_2 - \bar{y}_1 - (\mu_2 - \mu_1)}{s(1/m + 1/n)^{1/2}}
$$
is the Student\( (m + n - 2) \) distribution; see Problem 7.
(a) Use Problem 8.2.6 to justify the preceding distribution for \( t \).
(b) Describe the procedure for testing the hypothesis: \( \mu_1 = \mu_2 \).
(c) Determine a 1 - \( \alpha \) confidence interval for \( \mu_2 - \mu_1 \).

(continuation) Use the data and model from Problem 8 to determine a 95 percent confidence interval for the difference \( \mu_2 - \mu_1 \) in mean lifetimes.

Eleven pieces of material were sampled from a lot; five chosen at random were subjected to a first treatment, and the remaining six to a second treatment. The resulting response measurement yields

\[
\bar{y}_1 = 0.275, \quad \bar{y}_2 = 0.293,
\]

\[
s_1^2 = 0.00045, \quad s_2^2 = 0.00039.
\]

On the basis of the model in Problem 9,
(a) Test the hypothesis \( \mu_1 = \mu_2 \).
(b) Form a 95 percent confidence interval for \( \delta = \mu_2 - \mu_1 \).

For the model as given in Problem 8.2.9, we have that

\[
\frac{\sigma_1^2}{s^2} = \frac{s_1^2}{\sigma_2^2} = \frac{s_2^2}{\sigma_2^2}
\]

is the \( F \) distribution\( (m - 1, n - 1) \).
(a) Describe the procedure for testing the hypothesis \( \sigma_1^2 = \sigma_2^2 \).
(b) Determine a 1 - \( \alpha \) confidence interval for \( \gamma = \sigma_1^2 / \sigma_2^2 \).

(continuation) For the data in Problem 11, test the hypothesis \( \sigma_1^2 = \sigma_2^2 \).

8.4

AN IMPORTANT STATISTIC

For the normal(\( \theta, \sigma_0 \)) model we have seen (Example 8.1.1) that each sample point \( (y_1, \ldots, y_n) \) that has a particular value for the function \( \bar{y} \) gives the same likelihood function; accordingly, as we have noted, the model says the same thing about each of these sample points. As a result we restricted our attention for statistical inference to \( \bar{y} \) and its marginal model.

For this example we can see that a particular value for \( \bar{y} \) determines the particular likelihood function, and, indeed, a particular likelihood function in return determines the particular value for \( \bar{y} \). Thus \( \bar{y} \) indexes the possible likelihood functions. Such an essentially unique statistic that indexes the likelihood functions is called the likelihood statistic. The likelihood statistic is a fundamental part of the analysis of most statistical models; indeed, it guides the analysis in almost all areas of application.

Of course, for this example the function \( \bar{y} \) is the maximum likelihood estimate (MLE). For some more general problems, however, we may need to know more than the MLE estimate in order to determine the particular likelihood function. Thus we should not expect generally that a maximum likelihood estimate will serve as the likelihood statistic; recall Exercise 8.1.4.

A THE LIKELIHOOD STATISTIC

First, consider the normal model with known variance \( \sigma_0^2 \). The likelihood function \( \ell(\gamma \mid \theta) \) from a possible sample \( y = (y_1, \ldots, y_n) \) is given by
(1) \[ L(y | \theta) = c \exp \left\{ -\frac{n}{2\sigma_0^2} (\bar{y} - \theta)^2 \right\}; \]

the arbitrary constant \( c \) generates the equivalence class of similarly shaped functions of \( \theta \). We can eliminate the constant by choosing, say, the representative that is standardized with respect to maximum likelihood:

(2) \[ L_2(y | \theta) = \exp \left\{ -\frac{n}{2\sigma_0^2} (\bar{y} - \theta)^2 \right\}. \]

Note that the likelihood function \( L_2(y | \cdot) \) has normal shape, is scaled by \( \sigma_0 / \sqrt{n} \), and has maximum value at the point \( \theta = \bar{y} \).

Each sample point \((y_1, \ldots, y_n)\) that has \( \bar{y} \) equal to \( k \) produces a particular likelihood function located at \( k \), and each point that has \( \bar{y} \) equal to a different \( k' \) produces a particular likelihood function located at a different point \( k' \); see Figure 8.10. Thus the values for \( \bar{y} \) can be put in one-to-one correspondence with the possible likelihood functions; that is, \( \bar{y} \) indexes the possible likelihood functions. Any other function that indexes the likelihood functions is of course one-to-one equivalent to \( \bar{y} \): for example \( \Sigma_i^n y_i \), or \( \bar{y}^2 \), or \( \exp \{ \bar{y} \} \); recall Section 3.1D. Accordingly, we call \( \bar{y} \) the likelihood statistic for the normal model with known variance; and indeed we can call \( \bar{y}^2 \) the likelihood statistic.

Now consider a statistical model on \( \mathbb{R}^n \) with density function \( f(\cdot | \theta) \) with \( \theta \) in \( \Omega \); we assume some regularity, as discussed in Section 8.1A. The likelihood function from a sample point \( y \) is \( L(y | \cdot) \), where

(3) \[ L(y | \theta) = cf(y | \theta); \]

the arbitrary constant \( c \) generates the equivalence class of similarly shaped functions of \( \theta \).

We now consider a particular likelihood function and we identify the preimage set of points that produce that likelihood function; in a similar way, we consider some other likelihood function and we identify the preimage set of points that produce that likelihood function; see Figure 8.11. Suppose that we can find a function \( s \) on \( \mathbb{R}^n \) that indexes the likelihood functions; then the set of points that produce a particular likelihood function is the set of points for which \( s(y) \) takes some value \( k \).

**FIGURE 8.10**

Likelihood function from points having \( \bar{y} = k \); likelihood function from points having \( \bar{y} = k' \).
**Figure 8.11**
Likelihood function from points having $s(y) = k$; likelihood function from points having $s(y) = k'$.

**Definition 1**

The likelihood statistic is a function on the sample space that indexes the possible likelihood functions; the likelihood statistic is essentially unique up to a one-to-one equivalence.

In conclusion we remark that this concept can be examined from a very general point of view. Let $\mathcal{L}$ be the set of possible likelihood functions. We can view $L(y \mid \cdot)$ as a function that carries a point $y$ in $\mathbb{R}^n$ to a particular likelihood function in $\mathcal{L}$; call this function the likelihood map. The likelihood map itself is an example of the likelihood statistic; in the preceding paragraph, however, we had in mind a simple function such as $s(y) = \bar{y}$.

**B Examples of Likelihood Statistics**

We now determine the likelihood statistic for several examples.

**Example 1**

The location-scale normal: Let $(y_1, \ldots, y_n)$ be a sample from the normal$(\mu, \sigma)$ distribution with $(\mu, \sigma)$ in $\mathbb{R} \times \mathbb{R}^+$. From Example 8.1.3 we know that $(\bar{y}, s^2_y)$ indexes the likelihood functions. Thus $s(y) = (\bar{y}, s^2_y)$ is the likelihood statistic. Note some equivalent forms: $(\bar{y}, \Sigma (y_i - \bar{y})^2), (\Sigma y_i, \Sigma y_i^2)$.

**Example 2**

The Bernoulli: Let $(x_1, \ldots, x_n)$ be a sample from the Bernoulli$(p)$ with $p$ in [0, 1]. From Example 8.1.2 we know that $\Sigma x_i$ indexes the likelihood functions. Thus $s(x) = \Sigma x_i$ is the likelihood statistic. An equivalent statistic is the sample proportion $\hat{p} = \Sigma x_i/n$.

**Example 3**

The uniform: Let $(y_1, \ldots, y_n)$ be a sample from a uniform$(0, \theta)$ distribution with $\theta$ in (0, $\infty$). For notation, let $\phi$ be the indicator function for the interval (0, 1).
FIGURE 8.12
Likelihood function $f(y|\theta)$ from a sample point $y$.

Then for a single response we have

$$f(y|\theta) = \theta^{-1}\phi(y/\theta),$$

and for a sample of $n$ we have

$$f(y|\theta) = \theta^{-n} \prod_{i=1}^{n} \phi(y_i/\theta).$$

Thus the likelihood function from a sample point $y$ is

$$L(y|\theta) = c\theta^{-n} \prod_{i=1}^{n} \phi(y_i/\theta).$$

This function has the form $c\theta^{-n}$ whenever the product of the $\phi$'s is equal to 1, that is, whenever $\theta$ is greater than all the $y$'s or whenever $\theta$ is greater than max $y_i = y_{(n)}$; and it is zero for $\theta \leq y_{(n)}$. See Figure 8.12. Note that $y_{(n)}$ determines the likelihood function and the likelihood function determines $y_{(n)}$. Thus the function $s(y) = y_{(n)}$ indexes the likelihood functions; it is the likelihood statistic. It follows that $y_{(n)}$ is all that is needed from an observed response.

EXAMPLE 4

Let $\{y_1, \ldots, y_n\}$ be a sample from a distribution given by $f(y|\theta)$ with $\theta$ in $\{\theta_1, \theta_2\}$. The likelihood function can be expressed conveniently by using the representative

$$L_2(y) = \frac{\prod_{i=1}^{n} f(y_i|\theta_2)}{\prod_{i=1}^{n} f(y_i|\theta_1)},$$

mentioned in Section 8.1D. For given $y$ note that $L_2(y)$ is just a real number or $+\infty$ (nonzero numerator, zero denominator); it is called the likelihood ratio and it gives the ratio of the probability at $y$ from $\theta_2$ to that from $\theta_1$. Note that $L_2(y)$ as a function on the sample space $\mathbb{R}^n$ maps into $\mathbb{R}$ or into $\mathbb{R} \cup \{+\infty\}$. Thus $s(y) = L_2(y)$ indexes the likelihood functions; it is a real-valued likelihood statistic.
C THE LIKELIHOOD STATISTIC
WITH INDEPENDENCE

Now consider the likelihood statistic when independent models are combined. We see first that likelihood functions from independent systems combine in a very simple way.

Consider a statistical model for \( y_1 \) given by \( f'(y_1 | \theta) \) with \( \theta \) in \( \Omega \) and a statistical model for an independent \( y_2 \) given by \( f''(y_2 | \theta) \) with the same \( \theta \) in \( \Omega \). The statistical model for the combined \( (y_1, y_2) \) is then given by \( f'(y_1 | \theta)f''(y_2 | \theta) \) with \( \theta \) in \( \Omega \).

The likelihood functions from the components \( y_1 \) and \( y_2 \) are

\[ L'(y_1 | \theta) = cf'(y_1 | \theta), \quad L''(y_2 | \theta) = cf''(y_2 | \theta). \]

And the likelihood function from the combined \( (y_1, y_2) \) is

\[ L(y_1, y_2 | \theta) = cf'(y_1 | \theta)f''(y_2 | \theta) = L'(y_1 | \theta)L''(y_2 | \theta). \]

Thus when independent models are combined, the likelihood functions are multiplied. Correspondingly, the log-likelihoods or the scores are added:

\[ \log L(y_1, y_2 | \theta) = \log L'(y_1 | \theta) + \log L''(y_2 | \theta). \]

\[ S(y_1, y_2 | \theta) = S'(y_1 | \theta) + S''(y_2 | \theta). \]

Thus when two statistical models are combined, we have the simple formulas (6), (7), and (8) for combining the likelihood, log-likelihood, or score functions. Accordingly, we will almost always determine the likelihood statistic for a combined model directly from properties of the simply obtained likelihood function.

D THE LIKELIHOOD STATISTIC
IS SUFFICIENT

We have seen that the likelihood function presents all that the model has to say about an observed response. We have also seen that a range of sample points can have the same likelihood function—that the model is saying the same thing about each of the points in the range. Accordingly, we used \( \bar{y} \) with the normal(\( \mu, \sigma_0 \)) and we used \( (\bar{y}, s^2_0) \) with the normal(\( \mu, \sigma \))—on the grounds that the model does not distinguish among the points having a particular value for the likelihood statistic. And, of course, any implications concerning the true parameter value would be the same from each of the points in the range.

We now consider a distribution property of the likelihood statistic, a property that is somewhat weaker than the preceding result. Suppose that we have an investigator who has collected data from a system being investigated and has reported to us the value of the likelihood statistic. With this value in hand we are, of course, free to contemplate the possibilities for the antecedent data, that is, to contemplate the conditional distribution given the value of the likelihood statistic. This conditional distribution cannot involve the parameter \( \theta \), for if it did, the model would have to say different things about various possibilities for the antecedent data. We examine this more specifically.

Consider a statistical model on \( R^n \) with density function \( f(\cdot | \theta) \) with \( \theta \) in \( \Omega \); we assume some regularity, as discussed in Section 8.1A. In the preceding paragraph we
mentioned a conditional distribution being independent of a parameter; for this we introduce the following definition.

**DEFINITION 2**

A function $t$ on the sample space $\mathbb{R}^n$ is a **sufficient statistic** if the conditional distribution given $t$ does not depend on the parameter $\theta$.

As support for the term "sufficient," suppose that we have been given the value $t(y) = t'$ for the sufficient statistic in an application. Then it is reasonable to inquire about the antecedent data. The definition says that the distribution describing possible antecedent data does not involve $\theta$. This suggests that the antecedent data would be of no additional use—that the statistic $t$ is sufficient. Of course, $t$ can be vector-valued.

Consider briefly an example.

**EXAMPLE 5**

Let $(y_1, y_2)$ be a sample of $n = 2$ from the normal$(\theta, \sigma_0)$ with $\theta$ in $\mathbb{R}$. From Section 8.2C we know that

$$u_1 = \frac{y_1 + y_2}{\sqrt{2}}, \quad u_2 = \frac{y_1 - y_2}{\sqrt{2}}$$

are statistically independent, that $u_1$ is normal$(\sqrt{2}\theta, \sigma_0)$, and that $u_2$ is normal$(0, \sigma_0)$. We examine the conditional distribution given $u_1$; the conditional distribution is described by $u_2$ and is normal$(0, \sigma_0)$; it does not involve $\theta$. Thus $u_1$ is sufficient. Of course, we note that $u_1 = \sqrt{2} \bar{y}$ is the likelihood statistic.

Our introductory remarks can now be formalized as the following theorem.

**THEOREM 3**

The likelihood statistic is sufficient.

**Proof**  Our consideration of conditional distributions in Section 4.2 was limited to cases in which we could find a complementing function $u$ such that there is a one-to-one correspondence, $y \leftrightarrow (s(y), u(y))$; and in the absolutely continuous case, we require a continuous Jacobian each way. We record the proof for this absolutely continuous case; the discrete case is a direct analog without the Jacobians.

The likelihood function $L(y \mid \cdot)$ is determined by the likelihood statistic $s(y)$. Let $L^*(s \mid \cdot)$ be a particular representative in the class of similarly shaped likelihood functions determined by $s(y) = s$. Note that $f(y \mid \theta)$ and $L^*(s(y) \mid \theta)$ are similarly shaped functions of $\theta$; let $a(y)$ give the proportionality constant,

$$f(y \mid \theta) = a(y)L^*(s(y) \mid \theta).$$

Thus the probability differential for $y$ can be written

$$f(y \mid \theta) \, dy = a(y)L^*(s(y) \mid \theta) \, dy.$$
The joint distribution for \((s, u)\) can be obtained from Section 2.4E. Let
\[ J(s, u) = \left| \frac{\partial y}{\partial (s, u)} \right|_+ \]
be the Jacobian determinant. Then substitution in (10) gives the probability differential for \((s, u)\),
\[ (11) \quad a(y)L^*(s|\theta)J(s, u) \, ds \, du. \]

The marginal distribution for \(s\) can be obtained by integrating out the variable \(u\). The variable \(u\) occurs only in the factors \(a(y)\) and \(J(s, u)\); hence we obtain the marginal differential
\[ (12) \quad L^*(s|\theta)k(s) \, ds, \]
where \(k\) is the obvious integral involving \(a(y)\) and \(J(s, u)\).

The conditional distribution for \(u\) given \(s\) can be obtained by dividing (11) by (12); we obtain
\[ (13) \quad \frac{a(y)J(s, u)}{k(s)} \, du. \]

Note that this conditional distribution does not depend on \(\theta\); thus \(s\) is a sufficient statistic.

The preceding proof has an immediate corollary:

**COROLLARY 4**

*The likelihood function obtained from an observed \(y\) is the same as the likelihood function obtained from the likelihood statistic \(s(y)\) and the marginal model for \(s\).*

**Proof** The likelihood function from \(s(y)\) and its marginal model is available from (12):
\[ (14) \quad cL^*(s(y)|\theta)k(s(y)) = cL^*(s(y)|\theta). \]

But this is just the likelihood function from \(y\) as produced by (9).

**E THE LIKELIHOOD STATISTIC IS MINIMAL SUFFICIENT**

A sufficient statistic has an attractive property: it provides a reduction or simplification on the original response, and yet it records all that seems necessary concerning the original response. Clearly we would be interested in a sufficient statistic that makes the largest reduction — records as little as possible concerning the original response and yet records all that is necessary. For this we present the following definition:

**DEFINITION 5**

A sufficient statistic \(s\) is **minimal** if for any other sufficient statistic \(t\) we can find a function \(h\) so that \(s(y) = h(t(y))\).

The definition says that a minimal sufficient statistic is a reduction on any other sufficient statistic. We now have the following property for the likelihood statistic.
THEOREM 6

The likelihood statistic is the minimal sufficient statistic.

Proof Consider a sufficient statistic \( t \). We confine our proof as with Theorem 3 to the case where we have a complementing function \( u \) such that there is a one-to-one correspondence \( y \leftrightarrow (t(y), u(y)) \); and in the absolutely continuous case we require a continuous Jacobian each way. We then obtain

\[
(15) \quad f(y | \theta) \, dy = f(y | \theta) \, J(t, u) \, dt \, du = f_1(u : t | \theta) \, f_2(t | \theta) \, dt \, du,
\]

where the middle expression comes by a change of variable and the last expression is an ordinary conditional-times-marginal expression in which the conditional density \( f_1(u : t) \), by sufficiency, does not depend on \( \theta \). The likelihood function \( L(y | \cdot) \) can now be calculated and we have

\[
(16) \quad L(y | \theta) = cf(y | \theta) = cf_2(t(y) | \theta),
\]

where we have combined \( f_1(u : t) \) with the multiplicative constant \( c \). Note that the likelihood function can be calculated from \( t \) alone, and thus the likelihood statistic can be expressed as a function of \( t \).

The theorem states in effect that the likelihood statistic is the best sufficient statistic. The concept of a sufficient statistic has been widely used in statistical theory. The concept of the likelihood statistic, however, has recently emerged as a more appropriate concept: The likelihood statistic has the strong justification based on the likelihood function in Section 8.1, a justification that is substantially stronger than the distribution property of a sufficient statistic; the likelihood function itself admits direct interpretation and immediate usefulness in statistical inference; and, besides, the likelihood statistic is the fastest route to the minimal sufficient statistic.

F EXERCISES

1. **Poisson distribution:** Let \( (y_1, \ldots, y_n) \) be a sample from the Poisson(\( \lambda \)) distribution with \( \lambda \) in \((0, \infty)\). Show that the likelihood statistic can be given by \( \Sigma y_i \) or by \( \bar{y} \).

2. **Binomial distribution:** Let \( (x_1, \ldots, x_n) \) be a sample from the Bernoulli(\( p \)) distribution with \( p \) in \([0, 1]\). Show that the likelihood statistic can be given by \( \Sigma x_i \) or by \( \bar{x} \).

3. **Scale normal:** Let \( (y_1, \ldots, y_n) \) be a sample from the normal(\( \mu \), \( \sigma^2 \)) distribution with \( \sigma^2 \) in \((0, \infty)\). Show that the likelihood statistic is \( \Sigma (y_i - \mu)^2 \).

4. **Uniform(\( \theta, \theta + 1 \)):** Let \( (y_1, \ldots, y_n) \) be a sample from the uniform(\( \theta, \theta + 1 \)) distribution with \( \theta \) in \( \mathbb{R} \). Show that the likelihood statistic is \( \Sigma y_i \).

5. **Scale-exponential:** Let \( (y_1, \ldots, y_n) \) be a sample from the exponential(\( \theta \)) with \( \theta \) in \((0, \infty)\). Show that the likelihood statistic is \( \Sigma y_i \).

6. **Scale gamma or chi-square:** Let \( (y_1, \ldots, y_n) \) be a sample from the scale gamma having density \( f(y | \theta) = \Gamma^{-1}(p; \theta^2 y^{-p-1}) \exp (-y/\theta) \) on \((0, \infty)\) and 0 otherwise, with \( \theta \) in \((0, \infty)\). Show that the likelihood statistic is \( \Sigma y_i \).

7. **Pareto:** Let \( (y_1, \ldots, y_n) \) be a sample from the Pareto distribution \( f(y | \theta) = \theta(1 + y)^{-\theta-1} \) on \((0, \infty)\) and 0 otherwise, with \( \theta \) in \((0, \infty)\). Show that the likelihood statistic is \( \Pi (1 + y_i) \).

8. **Order statistic:** Let \( (y_1, \ldots, y_n) \) be a sample of \( n \) from an arbitrary continuous density on \( \mathbb{R} \). Determine the conditional distribution of \( (y_1, \ldots, y_n) \) given \( (y_1(n), \ldots, y_n) \); and deduce that the order statistic \( (y_1(n), \ldots, y_n) \) is sufficient; see Section 3.6E.
G PROBLEMS

9 Negative binomial distribution: Let \((x_1, x_2, \ldots)\) be a sample sequence from the Bernoulli(\(p\)) distribution with \(p\) in \((0, 1)\), and suppose the sequence is terminated when \(\sum_{i=1}^{p} x_i\) first equals a given \(k\). Show that the likelihood statistic is given by \(n\).

10 Hypergeometric: Let \((x_1, \ldots, x_n)\) be a random sample from a finite population of \(D\) 0's and \(N - D\) 1's with \(D\) in \([0,1,\ldots, N]\). Show that the likelihood statistic is \(\sum x_i\) (the hypergeometric variable).

11 Multinomial: Let \((y_1, \ldots, y_n)\) be a sample from the multinomial \((n, p_1, \ldots, p_k)\) with \(\sum p_i = 1, p_i > 0\). Show that the likelihood statistic is \(\sum y_i\).

12 Multivariate hypergeometric: Let \((x_1, \ldots, x_n)\) be a random sample from a finite population containing \(Np_1\) elements \(b_1, \ldots, Np_k\) elements \(b_k\), where the possible \(p_1, \ldots, p_k\) satisfy \(\sum p_i = 1, p_i > 0\). Show that the likelihood statistic is \((t_1, \ldots, t_k)\), where \(t_j\) is the number of \(b_j\) in \((x_1, \ldots, x_n)\).

13 Variable carrier: Let \((y_1, \ldots, y_n)\) be a sample from \(f(y|\theta) = k(\theta)c(y - \theta)h(y)\), where \(c\) is the positive indicator function defined in Problem 7.4.8. Show that the likelihood statistic is \(y_{(n)}\).

14 Uniform(\(\theta_1, \theta_2\)): Let \((y_1, \ldots, y_n)\) be a sample from the uniform distribution \((\theta_1, \theta_2)\) with \(\theta_1 < \theta_2\) and \(\theta_1, \theta_2\) in \(R\). Show that the likelihood statistic is \((y_{(1)}, y_{(n)})\).

15 Location-scale exponential: Let \((y_1, \ldots, y_n)\) be a sample from \(f(y|\theta, \tau) = \tau^{-1} \exp \left(-\frac{y - \theta}{\tau}\right)\) on \((\theta, \infty)\) and 0 otherwise, with \((\theta, \tau)\) in \(R \times (0, \infty)\). Show that the likelihood statistic is \((y_{(1)}, \Sigma y_i)\).

16 A drug is administered to \(n\) animals at each of \(r\) doses \(x_1, \ldots, x_r\). The ‘‘logistic’’ model is sometimes used for the probability \(p_i\) of reaction at dose \(x_i\).

\[
p_i = \frac{\alpha \beta x_i}{1 + \alpha \beta x_i}
\]

with parameters \(\alpha\) and \(\beta\). Let \(y_1, \ldots, y_r\) be the number of reactions, respectively, at doses \(x_1, \ldots, x_r\). Show that the likelihood statistic is \((\Sigma y_i, \Sigma x_i y_i)\).

17 Let \(p_i\) be the probability of exactly \(i\) female children in a family having 3 children, assuming independence of successive births, and let \(\theta\) be the probability of a female child. For a sample of \(n\) families having 3 children, let \(y_i\) be the number of families having \(i\) female children. Show that \(y_1 + 2y_2 + 3y_3\) is the likelihood statistic.

18 (a) Let \((x_1, \ldots, x_n)\) be a sample from the Bernoulli(\(p\)) distribution and write \(y = \Sigma x_i\). Determine the likelihood function.

(b) Let \(x_1, x_2, \ldots\) be statistically independent with the Bernoulli(\(p\)) distribution. Let \(n\) be the coordinate number at which \(\Sigma x_i = y\), where \(y\) is specified. Determine the likelihood function.

(c) If the values of \((n, y)\) in (a) and (b) are the same, show that the likelihood functions are the same.

19 Consider a population containing \(D\) 1's and \(N - D\) 0's with parameter \(D\) in \(\Omega = \{0, 1, \ldots, N\}\).

(a) Let \(y\) be the number of 1's in a random sample of \(n\). Determine the likelihood function from an observed \(y\).

(b) Suppose that elements are sampled randomly until a specified number \(y\) of 1's has been obtained. Let \(n\) be the number of elements sampled. Determine the likelihood function from an observed \(n\).

(c) If the values of \((n, y)\) in (a) and (b) are the same, show that the likelihood functions are the same.

20 Suppose that the parameter space \(\Omega\) is a connected open set in \(R^k\). Then show that the score function \((S_1(y|\theta), \ldots, S_k(y|\theta))\) on \(\Omega\) can be calculated from the log-likelihood, and conversely; the score components are

\[
S_i(y|\theta) = \frac{\partial}{\partial \theta_i} \ln f(y|\theta).
\]
21 From formula (16) we know that the likelihood function from an observed value of a sufficient statistic as calculated from the marginal model is the same as the likelihood function from the antecedent response as calculated from the full model. Let \( y_1, \ldots, y_n \) be a sample from a distribution with density \( f(y | \theta) \) with \( \theta \) in \( \{ \theta_1, \theta_2 \} \) and let \( L_2(y) \) be the likelihood ratio given in formula (5). Then show that the likelihood function (third form) from the observed \( L_2(y) \) is obtained from \( L_3 \) by the identity function.

22 Let \( T_1, \ldots, T_n \) be a sample from the canonical Student(\( \lambda \)) distribution with parameter \( \lambda \) in \((0, \infty)\). Determine the likelihood statistic.

23 Let \( G_1, \ldots, G_n \) be a sample from the canonical \( F(\lambda_1, \lambda_2) \) distribution with parameter \( \lambda_1, \lambda_2 \) in \((0, \infty) \times (0, \infty)\). Determine the likelihood statistic.

8.5 MODELS THAT HAVE EXPONENTIAL FORM

We have seen that, for a sample \( y_1, \ldots, y_{100} \) of 100 from the normal(\( \mu, \sigma \)) model, it suffices to record the value of \( (\bar{y}, s_y^2) \) for purposes of inference. This is a huge reduction—from a point in \( \mathbb{R}^{100} \) to a point in \( \mathbb{R}^2 \). In this section we investigate what kind of functional form for the model makes such reductions possible—in short, what makes the normal work.

Consider this normal example in detail. The density for a single \( y \) from the normal(\( \mu, \sigma \)) can be expressed in the form

\[
(1) \quad f(y | \mu, \sigma) = (2\pi\sigma^2)^{-1/2} \exp \left( -\frac{(y - \mu)^2}{2\sigma^2} \right) \cdot \exp \left( y \frac{\mu}{\sigma^2} - y^2 \frac{1}{2\sigma^2} \right).
\]

Note the simple form in the right-hand exponent—a linear combination of statistics and parameters. We obtain the density for a sample of \( n \) by adding the exponents:

\[
(2) \quad f(y | \mu, \sigma) = (2\pi\sigma^2)^{-n/2} \exp \left( -\frac{n(y - \mu)^2}{2\sigma^2} \right) \cdot \exp \left( \sum y_i \frac{\mu}{\sigma^2} - \sum y_i^2 \frac{1}{2\sigma^2} \right).
\]

Recall that \( (\Sigma y_i, \Sigma y_i^2) \) is the likelihood statistic.

Consider the Bernoulli(\( \rho \)) model. The density for a single \( x \) can be expressed in the form

\[
(3) \quad f(x | \rho) = \rho^x q^{1-x} = q \cdot \exp \left( x(\ln \rho - \ln q) \right).
\]

Note the simple form in the exponent—a linear combination of a statistic and a parameter. We obtain the density for a sample of \( n \) by adding the exponents:

\[
(4) \quad f(x | \rho) = \rho^{\Sigma x_i} q^{n-\Sigma x_i} = q^n \cdot \exp \left( \sum x_i(\ln \rho - \ln q) \right).
\]

Recall that \( \Sigma x_i \) is the likelihood statistic.

In this section we investigate statistical models that have this linear form in the exponent, and we see how this form relates to having a simple likelihood statistic regardless of the sample size.

A EXPONENTIAL FORM

The linear properties in the exponent lead to the following definition of exponential form.
DEFINITION 1

A statistical model \( \{f(y|\theta) : \theta \in \Omega \} \) has exponential form if the density function can be expressed as

\[
(5) \quad f(y|\theta) = \gamma(\theta) \exp \left( a_1(y)\psi_1(\theta) + \cdots + a_r(y)\psi_r(\theta) \right) h(y).
\]

Note in particular that the range of the distribution does not depend on \( \theta \).

It is quite possible for such a model to be written with more \( \psi \)-functions than are really needed. For this the following definition is convenient:

DEFINITION 2

A model having exponential form is in a reduced form if the functions \( 1, \psi_1(\theta), \ldots, \psi_r(\theta) \) (on \( \Omega \)) are linearly independent.

For the normal \((\mu, \sigma)\) example, it is easy to see that \( 1, \mu/\sigma^2, \) and \( 1/2\sigma^2 \) are linearly independent functions on \( R \times R^+ \). And for the Bernoulli \((p)\) example, it is immediate that \( 1 \) and \( \ln p - \ln q \) are linearly independent.

A model having exponential form can always be put in a reduced form. For if the \( \psi \)-functions are linearly dependent, we can solve for one in terms of the others, we can substitute in the density expression, and we can simplify, thus obtaining exponential form with one less \( \psi \)-function. This is repeated until a reduced form is obtained.

B THE LIKELIHOOD STATISTIC FOR
A MODEL HAVING EXPONENTIAL FORM

The likelihood statistic is available immediately for a model having reduced exponential form.

THEOREM 3

If an exponential model (5) is in a reduced form, then \( (a_1(y), \ldots, a_r(y)) \) is the likelihood statistic.

**Proof** Consider an exponential model (5) in a reduced form. The log-likelihood \( l(y|\cdot) \) from a sample point \( y \) has the form

\[
(6) \quad l(y|\theta) = \ln \gamma(\theta) + a \cdot 1 + a_1(y)\psi_1(\theta) + \cdots + a_r(y)\psi_r(\theta).
\]

Clearly the value of \( (a_1(y), \ldots, a_r(y)) \) determines the likelihood function; see Figure 8.13.

Now suppose that the values \( (a_1', \ldots, a_r') \) and \( (a_1'', \ldots, a_r'') \) for \( (a_1(y), \ldots, a_r(y)) \) both produce the same likelihood function. Then

\[
\ln \gamma(\theta) + a' \cdot 1 + a'_1\psi_1(\theta) + \cdots + a'_r\psi_r(\theta)
\]

and hence

\[
a \cdot 1 + a_1\psi_1(\theta) + \cdots + a_r\psi_r(\theta) = 0,
\]

and hence

\[
a \cdot 1 + a_1\psi_1(\theta) + \cdots + a_r\psi_r(\theta) = 0,
\]
FIGURE 8.13
Adjusted log-likelihood function \( l(y|\cdot) = \ln \gamma(\cdot) = a_1 + a_1 \psi_1(\cdot) + a_2 \psi_2(\cdot) \) has coordinates \( a_1 \) and \( a_2 \) with respect to the functions \( \psi_1 \) and \( \psi_2 \); the arbitrary value for \( a \) produces the line as shown.

where \( a = a' - a'' \) and \( a_1 = a'_1 - a''_1 \). The linear independence then says that the \( a_1 = 0 \) and hence that
\[
(a'_1, \ldots, a'_r) = (a''_1, \ldots, a''_r).
\]
Thus the likelihood function determines the value of \( (a_1(y), \ldots, a_r(y)) \). The two results show that \( (a_1(y), \ldots, a_r(y)) \) indexes the likelihood functions and is thus the likelihood statistic.

For the normal\((\mu, \sigma)\) example with density function given by (2), this theorem states that \( (\Sigma y_i, \Sigma y_i^2) \) is the likelihood statistic. Note that for a sample of \( n = 1 \), this says that \( (y, y^2) \) is the likelihood statistic; of course, this is equivalent to \( y \) itself, since \( y^2 \) is redundant.

For the Bernoulli\((p)\) example with density function given by (4), this theorem states that \( \Sigma x_i \) is the likelihood statistic; this is, of course, equivalent to the proportion or estimated probability \( \hat{p} = \Sigma x_i / n \).

C  SAMPLING FROM A MODEL THAT HAS EXPONENTIAL FORM

A model that has exponential form has some very convenient properties under sampling. Consider a model for \( y \) that has exponential form with density
\[
(7) \quad f(y|\theta) = \gamma(\theta) \exp \left( a_1(y)\psi_1(\theta) + \cdots + a_r(y)\psi_r(\theta) \right) h(y).
\]
Then the model for a sample \( (y_1, \ldots, y_n) \) has the density function
\[
(8) \quad f(y|\theta) = \gamma^*(\theta) \exp \left( \sum a_1(y_i)\psi_1(\theta) + \cdots + \sum a_r(y_i)\psi_r(\theta) \right) \prod h(y_i).
\]
But this model also has exponential form and uses the same functions \( \psi_1(\theta), \ldots, \psi_r(\theta) \).

Now suppose that the model (7) or (8) is in reduced exponential form. Then by
Theorem 3 it follows that \( \left( \sum a_1(y_i), \ldots, \sum a_r(y_i) \right) \) is the likelihood statistic for the sample \( (y_1, \ldots, y_n) \). This establishes the following theorem:

**THEOREM 4**

If \( (y_1, \ldots, y_n) \) is a sample from a model having reduced exponential form, then \( \left( \sum a_1(y_i), \ldots, \sum a_r(y_i) \right) \) is the likelihood statistic.

Consider a simple example.

**EXAMPLE 1**

Let \( (y_1, \ldots, y_n) \) be a sample from a normal\((\mu, \sigma)\) distribution with \( \sigma \) in \( R^+ \). The density function for a single response \( y \) can be written in the form

\[
f(y|\sigma) = (2\pi\sigma^2)^{-1/2} \exp \left\{ -\frac{(y - \mu)^2}{2\sigma^2} \right\}.
\]

This has reduced exponential form with one \( \psi \)-function \( 1/2\sigma^2 \). Thus for a sample \( (y_1, \ldots, y_n) \) it follows from Theorem 4 that \( \sum a(y_i - \mu)_2 \) is the likelihood statistic; compare with Problem 8.4.3.

**D WHEN DOES THE LIKELIHOOD STATISTIC HAVE FIXED DIMENSION?**

For a sample from the normal\((\mu, \sigma)\) model which has two \( \psi \)-functions, we have seen that the likelihood statistic is \( \left( \sum y_i, \sum y_i^2 \right) \), which has just two dimensions regardless of how large the sample size is.

We have seen this property with another kind of model. Let \( (y_1, \ldots, y_n) \) be a sample from the uniform\((0, \theta)\) distribution. Then by Example 8.4.3 the function \( y_{(n)} \) is the likelihood statistic. This function has a single coordinate regardless of how large the sample size is. But note that this model exemplifies a rather different kind of model—a model in which the region of positive density varies with the parameter, a variable carrier model. However, such models are appropriate only in very special circumstances.

We now restrict our attention to statistical models that have a fixed region of positive density; that is, the limits for the distribution do not vary with \( \theta \). Let \( f(y|\theta) \) be the density function relative to the length measure on \( R \), and suppose that \( \ln f(y|\theta) \) is continuously differentiable with respect to \( y \) on \( R \); note that this implies that the logarithm exists and that \( f \) is nonzero on \( R \). We now see that fixed dimension for the likelihood statistic occurs only for a model having exponential form:

**THEOREM 5**

If \( \{f(y|\theta) : \theta \in \Omega\} \) is a statistical model with the regularity conditions in the preceding paragraph, then the likelihood statistic has maximum dimension \( r \) under sampling if and only if the model is exponential with \( r \) \( \psi \)-functions in a minimal reduced form.
The proof is recorded in Section E. Minimal means smallest value for \( r \).

In Chapters 9 and 10 we will see that a large part of statistical theory works only if there is a likelihood statistic with one, two, or several coordinates—but not more than the number of parameter coordinates. By Theorem 5 this property occurs only for models that have exponential form. We of course then want to know what common statistical models have exponential form.

For continuous models we have seen that the \( \text{normal}(\mu_0, \sigma) \) model and the \( \text{normal}(\mu, \sigma) \) model have the exponential form just described. For discrete models we have seen that the \( \text{binomial}(\rho) \) model and the \( \text{Poisson}(\theta) \) model have the exponential form. All the frequency-count models, however, are basically simple and hardly need the somewhat elaborate concepts that we have been discussing in this chapter.

This brings us back to asking what common continuous models have exponential form—other than the normal models we have just mentioned. A common statistical model is concerned with the general level of the response in the case where the distribution for variation is known; the response model is

\[
\{ f(y - \theta) : \theta \in R \},
\]

with \( f \) known. It can be shown that such a model has exponential form with one \( \psi \)-function if and only if (a) \( y \) has a normal distribution \( (\equiv a + cz) \), or (b) \( y \) has the distribution form of the logarithm of a \( \chi^2 \) variable \( (\equiv a + c \ln \chi^2) \). In effect, these are just the two normal examples that have a single parameter: the \( \text{normal}(\mu, \sigma) \) with location \( \mu \); and a logarithmic transform of the \( \text{normal}(\mu_0, \sigma) \) with location \( \ln \sigma \) (this latter case does, of course, include the scale exponential and scale gamma).

An even more general statistical model is concerned with the general level and the scaling of the response in cases where the form of the variation is known; the response model is

\[
\left\{ \frac{1}{\sigma} f \left( \frac{y - \mu}{\sigma} \right) : (\mu, \sigma) \in R \times R^+ \right\},
\]

with \( f \) known. It can be shown that such a model has exponential form with two \( \psi \)-functions if and only if \( y \) has a normal distribution. This is the third of the three normal examples: the \( \text{normal}(\mu, \sigma) \) with location \( \mu \) and scale \( \sigma \).

For other common statistical models we need some linear structure that is described in Chapter 9. For these models, also, it can be shown that exponential form occurs if and only if the distribution is normal.

It thus follows that a large part of the theory in the next two chapters works only for the three familiar normal examples, for the common frequency-count examples, and certain variable carrier examples.

**Supplement: Proof of Theorem 5**

Consider a density function \( f(y | \theta) \) satisfying the regularity conditions preceding Theorem 5. The log-likelihood function for a single response of \( y \) can be presented as

\[
\ell(y | \theta) = \ln f(y | \theta) - \ln f(y | \theta),
\]

where we are using the form \( I \) given by formula (8.1.13). The log-likelihood then for a sample \( (y_1, \ldots, y_n) \) is

\[
I_n(y | \theta) = \ell(y_1 | \theta) + \cdots + \ell(y_n | \theta).
\]
We examine the log-likelihood function for points in the neighborhood of a point \( y \):

\[
(12) \quad l_n(y + t | \theta) = l_n(y | \theta) + \sum_{i=1}^{n} t_i \frac{\partial}{\partial y_i} l(y_i | \theta) + o(t),
\]

where \( o(t) \) is a function that goes to zero as a proportion of \(|t|\); note that the derivative \( \partial / \partial y_i \) applied to (11) catches only the term \( l(y_i | \theta) \) on the right side. See Figure 8.14.

Now suppose that the likelihood statistic has maximum dimension \( r \) and that \( y \) is a point at which this dimension \( r \) is attained. The possible changes in the log-likelihood function as determined near \( y \) are given by

\[
(13) \quad \sum_{i=1}^{n} t_i \frac{\partial}{\partial y_i} l(y_i | \cdot),
\]

a linear combination of \( n \) functions. But the likelihood statistic has maximum dimension \( r \); hence the functions produced by formula (13) must have a basis consisting of just \( r \) functions, say \( \psi_1(\cdot), \ldots, \psi_r(\cdot) \); accordingly, all the functions (13) can be generated as linear combinations of these \( r \) \( \psi \)'s.

Now consider a sample \( (y_1, \ldots, y_n, y) \) of \( n + 1 \) let \( (y_1, \ldots, y_n) \) be the point described above and let \( y \) be arbitrary. Then, as above, the changes in the log-likelihood function as determined near this new point are given by

\[
\sum_{i=1}^{n} t_i \frac{\partial}{\partial y_i} l(y_i | \cdot) + t \frac{\partial}{\partial y} l(y | \cdot).
\]

But the likelihood statistic cannot have dimension greater than \( r \). It has dimension \( r \) on the basis of just the first terms; hence the function

\[
\frac{\partial}{\partial y} l(y | \cdot)
\]

the set of possible log-likelihood functions

\[
\frac{\partial}{\partial y_1} l(y_1 | \cdot) + t_2 \frac{\partial}{\partial y_2} l(y_2 | \cdot)
\]

FIGURE 8.14
Set of possible log-likelihood functions in the neighborhood of a particular likelihood function \( l_n(y | \cdot) \). The tangent plane at the particular likelihood function has coordinates \( t_1 \) and \( t_2 \) with respect to \( \partial l(y_1 | \cdot)/\partial y_1 \) and \( \partial l(y_2 | \cdot)/\partial y_2 \).
must be one of those generated by the \( r \psi \)-functions.

The preceding paragraph shows that

\[
\frac{\partial}{\partial \theta} l(y | \theta) = A_1(y) \psi_1(\theta) + \cdots + A_r(y) \psi_r(\theta).
\]

Integration then gives

\[
\ln \frac{f(y | \theta)}{f(y | \theta, y)} = l(y | \theta) = a_1(y) \psi_1(\theta) + \cdots + a_r(y) \psi_r(\theta) + \phi(\theta).
\]

Hence \( f(y | \theta) \) has exponential form based on \( r \psi \)-functions.

The converse follows from Theorem 3 applied to a reduced exponential model with minimum value for \( r \).

---

**F Exercises**

1. **Binomial distribution:** Let \( x_1, \ldots, x_n \) be a sample from the Bernoulli(\( p \)) distribution. Use Theorem 4 to determine the likelihood statistic.

2. **Poisson distribution:** Let \( y_1, \ldots, y_n \) be a sample from the Poisson(\( \theta \)) distribution. Use Theorem 4 to determine the likelihood statistic.

3. **Scale gamma:** Let \( y_1, \ldots, y_n \) be a sample from the scale gamma in Exercise 8.4.6. Use Theorem 4 to determine the likelihood statistic.

4. **Pareto:** Show that the Pareto distribution in Exercise 8.4.7 has exponential form. Use Theorem 4 to determine the likelihood statistic.

---

**G Problems**

5. **Multinomial:** Show that the multinomial\((n, p_1, \ldots, p_k)\) distribution with \( \sum p_i = 1 \) and \( p_i \geq 0 \) has exponential form. Determine the \( \psi \)-functions and for a sample verify the likelihood statistic in Problem 8.4.11.

6. **Beta:** Let \( u_1, \ldots, u_n \) be a sample from the beta\((\alpha, \beta)\) distribution with \( (\alpha, \beta)\) in \( R^+ \times R^+ \). Show that the model has exponential form. Determine the \( \psi \)-functions and record a reasonable likelihood statistic.

7. **Normal linear model:** Consider a response vector \( y \) in \( R^n \) having a distribution with mean located in an \( r \)-dimensional subspace \( X(x_1, \ldots, x_r) \) and with variation \( \sigma z \), where \( z \) is a sample from the standard normal; that is, \( y = X\beta + \sigma z \), where \( X = (x_1, \ldots, x_r) \) is an \( n \times r \) matrix of rank \( r \) and the parameter \((\beta, \sigma)\) is in \( R^r \times R^+ \). Show that the response density function is

\[
 f(y | \beta, \sigma) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta) \right\}
\]

and that it has exponential form. Determine a set of \( \psi \)-functions and show that \( (y'y, y'X, \ldots, y'x_j) = (y'y, y'X) \) is a likelihood statistic.

8. **Multivariate normal:** Consider the multivariate normal \((\mu, \Sigma)\) with \( \mu \) in \( R^k \) and with essential coordinates of \( \Sigma \) in a region of \( R^{k(k+1)/2} \) corresponding to positive definite matrices.

(a) Show that the model has exponential form. Determine the \( \psi \)-functions.

(b) Let \( y_1, \ldots, y_n \) be a sample of \( n \) from the multivariate normal\((\mu, \Sigma)\) as described. Show that \((\Sigma y_1, \Sigma y_1 y_1)'\) is the likelihood statistic; it has essentially \( k + k(k + 1)/2 \) coordinates.
8.6
INFERENCES FROM LARGE SAMPLES
USING LIKELIHOOD

We have seen with a parametric model that the likelihood function from an
observed response presents all that the model has to say about the observed response.
In Section 8.1C we discussed some inference methods based on a direct assessment of
the observed likelihood function.

We have also seen how to group together sample points that produce the same
likelihood function. This led us in Section 8.4 to the definition of the likelihood
statistic, a function that gives an important reduction on the sample space.

In this section we examine these two concepts together: for any contemplated
parameter value we consider the distribution of possible likelihood functions obtained
from samples. This statistical model for possible likelihood functions is central to our
present examination of inference for parametric models. Recall the observed likelihood
function for the electrolytic crucibles in Section 8.1B; it is natural to contemplate the
likelihood functions that we might have obtained given any particular distribution for
 crucible life.

For a sample from the normal(θ, σ₀) model, the distribution for possible likelihood
functions is easily derived. We examine this in detail in Section A.

For a sample from some other model, the distribution for possible likelihood
functions may be complicated to derive or awkward to express. Fortunately, however,
for large samples the distribution is easily obtained and is almost a copy of that for the
normal(θ, σ₀) model. We examine this in Sections B and C.

We then examine some large-sample inference methods in Sections D and E. Our
large-sample inference methods in Section 7.2 were developed for the mean of an
essentially arbitrary distribution. The present methods provide the essential inference
procedures for the general parameter of a parametric model. Indeed, they provide
some basic procedures that work when other methods fail.

A  THE NORMAL WITH KNOWN VARIANCE

Let (y₁, . . . , yₙ) be a sample from a normal(θ, σ₀) distribution with θ in R. The
likelihood, log-likelihood, and score functions are given by

\[ L_n(y | \theta) = c \exp \left( -\frac{1}{2\sigma^2_0/n} (\bar{y} - \theta)^2 \right), \]

\[ l_n(y | \theta) = -\frac{1}{2\sigma^2_0/n} (\bar{y} - \theta)^2 + a. \]

\[ s_n(y | \theta) = \frac{\bar{y} - \theta}{\sigma^2_0/n}; \]

the subscript n is used to indicate dependence on the sample size.

We are now going to consider the distribution of possible likelihood functions. For
this we face a small notational problem. On the one hand, we have used \( \theta \) for a
parameter value that gives the distribution for the sample \( (y_1, . . . , y_n) \). And on the
other hand, we have used \( \theta \) as a free variable when examining the likelihood function
from an observed response. Now that we are going to examine the distribution of possible likelihood functions, we must face these two roles for \( \theta \) simultaneously. Accordingly, we will adopt distinguishing notations where needed. For the present section we will use \( \theta \) for the free variable when examining the likelihood function. And then when we need to distinguish the parameter value that gives the distribution for the sample, we will use \( \theta_* \) and refer to it as the true value of the parameter (of course, we can still use \( \theta_* \) as a free variable when studying response distributions under various conditions).

We also face a small problem connected with the arbitrary constant in the expression for a likelihood function. The direct way of eliminating the constant is to choose a representative or standardize the likelihood function in some convenient way. We will do this, but we will do it in a very special and convenient way: we will standardize with respect to the true value \( \theta_* \). In an application we would not be able to do this, but for theoretical discussions here it is very convenient—and besides it is only a device for examining likelihood ratios or log-likelihood differences. Accordingly, in the present context we take the log-likelihood function to be

\[
I_n(y|\theta) = -\frac{n}{2\sigma_0^2} (\bar{y} - \theta)^2 + a = -\frac{n}{2\sigma_0^2} (\bar{y} - \theta_*^2) + \frac{n}{2\sigma_0^2} (\bar{y} - \theta_*^2)
\]

\[
= -\frac{1}{2\sigma_0^2} (\delta - \delta^*)^2 + \frac{1}{2\sigma_0^2} \delta^2.
\]

This log-likelihood function and the corresponding score function are plotted in Figure 8.15; note the adjusted parameter \( \delta = \sqrt{n}(\theta - \theta_*) \) and the adjusted variable \( \delta^* = \sqrt{n}(\bar{y} - \theta_*) \).

\[\text{FIGURE 8.15}\]

Normal(\( \theta, \sigma_0 \)). Distribution of the log-likelihood and score in terms of a standard normal variable \( z \). Note that \( \delta = \sqrt{n}(\theta - \theta_*) \) and \( \delta^* = \sqrt{n}(\bar{y} - \theta_*) \).
The distribution theory in Section 8.2 gives immediately the distribution theory for
the log-likelihood function and for the score since both are functions of the likelihood
statistic \( \bar{y} \). The distribution of the likelihood statistic \( \bar{y} \) is normal(\( \theta_* \), \( \sigma_0 / \sqrt{n} \)); thus we
write \( \bar{y} = \theta_* + \sigma_0 / \sqrt{n} \) or \( \delta = \sigma_0 z \) in terms of a standard normal \( z \). Now first we note
that the distribution of the log-likelihood function (1) is given by its dependence on \( \bar{y} \)
or on \( \delta \). Then we note the distribution of several salient characteristics: the deviation
from \( \theta_* \) to the point of maximum log-likelihood is

\[
(2) \quad \bar{y} - \theta_* = \frac{\sigma_0}{\sqrt{n}} z, \quad \text{or, as adjusted,} \quad \delta = \sigma_0 z;
\]

the slope of the log-likelihood at the true \( \theta_* \) is

\[
(3) \quad S_n(\bar{y} | \theta_*) = \frac{\bar{y} - \theta_*}{\sigma_0^2 / n} = \frac{\sqrt{n}}{\sigma_0} z;
\]

and the height to the maximum log-likelihood is

\[
(4) \quad I_n(\bar{y} | \bar{y}) - I_n(\bar{y} | \theta_*) = \frac{1}{2\sigma_0^2 / n} (\bar{y} - \theta_*)^2 = (1/2)z^2;
\]

see Figure 8.15. For large samples we will see that these same distribution properties
are obtained with the general parametric model.

**B THE DISTRIBUTION OF THE SCORE: MEANS AND VARIANCES**

Consider a statistical model on a sample space \( S \) with density function \( f(y | \theta) \) with
\( \theta \) in \( \Omega = R \). We assume that in \( f(y | \theta) \) can be differentiated several times with respect
to \( \theta \) and that this differentiation can be carried through a sample-space integration sign;
in particular, this excludes variable-carrier models. The precise assumptions are
recorded at the end of this section. Note the following expression for the derivative of
the density function:

\[
(5) \quad \frac{\partial}{\partial \theta} f(y | \theta) = S(y | \theta)f(y | \theta).
\]

First, we differentiate the integral

\[
\int f(y | \theta) \, dy = 1
\]

with respect to \( \theta \):

\[
(6) \quad \int \frac{\partial}{\partial \theta} f(y | \theta) \, dy = \int S(y | \theta) \cdot f(y | \theta) \, dy = 0.
\]

Thus we obtain

\[
(7) \quad E(S(y | \theta) | \theta) = 0,
\]

which says that the mean value of the score function at the true value is zero, or the
mean slope of the log-likelihood at the true value is zero. This holds for all values for
the parameter \( \theta \).

Consider the notation \( E(h(y) | \theta) \), which designates the mean value of \( h(y) \) calculated
using the distribution with parameter value \( \theta \). The expression \( E(S(y | \theta) | \theta) \) is the mean
value of \( h(y) = S(y | \theta) \), the score calculated at the true parameter value \( \theta \). Note here
that the parameter argument for the score function is the same as the parameter of the
distribution; thus we have no need for the specialized notation.
Second, we differentiate a second time working from the middle expression in formula (6) and again use the adjustment (5) in terms of the logarithmic derivative:

\[
\int \frac{\partial}{\partial \theta} S(y|\theta) \cdot f(y|\theta) \, dy + \int S(y|\theta) \frac{\partial}{\partial \theta} f(y|\theta) \, dy = 0,
\]

where the prime denotes a derivative with respect to \( \theta \). Thus using (7) we obtain

\[
\text{var} \{ S(y|\theta) | \theta \} = E(S^2(y|\theta) | \theta) = -E(S'(y|\theta) | \theta),
\]

which says that the variance of the score at the true value is equal to the negative mean of the derivative of the score at the true value—or to the negative mean of the second derivative of the log-likelihood at the true value. We write this

\[
\text{var} \{ S(y|\theta) | \theta \} = -E(S'(y|\theta) | \theta) = j(\theta),
\]

where \( j(\theta) \) is called the Fisher information as obtained from a single response. The significance of the term "information" will appear with results later in this section.

From a single \( y \) from the normal(\( \theta, \sigma_0^2 \)), we have

\[
S(y|\theta) = \frac{y - \theta}{\sigma_0^2}, \quad j(\theta) = \text{var} \{ S(y|\theta) | \theta \} = \frac{\sigma_0^2}{\sigma_0^2} = \frac{1}{\sigma_0^2}.
\]

Third, we consider a sample \( \{y_1, \ldots, y_n\} \) from the distribution given by \( f(y|\theta) \). Then

\[
\ln f(y|\theta) = \ln f(y_1|\theta) + \cdots + \ln f(y_n|\theta),
\]

(11) \[ S_n(y|\theta) = S(y_1|\theta) + \cdots + S(y_n|\theta), \]

\[ S'_n(y|\theta) = S'(y_1|\theta) + \cdots + S'(y_n|\theta). \]

We now examine mean values for the score and the score derivative at the true value \( \theta \). The terms on the right side of (11) are statistically independent; thus we obtain

\[
E(S_n(y|\theta) | \theta) = 0,
\]

(12) \[ \text{var} \{ S_n(y|\theta) | \theta \} = J(\theta) = n j(\theta) = -E(S'_n(y|\theta) | \theta), \]

where \( J(\theta) = n j(\theta) \) is the Fisher information as obtained from a sample of \( n \).

For a sample \( \{y_1, \ldots, y_n\} \) from the normal(\( \theta, \sigma_0^2 \)), we have

\[
S_n(y|\theta) = \frac{\overline{y} - \theta}{\sigma_0^2/n}, \quad J(\theta) = \frac{\sigma_0^2/n}{\sigma_0^2/n^2} = \frac{n}{\sigma_0^2}.
\]

The following assumption supports the preceding analysis.

**ASSUMPTION 1**

The derivatives of the density function satisfy \( |\partial f(y|\theta)/\partial \theta| < M_1(y) \) and \( |\partial^2 f(y|\theta)/\partial \theta^2| < M_2(y) \), where \( M_1(y) \), \( M_2(y) \) and

\[
\frac{\partial^2}{\partial \theta^2} \ln f(y|\theta) \cdot f(y|\theta)
\]

are integrable on the sample space \( \mathcal{S} \).
THE DISTRIBUTION OF THE LARGE-SAMPLE LIKELIHOOD FUNCTION

Consider a statistical model with density \( f(y|\theta) \) as discussed in the preceding section. Let \( (y_1, \ldots, y_n) \) be a sample of \( n \). We now examine the distribution of the large-sample likelihood function. We do this in terms of log-likelihood and as before we use \( \theta_* \) to refer to the true value of the parameter. The distribution is presented in the following theorem based on Assumption 1 and on Assumption 3 recorded with the proof in Section F.

**THEOREM 2**

As \( n \to \infty \), the log-likelihood has a limiting quadratic shape

\[
I_n(y|\theta) = -\frac{nJ(\theta_*)}{2} (\hat{\theta}(y) - \theta)^2 + a = -\frac{j(\theta_*)}{2} (\delta - \delta)^2 + \frac{j(\theta_*)}{2} \delta^2;
\]

this is the quadratic form (1) found in the normal(0, \( \sigma^2 \)) case but with the factor \( n/\sigma^2 \) replaced by \( nJ(\theta_*) \). Note the adjusted parameter \( \delta = \sqrt{n}(\hat{\theta}(y) - \theta_*) \) and adjusted variable \( \delta = \sqrt{n}(\hat{\theta}(y) - \theta_*) \).

The scaled deviation \( \sqrt{n}(\hat{\theta}(y) - \theta_*) \) of the maximum likelihood estimate from the true value has a limiting central normal distribution with variance 1/j(\theta_*).

The scaled slope \( S_n(y|\theta)/\sqrt{n} \) of the log-likelihood at the true value has a limiting central normal distribution with variance \( j(\theta_*) \). The height \( I_n(y|\theta) - I_n(y|\theta_*) \) to the maximum log-likelihood has a limiting(1/2) chi-square(1) distribution.

The proof is given in Section F.

The limiting distribution can be presented in a more convenient and informal way. See Figure 8.16 and note the close similarities to Figure 8.15. Let \( z \) designate a standard normal variable; the limiting distribution can be presented as follows:

\[
\hat{\theta}(y) - \theta_* \sim \frac{1}{\sqrt{nJ(\theta_*)}} z, \quad \text{or, as adjusted,} \quad \delta = \frac{1}{\sqrt{j(\theta_*)}} z.
\]

\[
S_n(y|\theta_*) \sim \frac{1}{\sqrt{nJ(\theta_*)}} z.
\]

\[
I_n(y|\theta) - I_n(y|\theta_*) \sim (1/2)z^2.
\]

Note that the large-sample likelihood is the same as if we had taken a sample of \( n \) from a normal distribution with mean \( \theta_* \) and reciprocal variance or precision \( j(\theta_*) \); the maximum-likelihood estimate takes the place of the sample average. Thus we see that \( j(\theta_*) \) is information in the sense of precision: each sample component contributes \( j(\theta_*) \) to the precision \( nJ(\theta_*) \) of the estimate \( \hat{\theta}(y) \), which locates the large-sample likelihood.

Also note in a parallel way that the limiting form of the log-likelihood is the same as would be obtained from the limiting marginal distribution of the maximum-likelihood estimate \( \hat{\theta} \).

**D TESTS AND CONFIDENCE INTERVALS FROM A LARGE SAMPLE**

Theorem 2 shows that the large-sample likelihood depends on a single sample characteristic \( \hat{\theta}(y) \). We can thus view \( \hat{\theta} = \hat{\theta}(y) \) as the large-sample likelihood statistic and accordingly use it as a basis for tests and confidence intervals.
Sec. 8.6: Inference from Large Samples Using Likelihood

FIGURE 8.16
General case: limiting distribution of the log-likelihood and score in terms of a standard normal variable $z$. Note $\delta = \sqrt{n} (\theta - \theta_0)$ and $\bar{\delta} = \sqrt{n} (\bar{\theta}(y) - \theta_0)$.

Let $\{y_1, \ldots, y_n\}$ be an observed large sample from a distribution satisfying Assumption 1 in Section B and Assumption 3 in Section F.

First, consider the hypothesis $\theta = \theta_0$. The hypothesis can be tested by calculating

$$ (nj(\theta_0))^{1/2} (\bar{\theta}(y) - \theta_0) $$

and comparing the value with the standard normal distribution. Alternatively, the hypothesis can be tested by calculating

$$ (nj(\theta_0))^{-1/2} S_n(y | \theta_0) $$

and comparing the value with the standard normal distribution. Or it can be tested by calculating

$$ 2 \ln \left( \prod \frac{f(y_i | \bar{\theta}(y))}{f(y_i | \theta_0)} \right) $$

and comparing the value with the chi-square(1) distribution; large values denote low likelihood at $\theta_0$ and thus provide evidence against the hypothesis. In the limit these tests are equivalent; for moderate samples they may give slightly different results, and a choice will often be based on the ease with which computations can be made.

Second, consider the formation of a confidence interval for $\theta$. A $1 - \alpha$ confidence interval can be based on

$$ P(-z_{\alpha/2} < (nj(\theta))^{1/2} (\bar{\theta}(y) - \theta) < z_{\alpha/2}) \approx 1 - \alpha, $$

or

$$ P\left( -z_{\alpha/2} < \frac{S_n(y | \theta)}{(nj(\theta))^{1/2}} < z_{\alpha/2} \right) \approx 1 - \alpha. $$
or

\[(22) \quad P \left( 2 \ln \frac{\prod f(y_i|\hat{\theta}(y))}{\prod f(y_i|\theta)} \leq \chi^2_n \right) \approx 1 - \alpha.\]

The observed values would be substituted in the relations; the relations would then be rearranged (as we have discussed before) to give the interval of acceptable values, an approximate $1 - \alpha$ confidence interval for $\theta$. An approximate $1 - \alpha$ confidence interval for $\theta$ can be obtained directly from the first expression (20):

\[(23) \quad \hat{\theta}(y) \pm z_{\alpha/2}(n\hat{f}(\hat{\theta}(y)))^{-1/2};\]

compare this with the results in Section 7.2C and 7.2F.

**E. BINOMIAL TESTS AND CONFIDENCE INTERVALS**

Let $(x_1, \ldots, x_n)$ be a sample from the Bernoulli($\rho$) distribution with $\rho$ in (0, 1). The density function for a single $x$ is

\[f(x|\rho) = \rho^x q^{1-x}, \quad x = 0, 1.\]

Thus

\[l(x|\rho) = x \ln \rho + (1 - x) \ln q,\]

\[S(x|\rho) = \frac{x}{\rho} - \frac{1 - x}{q},\]

\[S'(x|\rho) = -\frac{x}{\rho^2} - \frac{1 - x}{q^2},\]

from which we obtain

\[E(S(x|\rho)|\rho) = \frac{\rho}{\rho} - \frac{q}{q} = 0;\]

\[E(S'(x|\rho)|\rho) = -\frac{\rho}{\rho^2} - \frac{q}{q^2} = -\frac{1}{\rho q};\]

\[\text{var} (S(x|\rho)|\rho) = \frac{\rho q}{(pq)^2} = \frac{1}{\rho q} = i(\rho).\]

The maximum-likelihood estimate from the sample is obtained by solving $S_x(x|\rho) = 0$; thus we obtain $\hat{\rho} = \Sigma x_i/n$ as in Example 8.1.2.

The hypothesis $\rho = \rho_0$ can be tested by calculating, say (17),

\[(24) \quad \frac{\hat{\rho} - \rho_0}{(\rho_0 q_2/n)^{1/2}},\]

and comparing the value with the standard normal.

A confidence interval for $\rho$ can be calculated on the basis of say (20). At the 95 percent level we have

\[-1.96 \leq \frac{\hat{\rho} - \rho}{(\rho q/n)^{1/2}} \leq 1.96.\]
This can be rearranged to give

\[ p_1(\hat{\rho}) < p < p_2(\hat{\rho}), \]

where \( p_1(\hat{\rho}) \) and \( p_2(\hat{\rho}) \) are the solutions of the equation

\[ (\hat{\rho} - p)^2 = (1.96)^2 \frac{pq}{n}. \]

Thus

(25) \[ P(p_1(\hat{\rho}) < p < p_2(\hat{\rho})) = 95\% \]

and \( (p_1(\hat{\rho}), p_2(\hat{\rho})) \) is a 95 percent confidence interval for \( p \). At the \( 1 - \alpha \) level we have

\[-z_{\alpha/2} < \frac{\hat{\rho} - p}{(pq/n)^{1/2}} < z_{\alpha/2}.\]

This can be rearranged to give the \( 1 - \alpha \) confidence interval \( (p_1(\hat{\rho}), p_2(\hat{\rho})) \), where \( p_1(\hat{\rho}) \) and \( p_2(\hat{\rho}) \) are the solutions of the equation

\[ (\hat{\rho} - p)^2 = z_{\alpha/2}^2 \frac{\rho(1 - \rho)}{n}; \]

these solutions can be given explicitly as

(26) \[ p_1(\theta) = \hat{\rho} + \frac{z_{\alpha/2}^2 \left( \hat{\theta} - \hat{\rho} \right)}{2n} + \frac{z_{\alpha/2}^2}{\sqrt{n}} \left( \hat{\theta} \hat{\theta} + \frac{z_{\alpha/2}^2}{4n} \right)^{1/2} \frac{1 + z_{\alpha/2}^2/n}{1 + z_{\alpha/2}^2/n}. \]

If \( n \) is reasonably large, the confidence interval can be approximated as

\[ \hat{\rho} \pm z_{\alpha/2} \left( \frac{\hat{\theta} \hat{\theta}}{n} \right)^{1/2}. \]

**F SUPPLEMENT: PROOF OF THEOREM 2**

We now record the proof of Theorem 2. For this it is convenient to follow the notational device in Section A and standardize the likelihood function with respect to the true value \( \theta_* \). Thus the log-likelihood can be written without an arbitrary constant in the form

\[ l_n(y | \theta) = \ln f(y | \theta) - \ln f(y | \theta_*). \]

First, we expand the log-likelihood in a Taylor series about the true value \( \theta_* \):

\[ l_n(y | \theta) = (\theta - \theta_*) S_n(y | \theta_*) + \frac{(\theta - \theta_*)^2}{2!} S_n'(y | \theta_*) + \frac{(\theta - \theta_*)^3}{3!} R \sum_1^n N(y_i), \]

where \( |R| \leq 1 \). The coefficient \( S_n(y | \theta_*) \) has standard deviation \( (n/\theta_*)^{1/2} \); accordingly, we examine \( \theta = \theta_* + \delta n^{-1/2} \), where \( \delta \) is deviation in units of length \( n^{-1/2} \):

\[ l_n(y | \theta_* + \delta n^{-1/2}) = \delta \frac{S_n(y | \theta_*)}{n^{1/2}} + \delta^2 \frac{S_n'(y | \theta_*)}{2 \sqrt{n}} + \frac{\delta^3}{3 \sqrt{n}} R \sum_1^n N(y_i) \]
Note the correspondence between $\theta$ values and $\delta$ values:

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>$\theta_*$</th>
<th>$\hat{\theta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta$</td>
<td>0</td>
<td>$\hat{\delta}$</td>
</tr>
</tbody>
</table>

By the Central Limit Theorem 6.2.2 we have that

$$ w = \frac{S_n(y|\theta_*)}{n^{1/2}} $$

has a limiting normal($0$, $\sqrt{j(\theta_*)}$) distribution. By the Strong Law of Large Numbers (Theorem 6.3.12), we have that

$$ S_n'(y|\theta_*) + \frac{\delta R}{3 \sqrt{n}} \sum_i N(y_i) $$

has limit $-j(\theta_*)$ with probability 1. Thus the limiting form of the log-likelihood is given by

$$ \delta w - \frac{\delta^2}{2} j(\theta_*) = -\frac{j(\theta_*)}{2} (\hat{\delta} - \delta)^2 + \frac{j(\theta_*)}{2} \hat{\delta}^2, $$

where the maximum-likelihood estimate $\hat{\delta} = w/j(\theta_*) = \sqrt{n}(\hat{\theta}(y) - \theta_*) = z/\sqrt{j(\theta_*)}$ has a limiting normal($0$, $1/\sqrt{j(\theta_*)}$) distribution. This is pictured in Figure 8.16. Note that $\hat{\delta} = \sqrt{n}(\hat{\theta} - \theta_*)$ uses a new origin at $\theta_*$ with scaling unit $n^{-1/2}$. This proof uses the following assumption together with Assumption 1:

**ASSUMPTION 3**

The information $j(\theta) > 0$ and $|\partial^3 \ln f(y|\theta)/\partial \theta^3| < N(y)$, where $N(y)f(y|\theta)$ is integrable on the sample space $S$.

**G** **SUPPLEMENT: LARGE-SAMPLE LIKELIHOOD WITH A VECTOR PARAMETER**

Now consider the case of a vector parameter $\theta = (\theta_1, \ldots, \theta_r)^T$ in $R^r$ or in an open set of $R^r$. The score function has a coordinate for each coordinate of $\theta$:

$$ S_1(y|\theta) = \frac{\partial}{\partial \theta_1} \ln f(y|\theta), $$

$$ \vdots $$

$$ S_r(y|\theta) = \frac{\partial}{\partial \theta_r} \ln f(y|\theta). $$

The score function is thus a vector-valued function $S(y|\theta)$ and it records the gradient of the log-likelihood function. By the methods in Section B we obtain

$$ E(S(y|\theta)|\theta) = \mathbf{0}, \quad \text{VAR} (S(y|\theta)|\theta) = j(\theta), $$

where the matrix $j(\theta)$ can also be calculated from
\[ j(\theta) = -E\left( \begin{pmatrix} \frac{\partial^2}{\partial \theta_1^2} \ln f(y|\theta) & \cdots & \frac{\partial^2}{\partial \theta_1 \partial \theta_r} \ln f(y|\theta) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial \theta_r \partial \theta_1} \ln f(y|\theta) & \cdots & \frac{\partial^2}{\partial \theta_r^2} \ln f(y|\theta) \end{pmatrix} \right) | \theta). \]

Now consider a large sample \( y = (y_1, \ldots, y_n)' \). By the Central Limit Theorem 6.2.5 we have that the scaled slope
\[ w = \frac{S(y|\theta_0)}{n^{1/2}} \]

of the log-likelihood at the true value has a limiting multivariate normal(0; \( j(\theta_0) \)) distribution.

Then by the methods in Section F we can show that the log-likelihood standardized with respect to the true value \( \theta_* \) has the limiting form
\[ \delta' w - \frac{1}{2} \delta' j(\theta_0) \delta = \frac{1}{2} (\delta - \hat{\delta})' j(\theta_0)(\delta - \hat{\delta}) + \frac{1}{2} \hat{\delta}' j(\theta_0) \hat{\delta}, \]

where \( \delta = \sqrt{n}(\theta - \theta_0) \) and where \( \hat{\delta} \) is given by
\[ \hat{\delta} = j^{-1}(\theta_0) w = \sqrt{n}(\hat{\theta}(y) - \theta_0). \]

From this it follows that the offset \( \hat{\delta} = j^{-1}(\theta_0) w \) to the point of maximum likelihood has a multivariate normal(0; \( j^{-1}(\theta_0) \)) distribution; thus we can describe \( \hat{\theta}(y) \) as multivariate normal located at the true \( \theta_* \) with variance matrix \( J^{-1}(\theta_0) = n^{-1} j^{-1}(\theta_0) \).

It also follows that the height to the maximum is distributed as \( (1/2) \chi^2 \) where \( \chi^2 \) has a chi-square(r) distribution; thus the quadratic part \( \delta' J(\theta_0) \delta \) of the multivariate normal density becomes \( \Sigma_j z_j^2 \) when related to standard normal variables.

The distribution properties for the likelihood can be summarized informally as
\[ \hat{\theta}(y) \sim \text{multivariate normal}(\theta_*; n^{-1} j^{-1}(\theta_0)). \]
\[ S(y|\theta_0) \sim \text{multivariate normal}(0; nj(\theta_0)). \]
\[ 2 \ln \frac{\prod f(y_i|\hat{\theta}(y))}{\prod f(y_i|\theta_0)} \sim \chi^2(r). \]

**H EXERCISES**

1. A possibly biased coin was tossed 6400 times, yielding 3277 heads. Use the large-sample results in Sections C and E:
   (a) Test the hypothesis: \( p = 1/2 \).
   (b) Calculate a 95 percent confidence interval for \( p \).

2. Calculate \( S(y|\theta) \), \( S'(y|\theta) \), and \( j(\theta) \) for the Poisson(\( \theta \)) distribution.

3. Assuming that traffic fatalities in a particular city are approximately Poisson, derive a 95 percent confidence interval for the mean annual rate \( \theta \) given the observed count \( y = 113 \) for one year. Use Exercise 2 and the asymptotic results even though \( n = 1 \) (the annual count can be viewed as a sum of say 52 weekly counts).

4. A sample of nine from a Poisson(\( \theta \)) distribution yielded \( \bar{y} = 28\frac{1}{3} \). Use Exercise 2 and the large-sample results in Section D:
   (a) Test the hypothesis \( \theta = 25 \).
   (b) Form a 95 percent confidence interval for \( \theta \).
5 Let \( \{y_1, \ldots, y_n\} \) be a sample from the normal \((0, \sigma)\) with \( \theta = \sigma^2 \) in \( R^+ \). Calculate \( S(y|\sigma^2), S'(y|\sigma^2), \) and \( n\sigma(\sigma^2) \). Calculate \( 2 \ln \left( \frac{\prod f(y_i|\sigma^2)}{\prod f(y_i|\sigma^2)} \right) \) and quote its limiting distribution under the hypothesis.

I PROBLEMS

6 Calculate \( S(y|\theta), S'(y|\theta), \) and \( j(\theta) \) for the location Cauchy \( f(y|\theta) = \sigma^{-1}(1 + (y - \theta)^2)^{-1} \) with \( \theta \) in \( R \); use \( \int_{-\infty}^{\infty} (1 - t^2)(1 + t^2)^{-3} \, dt = \pi/4 \). Determine \( \sigma_0 \) so that the normal(\( \theta, \sigma_0 \)) distribution has the same information as the location Cauchy.

7 Calculate \( S(y|\theta), S'(y|\theta), \) and \( j(\theta) \) for the scale exponential \( f(y|\theta) = \theta^{-1} \exp \left( -y/\theta \right) \) on \( R^+ \) with \( \theta \) in \( R^+ \).

8 (continuation) For a large sample \( \{y_1, \ldots, y_n\} \) determine a 95 percent confidence interval for \( \theta \) based on the limiting likelihood.

9 Calculate \( S(y|\theta), S'(y|\theta), \) and \( j(\theta) \) for the scale gamma \( f(y|\theta) = \Gamma^{-1}(\theta)(y/\theta)^{\beta-1} \exp \left( -y/\theta \right)/\theta \) on \( R^+ \) with \( \theta \) in \( R^+ \).

10 (continuation) For a sample of \( n \), calculate \( 2 \ln \left( \frac{\prod f(y_i|\theta)}{\prod f(y_i|\theta_\ast)} \right) \). Quote its limiting distribution from the results in this section.

11 Calculate \( S(y|\theta), S'(y|\theta), \) and \( j(\theta) \) for the location(\( \theta \)) logistic distribution. Determine the normal distribution \( \{\theta, \sigma_0\} \) that has the same information as the location logistic.

12 Calculate \( S = (S_1(y|\mu, \sigma^2), S_2(y|\mu, \sigma^2)) \) for the normal(\( \mu, \sigma^2 \)) in \( R \times R^+ \). Calculate the derivatives of \( S_1 \) and \( S_2 \) and thus determine the information matrix \( J(\mu, \sigma^2) \).

13 (continuation) For a sample of \( n \), calculate \( 2 \ln \left( \frac{\prod f(y_i|\mu, \sigma^2)}{\prod f(y_i|\mu_\ast, \sigma^2)} \right) \). Quote its limiting distribution from the results in this section.

14 Let \( \{y_1, \ldots, y_n\} \) be a sample from a distribution with density \( f(y|\theta) \) with \( \theta \) in \( \Omega = \{\theta_1, \theta_2\} \). Let \( \mu_1 \) and \( \sigma_1^2 \) be the mean and variance of \( l_1(y) = \ln f(y|\theta_1) - \ln f(y|\theta_2) \) under the distribution \( f(y|\theta_1) \). Let \( \mu_2 \) and \( \sigma_2^2 \) be the mean and variance under the distribution \( f(y|\theta_2) \); suppose that \( \sigma_1 > 0 \) and \( \sigma_2 > 0 \). Show that \( l_2(y) \) has a distribution with limiting normal form \( (\mu_1 - \mu_2) / \sqrt{\sigma_1^2 + \sigma_2^2} \) when the distribution being sampled is \( f(y|\theta_1) \).

NOTES AND REFERENCES

Recently the observed likelihood function has been recognized as a basic means for assessing data. The acceptance of the likelihood function ranges from the Bayesians, who use only the likelihood function, to the traditional statisticians, who find that the likelihood concept provides feasible new ways of determining sufficient information and minimal sufficiency. In between there is a wide range of statisticians who use the likelihood function for some data assessment; for this in an extreme form, see Edwards (1972).

The likelihood function leads us to \( (\bar{y}, s^2) \) for the normal(\( \mu, \sigma \)) model. Sections 8.2 and 8.3 examine distribution theory and inference methods for the normal model.

In a similar way, the likelihood function leads to the likelihood statistic for a general statistical model; see Section 8.4. This provides a simple direct route to the usual sufficiency and minimal sufficiency results. Indeed, the likelihood statistic seems to replace the minimal sufficient statistic as the basic reduction method of statistics.

 Sufficiency for exponential models follows almost trivially by the use of the likelihood statistic; compare, for example, with Barankin and Maitra (1963).

The chapter concludes with some large-sample results for the likelihood function. We see that the general statistical model using a large sample can be analyzed in a way quite similar to that for the normal.

The proper definition of the likelihood function is essential for these results. The reader is to be alerted to the importance of the "arbitrary multiplicative constant" in the definition of the
likelihood function—it's absence in many current statistical texts prevents the proper and full use of the concept.

On the likelihood function:


On large-sample likelihood:


On sufficiency:


On sufficiency and exponential models:


In Chapter 7 we introduced estimation, tests of significance, and confidence intervals as the three basic concepts of statistical inference. We examined the concepts informally in that chapter, using the large-sample and symmetry results from the probability chapters. Then in Chapter 8, as part of examining likelihood with statistical models, we examined the three concepts in relation to the normal statistical model.

In Chapter 9 we shall examine estimation in detail. First we investigate briefly three traditional methods of obtaining an estimate, least squares, moments, and maximum likelihood (Section 9.1). We then examine least squares in greater detail for models having a certain linear structure for location (Section 9.2). We investigate ways of combining direct estimates to obtain an overall estimate, weighting by reciprocal variance, and of combining indirect estimates, the Gauss–Markov method (Section 9.3). We investigate ways of obtaining best estimates with a one- or a several-parameter model, by the use of the information inequality (Section 9.4) and by use of sufficiency and completeness (Section 9.5). The Bayesian method of inference is examined in Supplement Section 9.6.
9

ESTIMATION

9.1

SOME TRADITIONAL METHODS

A surveyor investigating the distance between two points may make several measurements or determinations of the distance and then combine the measurements in some suitable way to obtain an estimate of the distance. This is an example of the direct measurement of a physical quantity.

Or, more generally, a surveyor investigating the boundary of a piece of property may make a succession of measurements of angle and distance progressing from a starting point around the boundary and back to the starting point. Each measurement would directly refer to a physical quantity connected with the boundary. However, a relation exists among the physical quantities in the sense that the finishing point is, in fact, the starting point; the surveyed line has been closed. Accordingly, each measurement has some bearing on a range of physical quantities and there is thus some indirect measurement of physical quantities. Various tie lines or check lines could also be included in the survey and there would then be further indirect measurement on the physical quantities.

In this section we examine three traditional methods for obtaining an estimate or plausible value for a parameter. Direct measurement is somewhat straightforward; these methods are primarily concerned with indirect measurement.

A THE METHOD OF LEAST SQUARES

The method of least squares was proposed by Legendre and Gauss at the beginning of the nineteenth century. The method is concerned with measurements on physical quantities that satisfy a relationship or are expressed in terms of more elementary quantities. The method produces adjusted measurements that do satisfy the relationships and it produces estimates of the elementary quantities.

Consider an application with an observed response vector \( y = (y_1, \ldots, y_n)' \). Some related theory prescribes a location for the response vector: \( v(\theta) = (v_1(\theta_1, \ldots, \theta_r), \ldots, v_n(\theta_1, \ldots, \theta_r))' \), where \( \theta \) is a parameter having \( r \) coordinates; these coordinates may be more elementary quantities of the system being investigated. For example, we might have measurements \( (y_1, y_2, y_3) \) for two sides and the hypotenuse of a right-angled triangle and we could have \( v(\theta) = (\theta_1, \theta_2, \sqrt{\theta_1^2 + \theta_2^2}) \). The method produces a fitted response vector \( \hat{y} = (\hat{y}_1, \ldots, \hat{y}_n) \) that satisfies the relations indicated by the location vector, and it produces an estimate \( \hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_r) \) for the parameter.
FIGURE 9.1
Observation $y$. The set $\mathfrak{F}$ of possible locations $\nu(\theta)$. The minimizing point $\nu(\tilde{\theta})$ has deviation vector $y - \nu(\tilde{\theta})$ orthogonal to the tangent vectors in the direction of $\theta_1$ change, of $\theta_2$ change, . . . .

The least-squares method has been widely used in astronomy and geodesy, where a large number of measurements can be taken on a variety of physical dimensions. The geometry of the system gives relationships among the dimensions or presents the dimensions in terms of some independent components or parameters. The method of least squares can be viewed as adjusting or fitting the measurements in an orderly way so that they satisfy the geometry of the system.

The method of least squares also has important uses in statistical theory. In one direction it provides various computational procedures that are very useful. And in another direction it gives an estimation procedure that has certain optimum properties, depending on the form of the distribution.

We now describe the method. Let $y = (y_1, \ldots, y_n)'$ be an observed response. And let $\nu(\theta) = (\nu_1(\theta), \ldots, \nu_n(\theta))'$ be a possible location for the response; the model for location is then the set

$$\mathfrak{F} = \{ \nu(\theta) : \theta \in \Omega \}$$

of possible response locations obtained as the parameter $\theta$ ranges over the parameter space $\Omega$. We can picture $y$ and $\mathfrak{F}$ in Euclidean space as indicated in Figure 9.1:

<table>
<thead>
<tr>
<th>Y</th>
<th>$y_1$</th>
<th>$y_2$</th>
<th>$\cdots$</th>
<th>$y_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\nu(\theta)$</td>
<td>$\nu_1(\theta)$</td>
<td>$\nu_2(\theta)$</td>
<td>$\cdots$</td>
<td>$\nu_n(\theta)$</td>
</tr>
</tbody>
</table>

DEFINITION 1

The least-squares estimate of $\theta$ is the value $\hat{\theta} = \hat{\theta}(y)$ that minimizes the sum of squares

$$S^2 = (y_1 - \nu_1(\theta))^2 + \cdots + (y_n - \nu_n(\theta))^2 = (y - \nu(\theta))'(y - \nu(\theta)).$$

See Figure 9.1: $S^2$ is the squared distance from the point $\nu(\theta)$ to the point $y$; the method is to choose the value for $\theta$ that makes $\nu(\theta)$ closest to the observation $y$. If the set $\mathfrak{F}$
is closed in $\mathbb{R}^n$, then the minimizing parameter value will always exist: the closest point $\hat{y} = v(\hat{\theta})$ is called the fitted response vector.

Note that we have not mentioned any response distribution. Of course with a specification for the response distribution, the method may or may not be appropriate; some aspects of this will be examined later.

Now suppose that $v(\theta)$ is continuously differentiable with respect to $\theta$. If the minimizing value $\hat{\theta}$ is an interior point on the surface $\mathcal{P}$, then it can be calculated by differentiating with respect to $\theta$ and setting the derivative equal to zero:

$$
\begin{align*}
\frac{\partial S^2}{\partial \theta_1} &= 0, & \sum \frac{\partial v_i(\theta)}{\partial \theta_1} (y_i - v_i(\theta)) &= 0, \\
\vdots \\
\frac{\partial S^2}{\partial \theta_r} &= 0, & \sum \frac{\partial v_i(\theta)}{\partial \theta_r} (y_i - v_i(\theta)) &= 0;
\end{align*}
$$

the differentiation produces factors 2 and $-1$: these have been factored from the right-hand expressions.

The first equation states that the deviation vector,

$$(y_1 - v_1(\theta), \ldots, y_n - v_n(\theta)),$$

is orthogonal to the tangent vector,

$$\left( \frac{\partial v_1(\theta)}{\partial \theta_1}, \ldots, \frac{\partial v_n(\theta)}{\partial \theta_1} \right),$$

in the direction of $\theta_1$ change; $\ldots$; the last equation states that the deviation vector is orthogonal to the tangent vector,

$$\left( \frac{\partial v_1(\theta)}{\partial \theta_r}, \ldots, \frac{\partial v_n(\theta)}{\partial \theta_r} \right),$$

in the direction of $\theta_r$ change; see Figure 9.1.

In certain cases discussed in Section 9.2 these equations (2) are linear and can be solved directly. In other cases various iterative procedures can be used; for example Section 9.4D.

Consider a very simple example to illustrate the method; more substantial examples will be examined later.

**EXAMPLE 1**

Let $y = (y_1, \ldots, y_n)'$ be an observation with theoretical location $\beta x' = \beta(x_1, \ldots, x_n)' = (\beta x_1, \ldots, \beta x_n)'$, where the $x_i$ are known in value. The least-squares method is to choose the value $\beta = \hat{\beta}$ so as to minimize

$$S^2 = (y_1 - \beta x_1)^2 + \cdots + (y_n - \beta x_n)^2,$$

$$-\frac{1}{2} \frac{\partial S^2}{\partial \beta} = x_1(y_1 - \beta x_1) + \cdots + x_n(y_n - \beta x_n) = 0,$$

$$\sum x_i y_i - \beta \sum x_i x_i = x'y - \beta x' x = 0.$$
Thus the least-squares estimate of $\beta$ is

$$\hat{\beta} = \frac{x'y}{x'x} = (x'x)^{-1}x'y,$$

and the fitted response vector is

$$\hat{\beta}x = x\hat{\beta} = x(x'x)^{-1}x'y.$$

Note that $x(x'x)^{-1}x'$ is an $n \times n$ matrix that transforms the given response vector into the fitted response vector.

A clear picture of the geometry of least-squares procedures is particularly useful and informative. For the present example, see Figure 9.2 and recall the geometry in Sections 7.1E and 8.2C. The set $\mathcal{L} = \{bx : b \in R\}$ is the line generated by the vector $x$. The point on the line $\mathcal{L}(x)$ that is closest to $y$ is seen by equation (3) to be the projection $\hat{\beta}x$ of $y$ onto the line; the projection is, of course, the point whose residual vector $y - \hat{\beta}x$ is orthogonal to the line. Thus we have the orthogonal decomposition shown in Table 9.1.

**TABLE 9.1**

<table>
<thead>
<tr>
<th>Subspace</th>
<th>Dimension</th>
<th>Component</th>
<th>Squared Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{L}(x)$</td>
<td>1</td>
<td>$\hat{\beta}x$</td>
<td>$\hat{\beta}^2 \sum x_i^2$</td>
</tr>
<tr>
<td>$\mathcal{L}^{-1}(x)$</td>
<td>$n - 1$</td>
<td>$y - \hat{\beta}x$</td>
<td>$\sum (y_i - \hat{\beta}x_i)^2$</td>
</tr>
<tr>
<td></td>
<td>$n$</td>
<td>$y$</td>
<td>$y'y$</td>
</tr>
</tbody>
</table>
### B THE METHOD OF MOMENTS

The method of moments is a very old method of estimation. A large part of early statistics involved fitting a distribution form to data involving a moderately large number of observations. Various distribution forms with a small number of parameters have been proposed; prominent among these is a Pearson family of curves that includes the normal, gamma, beta, and Student distributions. The method of moments provides a rather straightforward method of fitting such a distribution form to a large number of observations.

Let \( y_1, \ldots, y_n \) be an observed sample from a distribution that has a parameter \( \theta = (\theta_1, \ldots, \theta_r)' \) with \( r \) coordinates. The moments of the distribution will, of course, depend on the parameter \( \theta \) of the distribution. The method of moments involves choosing as an estimate that parameter value \( \theta^* \) for which the first few moments of the distribution are equal to the corresponding moments of the sample; see Table 9.2.

<table>
<thead>
<tr>
<th>Sample Moment</th>
<th>Distribution Moment</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m_1 = \sum y_i/n )</td>
<td>( \mu_1(\theta) = E(y</td>
</tr>
<tr>
<td>( m_2 = \sum y_i^2/n )</td>
<td>( \mu_2(\theta) = E(y^2</td>
</tr>
<tr>
<td>[ \vdots ]</td>
<td>[ \vdots ]</td>
</tr>
<tr>
<td>( m_r = \sum y_i^r/n )</td>
<td>( \mu_r(\theta) = E(y^r</td>
</tr>
</tbody>
</table>

### DEFINITION 2

The **method-of-moments estimate** of \( \theta \) is the value \( \theta^* = \theta^*(y) \) obtained by solving the equations

\[
\begin{align*}
m_1 &= \mu_1(\theta), \\
\vdots \\
m_r &= \mu_r(\theta),
\end{align*}
\]

(6)

using the first few moments.

Typically as many moments will be needed as there are coordinates for \( \theta \).

### EXAMPLE 2

**The location scale normal**: Let \( y_1, \ldots, y_n \) be a sample from a normal(\( \mu, \sigma \)) distribution. The first two moments are shown in Table 9.3. Hence we obtain

\[
\bar{y} = \mu, \quad \bar{y}^2 = \mu^2 + \sigma^2; \quad \mu^* = \bar{y}, \quad \sigma^* = \frac{\sum y_i^2}{n} - \bar{y}^2 = \frac{n - 1}{n} s^2.
\]

Note that the method gives the sample second central moment as the estimate of variance.
TABLE 9.3
Sample and distribution moments

<table>
<thead>
<tr>
<th>Sample Moment</th>
<th>Distribution Moment</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{y} = \sum y_i / n )</td>
<td>( \mu )</td>
</tr>
<tr>
<td>( \bar{y}^2 = \sum y_i^2 / n )</td>
<td>( \mu^2 + \sigma^2 )</td>
</tr>
</tbody>
</table>

The method of moments can be very hazardous if more than just one or two parameters are estimated. Some indications for this come from the need to calculate higher powers of the observations; the higher powers are very sensitive to slight changes in the observations and the method can be unstable.

C  THE MAXIMUM-LIKELIHOOD METHOD

The method of maximum likelihood can handle both vector location as with least squares and distribution form as with the method of moments. It was proposed by R. A. Fisher in the 1920s as an alternative to the method of moments; some aspects of the method had been used by Gauss in the early nineteenth century. The method is very widely used in almost all areas of statistical application.

Consider a statistical model

\[ \{ f(\cdot | \theta) : \theta \in \Omega \} \]

and let \( y \) be an observed value. The likelihood function \( L(y | \cdot) \) and log-likelihood function \( l(y | \cdot) \) have been defined in Section 8.1D:

\[ L(y | \theta) = cf(y | \theta), \quad l(y | \theta) = \ln f(y | \theta) + a, \]

where \( c \) is in \( R^+ \) and \( a \) in \( R \).

DEFINITION 3

The maximum-likelihood estimate (MLE) of \( \theta \) is the value \( \hat{\theta} = \hat{\theta}(y) \) that maximizes the likelihood function \( L \):

\[ (7) \quad L(y | \hat{\theta}) = \sup_{\theta \in \Omega} L(y | \theta). \]

Note that the MLE estimate is the parameter value that maximizes the probability for the observed outcome.

If the likelihood function is continuously differentiable and if the maximum occurs on an open set of parameter values, then the MLE can be obtained as a solution of the score equation: for a real-valued parameter we have

\[ (8) \quad S(y | \theta) = \frac{\partial}{\partial \theta} \ln f(y | \theta) = 0; \]

and for a vector parameter \( \theta = (\theta_1, \ldots, \theta_p)^T \) we have
\[ S_1(y | \theta) = \frac{\partial \ln f(y | \theta)}{\partial \theta_1} = 0. \]

(9) \[ \vdots \]

\[ S_r(y | \theta) = \frac{\partial \ln f(y | \theta)}{\partial \theta_r} = 0. \]

Some examples involving MLE estimates were given in Section 8.1E. The large-sample importance of the MLE method is closely tied to the large-sample methods discussed in Section 8.6.

We now show a rather close connection between the least-squares method and the maximum likelihood method—for the case of normally distributed variation.

**THEOREM 4**

If \( y_1, \ldots, y_n \) are statistically independent and \( y_i - \nu_i(\theta) \) is normal(0, \( \sigma \)) with \( \sigma \) in \( R^+ \), then the maximum-likelihood estimate of \((\theta, \sigma^2)\) is given by

(a) the least-squares estimate \( \hat{\theta} \) for \( \theta \) obtained by minimizing

\[ S^2 = \sum_{i=1}^{n} (y_i - \nu_i(\theta))^2 \]

and (b) the average squared deviation

\[ s^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \nu_i(\hat{\theta}))^2 \]

for \( \sigma^2 \).

**Proof** The likelihood function from an observed \( y \) is given by

\[ L(y | \theta, \sigma^2) = c(2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} S^2 \right\}. \]

For given \( \sigma^2 \) this function is maximized by choosing \( \hat{\theta} \) to minimize \( S^2 \); thus \( \hat{\theta} \) is the least-squares estimate of \( \theta \).

Let \( S^2_0 \) be the minimized value of \( S^2 \). For maximization with respect to \( \sigma^2 \), we then have

\[ L(y | \hat{\theta}, \sigma^2) = c(\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} S^2_0 \right\}. \]

\[ l(y | \hat{\theta}, \sigma^2) = -\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} S^2_0 + \theta, \]

\[ \frac{\partial l(y | \hat{\theta}, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^2} S^2_0 = 0, \]

and hence \( \hat{\sigma}^2 = S^2_0 / n \).

**EXAMPLE 3**

The location-scale normal: Let \( (y_1, \ldots, y_n) \) be a sample from the normal(\( \mu, \sigma \)) with \( (\mu, \sigma^2) \) in \( R \times R^+ \). The least-squares estimate of \( \mu \) is obtained by minimizing
\[ S^2 = \sum_{i=1}^{n} (y_i - \mu)^2 \]

and is of course \( \hat{\mu} = \bar{y} \). Thus the maximum-likelihood estimate is

\[ (\hat{\mu}, \hat{\sigma}^2) = \left( \bar{y}, \sum (y_i - \bar{y})^2/n \right). \]

**D EXERCISES**

1. The observed response vector \((2, 4, -1, -1)\) has theoretical location \(\beta(1, 1, -1, -1) = \beta x'\).
   (a) Calculate the least-squares estimate of \(\beta\).
   (b) Calculate the fitted vector \(\hat{\beta}x\) and its squared length.
   (c) Calculate the residual vector \(y - \hat{\beta}x\) and its squared length.

2. (continuation) (a) The fitted vector \(\hat{\beta}x\) can be obtained by applying the projection matrix \(x(x'x)^{-1}x'\) to the response vector \(y\); calculate this \(4 \times 4\) projection matrix. (b) The residual vector \(y - \hat{\beta}x\) can be obtained by applying the projection matrix \(I - x(x'x)^{-1}x'\) to the response vector \(y\); calculate this \(4 \times 4\) projection matrix.

3. The observed response vector \((2, 2, 4, 5)\) has theoretical location \(\beta(1, 2, 3, 4)' = \beta x\).
   (a) Calculate the least-squares estimate of \(\beta\).
   (b) Calculate the fitted vector \(\hat{\beta}x\) and its squared length.
   (c) Calculate the residual vector \(y - \hat{\beta}x\) and its squared length.

4. (continuation)
   (a) The fitted vector \(\hat{\beta}x\) can be obtained by applying the projection matrix \(x(x'x)^{-1}x'\) to the response vector \(y\); calculate the \(4 \times 4\) projection matrix.
   (b) The residual vector \(y - \hat{\beta}x\) can be obtained by applying the projection matrix \(I - x(x'x)^{-1}x'\) to the response vector \(y\); calculate this \(4 \times 4\) projection matrix.

5. Let \((y_1, \ldots, y_n)\) be a sample from the Poisson(\(\lambda\)) distribution with \(\lambda\) in \(R^+\).
   (a) Calculate the least-squares estimate of the location \(\lambda\).
   (b) Calculate the method-of-moments estimate of \(\lambda\).
   (c) Calculate the maximum-likelihood estimate of \(\lambda\).

6. Let \((y_1, \ldots, y_n)\) be a sample from the binomial(\(n, p\)) distribution with \(p\) in \([0, 1]\).
   (a) Calculate the least-squares estimate of the location \(np\).
   (b) Calculate the method-of-moments estimate of \(np\).
   (c) Calculate the maximum-likelihood estimate of \(np\).

7. Let \((y_1, \ldots, y_n)\) be a sample from the normal(\(\mu_0, \sigma\)) distribution with \(\sigma\) in \(R^+\).
   (a) Calculate the method-of-moments estimate of \(\sigma^2\).
   (b) Calculate the maximum-likelihood estimate of \(\sigma^2\).

8. Let \((y_1, \ldots, y_n)\) be a sample from the uniform(0, 2\(\theta\)).
   (a) Calculate the least-squares estimate of \(\theta\).
   (b) Calculate the method-of-moments estimate of \(\theta\).
   (c) Calculate the maximum-likelihood estimate of \(\theta\).

**E PROBLEMS**

9. The vector \(\beta x\) was fitted by least squares in Example 1. Show that the squared length of the fitted vector is

\[ y'x\hat{\beta} = \sum y_i x_i \hat{\beta} \]

10. (continuation) Show that the squared length of the residual vector is

\[ \sum (y_i - \hat{\beta}x_i)^2 = \sum y_i^2 - \sum y_i x_i \hat{\beta} \]
11 Let \( y_{11}, \ldots, y_{1n}, y_{21}, \ldots, y_{2n} \) have location model \((\theta_1, \ldots, \theta_1, \theta_2, \ldots, \theta_2)\) with \((\theta_1, \theta_2)\) in \(\mathbb{R}^2\).

Find the least-squares estimate of \((\theta_1, \theta_2)\).

12 Let \( m_1 \) and \( m_2 \) be the first and second moments for a sample from the scale gamma distribution

\[
f(y|\rho, \beta) = \beta^{-\rho} \Gamma^{-1}(\rho) \rho^{-\rho} e^{-\rho y} y^{\rho-1}, \quad y > 0,
\]

with \((\rho, \beta)\) in \((0, \infty)^2\). Find the method-of-moments estimate of \((\rho, \beta)\).

13 Let \( y_1, \ldots, y_n \) have location given by \((m_1(\theta), \ldots, m_n(\theta))\). The least-absolute-deviations estimate is the value \( \hat{\theta} = \hat{\theta}(y) \) that minimizes

\[
\sum_{i=1}^{n} |y_i - m_i(\theta)|.
\]

Prove the following analog of Theorem 4: if \( y_1, \ldots, y_n \) are statistically independent and \( y_i - m_i(\theta) \) has the Laplace distribution

\[
f(v|\tau) = \frac{1}{2\tau} e^{-|v|/\tau}
\]

with \( \tau \) in \(\mathbb{R}^+\), then the maximum-likelihood estimate of \((\theta, \tau)\) is (a) the least-absolute-deviation estimate \( \hat{\theta} \) for \( \theta \), (b) the average absolute deviation

\[
\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} |y_i - m_i(\hat{\theta})|
\]

for \( \tau \). For example, see Problem 5.3.15.

9.2

LINEAR LEAST SQUARES

An application of least squares is concerned with a sequence of response values \( y_1, \ldots, y_n \) and a corresponding sequence of theoretical locations \( v_1(\theta), \ldots, v_n(\theta) \); the locations can be adjusted toward fitting the response values by varying a parameter \( \theta \). Least squares is particularly simple if the locations \( v_1(\theta), \ldots, v_n(\theta) \) are linear in the parameter \( \theta \). In this section we examine this linear least squares.

We introduce this least-squares theory here in order to have some substantial statistical examples to illustrate the inference methods in the next few chapters. The approach to inference in Chapters 9 and 10 is concerned with response values that deviate from a true location by error from a normal distribution; the approach in Chapter 11 covers the case with error from other distribution forms.

A EXAMPLES WITH LINEAR LOCATION

Consider a response \( y = (y_1, \ldots, y_n)' \) with location \( \nu(\theta) = (v_1(\theta), \ldots, v_n(\theta)) \) that is linear in the parameter \( \theta \). We consider briefly some examples that illustrate the great wealth of generality available with this linear model for location.

The general level of yield for trefoil on certain plots is approximately affine in the amount of nitrogen added; see Figure 9.3. Or, in general, consider a random system
where the location of the response is known to be an affine function of an input variable $x$ over a reasonable range. Then the location for the response would have the form

$$\beta_1 + \beta_2 x$$

with parameter $(\beta_1, \beta_2)$. Correspondingly, the location for the response vector $y$ with various input settings $x_1, \ldots, x_n$ would then be

$$\begin{align*}
v_1 &= \beta_1 + \beta_2 x_1, \\
&\vdots \\
v_n &= \beta_1 + \beta_2 x_n,
\end{align*}$$

$$\nu(\beta_1, \beta_2) = \beta_1 1 + \beta_2 x.$$ 

More generally, it may be necessary to allow for possible quadratic dependence on the input variable $x$. Then the location for the response would have the form

$$\beta_1 + \beta_2 x + \beta_3 x^2$$

with parameter $(\beta_1, \beta_2, \beta_3)$. Correspondingly, the location for the response vector $y$ with various input settings $x_1, \ldots, x_n$ would then be

$$\begin{align*}
v_1 &= \beta_1 + \beta_2 x_1 + \beta_3 x_1^2, \\
&\vdots \\
v_n &= \beta_1 + \beta_2 x_n + \beta_3 x_n^2,
\end{align*}$$

$$\nu(\beta_1, \beta_2, \beta_3) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

where we write

$$\begin{align*}
x_1 &= \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_n \\ 1 \end{pmatrix},
x_2 &= \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix},
x_3 &= \begin{pmatrix} x_1^2 \\ x_2^2 \end{pmatrix}.
\end{align*}$$

The general level of sales of a certain commodity has an annual cyclic effect. Or, in general, consider a random system where the location of the response is a simple cyclic function of a position or time variable $t$. If $t$ is conveniently scaled, the location for the response would have the form

$$\beta_1 + \beta_2 \sin t + \beta_3 \cos t = \beta_1 + \beta_4 \sin (t + \omega)$$

with parameter $(\beta_1, \beta_2, \beta_3, \beta_4)$. Correspondingly, the location for the response vector $y$ at various positions or times $t_1, \ldots, t_n$ would then be

$$\begin{align*}
v_1 &= \beta_1 + \beta_2 \sin t_1 + \beta_3 \cos t_1, \\
&\vdots \\
v_n &= \beta_1 + \beta_2 \sin t_n + \beta_3 \cos t_n,
\end{align*}$$

$$\nu(\beta_1, \beta_2, \beta_3) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$
where we write
\[ x_1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad x_2 = \begin{pmatrix} \sin t_1 \\ \vdots \\ \sin t_n \end{pmatrix}, \quad x_3 = \begin{pmatrix} \cos t_1 \\ \vdots \\ \cos t_n \end{pmatrix}. \]

The general level of output from a chemical process is approximately quadratic in terms of operating pressure and amount of a particular catalyst. Or, in general, consider a system with two input variables \( x_1 \) and \( x_2 \) and suppose that it is necessary to allow for possible quadratic dependence of the response level on the two variables. The location for the response would then have the form
\[ \beta_1 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_1 x_2 \]
with parameter \( (\beta_1, \ldots, \beta_6) \). Correspondingly, the location for the response vector \( y \) with various input settings \((x_{11}, x_{21}), \ldots, (x_{1n}, x_{2n})\) would be
\[ \nu(\beta_1, \ldots, \beta_6) = \beta_1 x_1 + \ldots + \beta_6 x_6, \]
where \( x_1, \ldots, x_6 \) record successive values of the variables \( 1, x_1, x_2, x_1^2, x_2^2, \) and \( x_1 x_2 \).

The general level of effectiveness of a language instruction program depends on which of three instructional methods is used. Or, in general, consider a system with a response \( y \) whose location depends on which of three treatments has been applied to the system. Let \( I_1, I_2, \) and \( I_3 \) be indicator variables for treatments 1, 2, and 3, respectively, and let \( \beta_1, \beta_2, \) and \( \beta_3 \) be the corresponding response locations. Then the location for the response can be presented as
\[ \beta_1 I_1 + \beta_2 I_2 + \beta_3 I_3 \]
with parameter \( (\beta_1, \beta_2, \beta_3) \). Correspondingly, the location for the response vector \( y \) with, say, two observations using treatment 1, two observations using treatment 2, and two observations using treatment 3 would be
\[
\begin{pmatrix}
\nu_1 \\
\nu_2 \\
\nu_3 \\
\nu_4 \\
\nu_5 \\
\nu_6
\end{pmatrix} = \begin{pmatrix}
1 & 0 & 0 \\
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 1
\end{pmatrix} \begin{pmatrix}
\beta_1 \\
\beta_2 \\
\beta_3 \\
0 \\
0 \\
0
\end{pmatrix}, \quad \nu(\beta_1, \beta_2, \beta_3) = \beta_1 I_1 + \beta_2 I_2 + \beta_3 I_3.
\]

where \( I_1, I_2, \) and \( I_3 \) are 1-vectors for the particular coordinates receiving the corresponding treatments.

**DEFINITION 1**

A model \( \nu(\beta) \) for the location of a response \( y \) is said to be linear if
\[ \nu(\beta) = X\beta = \beta_1 x_1 + \ldots + \beta_r x_r, \]
and \( \beta \in \mathbb{R}^r \); the vectors \( x_1, \ldots, x_r \) are assembled side by side to form the \( n \times r \) matrix \( X \).

For some matrix details, see Supplement Section 2.5.
B THE LEAST-SQUARES SOLUTION

Now consider an observed response $\mathbf{y} = (y_1, \ldots, y_n)^T$ and a linear location model

$$v(\beta) = X\beta,$$

where $X$ is an $n \times r$ matrix consisting of $r$ linearly independent vectors $x_1, \ldots, x_r$ (so that $X$ has rank $r$) and the parameter $\beta$ takes values in $\mathbb{R}^r$.

The set of possible response locations is given by

$$\mathcal{L}(X) = \{ b_1 x_1 + \cdots + b_r x_r : b_1, \ldots, b_r \in \mathbb{R} \};$$

it is the linear subspace generated by the column vectors of $X$; see Figure 9.4.

The least-squares estimate of $\beta$ is the value $\mathbf{b} = b(\mathbf{y})$ that minimizes

$$S^2 = \sum_{i=1}^{n} (y_i - b_1 x_{i1} - \cdots - b_r x_{ir})^2 = (\mathbf{y} - X\mathbf{b})^T(\mathbf{y} - X\mathbf{b}).$$

As pictured in Figure 9.4, $S^2$ is the squared distance from $\mathbf{y}$ to $X\mathbf{b}$; the method is to choose the value for $\mathbf{b}$ that gets $X\mathbf{b}$ closest to the observation $\mathbf{y}$.

The least-squares estimate can be obtained by differentiating with respect to $b_1, \ldots, b_r$ as in formula (9.1.2):

$$-\frac{1}{2} \frac{\partial S^2}{\partial b_1} = \sum_{i=1}^{n} x_{i1}(y_i - b_1 x_{i1} - \cdots - b_r x_{ir}) = 0,$$

$$\vdots$$

$$-\frac{1}{2} \frac{\partial S^2}{\partial b_r} = \sum_{i=1}^{n} x_{ir}(y_i - b_1 x_{i1} - \cdots - b_r x_{ir}) = 0.$$

These equations can be written in vector form as

$$x_1^T(\mathbf{y} - X\mathbf{b}) = 0$$

or

$$X^T(\mathbf{y} - X\mathbf{b}) = 0;$$

$$x_1^T(\mathbf{y} - X\mathbf{b}) = 0$$

FIGURE 9.4

The least-squares solution gives the projection $p_1$ of $\mathbf{y}$ on $\mathcal{L}(X)$; the residual is the projection $p_2$ of $\mathbf{y}$ on $\mathcal{L}^\perp(X)$. By orthogonality the sum $S_1^2 + S_2^2$ of the squared lengths of the projections is the squared length $\Sigma y_i^2$ of the original vector.
they state that the residual vector $\mathbf{y} - X\mathbf{b}$ is orthogonal to $x_1, \ldots, x_p$; that is, orthogonal to $L(X)$. Thus the least-squares solution gives the projection of $\mathbf{y}$ onto $L(X)$: it has an orthogonal residual $\mathbf{y} - X\mathbf{b}$, which lies in $L^\perp(X)$.

The equations (4) can be written as

(5) \[ X'y - X'X\mathbf{b} = 0 \]

and solved to give the least-squares solution,

(6) \[ \mathbf{b} = \mathbf{b}(\mathbf{y}) = (X'X)^{-1}X'y; \]

the coordinates of $\mathbf{b}$ are called **regression coefficients**, the regression coefficients of $\mathbf{y}$ on $x_1, \ldots, x_r$. Note that we assumed earlier that $X$ has rank $r$ and thus have that $X'X$ is a nonsingular symmetric $r \times r$ matrix with an inverse.

The least-squares fitted location is then given by

(7) \[ p_1 = X\mathbf{b}(\mathbf{y}) = X(X'X)^{-1}X'y = P_1\mathbf{y}; \]

this is the projection of $\mathbf{y}$ on $L(X)$; see Figure 9.4. Note that $p_1$ is obtained by applying the projection matrix

(8) \[ P_1 = X(X'X)^{-1}X' \]

to the vector $\mathbf{y}$. The squared length of the projection $p_1$ is easily obtained using equations (5) and (6):

(9) \[ S_1^2 = X'X\mathbf{b} = X'\mathbf{y}; \]

(10) \[ = y'X(X'X)^{-1}X'y. \]

The residual vector is given by

\[ \mathbf{p}_2 = \mathbf{y} - X\mathbf{b}(\mathbf{y}) = (I - X(X'X)^{-1}X')\mathbf{y} = \mathbf{P}_2\mathbf{y}, \]

where $I$ is the identity matrix. This is the projection of $\mathbf{y}$ on $L^\perp(X)$, for clearly its "residual" $X\mathbf{b}$ in $L(X)$ is orthogonal to $L^\perp(X)$. Note that $p_2$ is obtained by applying the projection matrix

(11) \[ \mathbf{P}_2 = I - X(X'X)^{-1}X' \]

to the vector $\mathbf{y}$. The squared length of the projection $\mathbf{p}_2$ is easily calculated. The orthogonality of $\mathbf{p}_1$ and $\mathbf{p}_2$ gives the decomposition

(12) \[ \mathbf{y}'\mathbf{y} = (\mathbf{p}_1 + \mathbf{p}_2)'(\mathbf{p}_1 + \mathbf{p}_2) = \mathbf{p}_1'\mathbf{p}_1 + \mathbf{p}_2'\mathbf{p}_2; \]

thus the squared length of $\mathbf{p}_2$ is

(13) \[ S_2^2 = \mathbf{y}'\mathbf{y} - S_1^2 = \mathbf{y}'\mathbf{y} - \mathbf{y}'X\mathbf{b}; \]

(14) \[ = \mathbf{y}'(I - X(X'X)^{-1}X')\mathbf{y}. \]

This decomposition into orthogonal components is presented in Table 9.4.

The least-squares method has been organized as a method of obtaining orthogonal components of a given vector. The geometry and algebra will allow us to examine some nontrivial and useful examples and will provide us with some basic results for the analysis of variance in Chapters 11 and 12.
TABLE 9.4
Components of vector $y$: corresponding squared lengths

<table>
<thead>
<tr>
<th>Subspace</th>
<th>Dimension</th>
<th>Component</th>
<th>Squared Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{E}(\chi)$</td>
<td>$r$</td>
<td>$p_1 = \chi b$</td>
<td>$S_1^2 = y' \chi b$</td>
</tr>
<tr>
<td>$\mathcal{E}^\perp(\chi)$</td>
<td>$n - r$</td>
<td>$p_2 = y - \chi b$</td>
<td>$S_2^2 = y'y - y' \chi b$</td>
</tr>
<tr>
<td></td>
<td>$\frac{n}{n}$</td>
<td>$y$</td>
<td>$y'y = \sum y_i^2$</td>
</tr>
</tbody>
</table>

C NUMERICAL EXAMPLE

The general level of yield for a chemical process is approximately affine in terms of pressure measured in deviation $x$ from some reference pressure. The yield $y$ and input pressure $x$ for five performances are as follows; also see Figure 9.5:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$-2$</th>
<th>$-1$</th>
<th>$0$</th>
<th>$1$</th>
<th>$2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>7</td>
<td>10</td>
</tr>
</tbody>
</table>

The location model is linear in terms of $1$ and $x$; thus it has the form

$$v(\beta_1, \beta_2) = \beta_1 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \beta_2 \begin{pmatrix} -2 \\ -1 \\ 0 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

with two vectors $1$ and $x$ recording the values of $1$ and $x$. We now fit this model using the least-squares formulas of Section B.

The vectors $1$ and $x$ have the following inner product matrix and inverse matrix:

$$X'X = \begin{pmatrix} 5 & 0 \\ 0 & 10 \end{pmatrix}, \quad (X'X)^{-1} = \begin{pmatrix} 1/5 & 0 \\ 0 & 1/10 \end{pmatrix}.$$

![FIGURE 9.5](image)

Fitting the line $y = b_1 + b_2 x$ to the five points as marked.
The vectors \( \mathbf{1} \) and \( x \) have the following inner product matrix with \( y \):

\[
X'Y = \begin{pmatrix} 25 \\ 20 \end{pmatrix}.
\]

The least-squares estimate of \( \beta \) is

\[
b = (X'X)^{-1}X'y = \begin{pmatrix} 1/5 & 0 \\ 0 & 1/10 \end{pmatrix} \begin{pmatrix} 25 \\ 20 \end{pmatrix} = \begin{pmatrix} 5 \\ 2 \end{pmatrix}.
\]

The fitted vector is the projection

\[
p_1 = Xb = \begin{pmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 5 \\ 2 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \\ 5 \\ 7 \\ 9 \end{pmatrix}
\]

onto \( \mathcal{L}(X) \), and the fitted line is \( y = 5 + 2x \). The squared length of the fitted vector is

\[
S_1^2 = y'Xb = (25, 20) \begin{pmatrix} 5 \\ 2 \end{pmatrix} = 165.
\]

The residual vector is the projection

\[
p_2 = y - Xb = \begin{pmatrix} 2 \\ 3 \\ 7 \\ 10 \end{pmatrix} - \begin{pmatrix} 1 \\ 3 \\ 7 \\ 9 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}
\]

onto \( \mathcal{L}^1(X) \). The squared length of the residual is

\[
S_2^2 = y'y - y'Xb = 171 - 165 = 6.
\]

The decomposition into orthogonal components is summarized in Table 9.5.

<table>
<thead>
<tr>
<th>Subspace</th>
<th>Dimension</th>
<th>Component</th>
<th>Squared Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathcal{L}^1(x) )</td>
<td>2</td>
<td>( 1, 3, 5, 7, 9 )</td>
<td>165</td>
</tr>
<tr>
<td>( \mathcal{L}^1(x) )</td>
<td>3</td>
<td>( 1, 0, -2, 0, 1 )</td>
<td>6</td>
</tr>
<tr>
<td>( \mathcal{L}^1(x) )</td>
<td>5</td>
<td>( 2, 3, 3, 7, 10 )</td>
<td>171</td>
</tr>
</tbody>
</table>

D  EXERCISES

1  The response vector \( y = (3, 7, 6, 5, 9)' \) was obtained with pressure settings given by \( x = (-2, -1, 0, 1, 2)' \); see the example in Section C.
(a) Calculate the projection \( p_1 \) and the squared length \( S_1^2 \); record the fitted line.
(b) Calculate the projection \( p_2 \) and the squared length \( S_2^2 \).
(c) Record the orthogonal decomposition in a table.

2  The general level of a yield \( y \) is approximately affine in terms of an input variable \( x \). The following data were obtained:
<table>
<thead>
<tr>
<th>x</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>7</td>
<td>10</td>
</tr>
</tbody>
</table>

Fit the linear location model \( m = \beta_1 1 + \beta_2 x \), where 1 and x record the values of 1 and x.
(a) Calculate the projection \( p_1 \) and the squared length \( S_1^2 \); record the fitted line.
(b) Calculate the projection \( p_2 \) and the squared length \( S_2^2 \).
(c) Record the orthogonal decomposition in a table.

The response vector \( y = (3, 7, 6, 5, 9)' \) was obtained with input settings \( x = (1, 2, 3, 4, 5)' \); see Exercise 2.
(a) Calculate the projection \( p_1 \) and the squared length \( S_1^2 \); record the fitted line.
(b) Calculate the projection \( p_2 \) and the squared length \( S_2^2 \).
(c) Record the orthogonal decomposition in a table.

For the example in Section C, calculate the projection matrices \( P_1 \) and \( P_2 \).

For the example in Exercise 2, calculate the projection matrices \( P_1 \) and \( P_2 \).

Let \( (y_{11}, y_{12}, y_{21}, y_{22}) \) have location model

\[
\begin{pmatrix}
  m_{11} \\
  m_{12} \\
  m_{21} \\
  m_{22}
\end{pmatrix} = \beta_0 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}.
\]

Find the least-squares estimate of \( (\beta_0, \beta_1) \) and record the projection \( p_1 \).

Let \( (y_{11}, \ldots, y_{1m}, y_{21}, \ldots, y_{2n}) \) have location model \( (\theta_1, \ldots, \theta_1, \theta_2, \ldots, \theta_2, \theta_3, \ldots, \theta_3) \) with \( (\theta_1, \theta_2, \theta_3) \) in \( \mathbb{R}^3 \); see the last example in Section A. Find the least-squares estimate of \( (\theta_1, \theta_2, \theta_3) \).
(a) Calculate the projection \( p_1 \) and the squared length \( S_1^2 \).
(b) Calculate the projection \( p_2 \) and the squared length \( S_2^2 \).

Suppose that we allow the possibility of quadratic dependence for the example in Section C.
Consider the response vector \( y = (2, 3, 3, 7, 10)' \) and the linear location model

\[
\begin{pmatrix}
  m \\
  \beta_1 \\
  \beta_2 \\
  \beta_3
\end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} -2 \\ -1 \\ 0 \\ 2 \end{pmatrix} + \beta_2 \begin{pmatrix} -1 \\ -1 \\ -2 \\ -2 \end{pmatrix} + \beta_3 \begin{pmatrix} 1 \\ 0 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ -1 \\ -1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.
\]

with three vectors recording the values of 1, \( x, x^2 - 2 \). Find the least-squares estimate of \( (\beta_1, \beta_2, \beta_3) \).
(a) Calculate the projection \( p_1 \) and the squared length \( S_1^2 \); record the fitted parabola \( y = b_1 + b_2 x + b_3(x^2 - 2) \).
(b) Calculate the projection \( p_2 \) and the squared length \( S_2^2 \).
(c) Record the orthogonal decomposition in a table.

Let \( Xb_1 \) and \( Xb_2 \) be two points in the location subspace \( \mathcal{L}(X) \). Show that the squared distance between them is given by the quadratic form

\[
(b_2 - b_1)'X'X(b_2 - b_1).
\]

E PROBLEMS

10 Consider the response \( y = (y_1, \ldots, y_n)' \) and the linear location model

\[
m = \beta_1 \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \beta_2 \begin{pmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix}.
\]
(a) Show that the least-squares estimate of \((\beta_1, \beta_2)\) is
\[
(16) \quad (b_1, b_2) = \left( \overline{y}, \frac{\left( \sum (x_i - \overline{x}) y_i \right)}{\left( \sum (x_i - \overline{x})^2 \right)} \right)
\]
and that the fitted line is \(y = \overline{y} + b_2(x - \overline{x})\).

(b) Show that the squared length of the residual is
\[
(17) \quad S_2^2 = \sum (y_i - b_1 - b_2(x_i - \overline{x}))^2 = \sum y_i^2 - n\overline{y}^2 - \frac{\left( \sum (x_i - \overline{x}) y_i \right)^2}{\sum (x_i - \overline{x})^2}.
\]

11 Consider the response \(y = (y_1, \ldots, y_n)'\) and the linear location model
\[m = \alpha \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \beta \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.\]

(a) Show that the least-squares estimate of \((\alpha, \beta)\) is
\[
(18) \quad (a, b) = (\overline{y} - bx, b),
\]
where \(b = (\sum (x_i - \overline{x}) y_i) / (\sum (x_i - \overline{x})^2)\), and that the fitted line is \(y = \overline{y} + b(x - \overline{x})\).

(b) Show that the squared length of the residual is
\[
(19) \quad S_2^2 = \sum (y_i - a - bx)^2 = \sum y_i^2 - n\overline{y}^2 - \frac{\left( \sum (x_i - \overline{x}) y_i \right)^2}{\sum (x_i - \overline{x})^2}.
\]

12 The projection \(P_1\) of \(y\) into \(L(X)\) is obtained by applying the projection matrix \(P_1 = X(X'X)^{-1}X'\) to the response \(y\); the projection \(P_2\) of \(y\) into \(L^2(X)\) is obtained by applying the projection matrix \(P_2 = I - X(X'X)^{-1}X'\) to the response \(y\).

(a) A matrix \(P\) is idempotent if \(PP = P\). Show that \(P_1\) and \(P_2\) are idempotent.

(b) An idempotent matrix is an orthogonal projection matrix if it is symmetric: \(P = P'\). Show that \(P_1\) and \(P_2\) are symmetric.

13 The least-squares solution for a linear model can be obtained by algebra rather than calculus. The sum of squares in formula (2) can be written as
\[
(20) \quad S^2 = (y - Xb)'(y - Xb)
\]
\[
= b'X'Xb - b'X'y - y'Xb + y'y
\]
\[
= (b - \cdots)'X'X(b - \cdots) + \cdots.
\]

Choose an entry for the blank locations so that the third line multiplied out produces the second line. Thus deduce (a) the least-squares solution for \(b\); (b) the squared length of the residual vector.

14 Consider a response \(y = (y_1, \ldots, y_n)\) and the linear location model \(X\beta\) with \(\beta\) in \(R^n\). If the various coordinates are expected to differ from the true location by differing amounts, then a weighted-least-squares method is sometimes used and is based on minimizing
\[
S^2 = (y - Xb)'M^{-1}(y - Xb),
\]
where \(M\) is some appropriate positive definite symmetric matrix. Use the method in Problem 13 to show that the weighted-least-squares estimate is
\[
b = (X'M^{-1}X)^{-1}X'M^{-1}y
\]
and the squared length of the residual is
\[
S_2^2 = y'M^{-1}y - y'M^{-1}Xb.
\]
Examples of this weighted least squares will be given later.
9.3

COMBINING UNBIASED ESTIMATORS

The theory of estimation is concerned with finding a function on the sample space whose values will cluster in a reasonable manner about the value of a particular parameter of interest. In an application there would be substantial hope, then, that the calculated value of such a function was a good approximation to the true value of the parameter.

The great wealth of functions on a sample space leads to a major problem in choosing one that is in some sense best. The concept that has been most fruitful both theoretically and in applications is that of unbiasedness, the central concept in this and the next two sections. There are other approaches, such as the decision theoretic, which provide extensions and alternatives, but none with the immediate relevance and usefulness of the unbiasedness approach. For example, with 10 measurements on the gravitational constant, should we use the average, \( \bar{y} \), or the median, \( \tilde{y} \), or some other sample function to obtain our combined estimate of the gravitational constant?

A UNBIASEDNESS

For the statistical model we let \( y \) be the response with values in a sample space \( \mathcal{S} \), and \( \theta \) be the parameter with values in a parameter space \( \Omega \). The parameter \( \theta \) indexes the various distributions for the response \( y \); in an application one of the parameter values is the true value which designates the distribution that is a reasonable approximation to the actual distribution of the response. We can now be somewhat more general than in Chapter 8: rather than record a density function for each parameter value, we merely assume that each parameter value designates a distribution on the sample space, a distribution that can be specified in some way—by density function, or by distribution function, or by probability measure.

Now suppose that we are interested in a particular real-valued parameter \( \beta(\theta) \) and that \( t(y) \) is some real-valued function that we are considering as an estimator for \( \beta(\theta) \). Note that \( t \) is a function from \( \mathcal{S} \) into \( R \) and \( \beta \) is a function from \( \Omega \) into \( R \). Then we define:

**DEFINITION 1**

A function \( t \) is an unbiased estimator for a parameter \( \beta \) if

\[
E(t(y) | \theta) = \beta(\theta)
\]

for each \( \theta \) in \( \Omega \).

As an example, consider a sample of \( n \) from some distribution on \( R \) that has a finite mean \( \mu \). The mean \( \mu \) can be viewed as a function of the distribution from which it is calculated. From formula (5.4.5) or (7.2.1) we have

\[
E(\bar{y}) = \mu
\]

for each distribution with finite mean. Thus in this general context \( \bar{y} \) is an unbiased estimator of \( \mu \).
As a second example, consider a sample of \( n \) from some distribution on \( R \) that has a finite variance \( \sigma^2 \). The variance \( \sigma^2 \) can be viewed as a function of the distribution from which it is calculated. From formula (5.4.9) or (7.2.13) we have

\[
E(s^2_0) = \sigma^2
\]

for each distribution with finite variance. Thus in this general context \( s^2_0 \) is an unbiased estimator of \( \sigma^2 \).

Now consider a much more specific model involving a sample from the normal(\( \mu, \sigma \)). Of course, \( \bar{Y} \) remains an unbiased estimator of \( \mu \), and \( s^2_n \) remains an unbiased estimator of \( \sigma^2 \). In fact, it is true generally: if \( t \) is an unbiased estimate of \( \beta \) for a model with \( \theta \) in \( \Omega \), then \( t \) is unbiased for the more specific model with \( \theta \) in \( \Omega^* \), where \( \Omega^* \subset \Omega \).

In Section 7.2 we gave a general justification for the use of unbiasedness: essentially that the values of \( t \) are sometimes larger than the true \( \beta \) and sometimes smaller than the true \( \beta \) but that the mean value is equal to \( \beta \): that is, the distribution of \( t \) is located in a certain sense at the parameter value corresponding to the underlying response distribution.

The definition of unbiasedness extends immediately to cover a vector estimator \( t(y) = (t_1(y), \ldots, t_r(y))' \) for a vector parameter \( \beta(\theta) = (\beta_1(\theta), \ldots, \beta_r(\theta))' \):

**DEFINITION 2**

A function \( t \) is an unbiased estimator for a parameter \( \beta \) if

\[
E(t(y) | \theta) = \beta(\theta)
\]

for each \( \theta \) in \( \Omega \); that is, if

\[
E(t_u(y) | \theta) = \beta_u(\theta), \quad u = 1, 2, \ldots, r,
\]

for each \( \theta \) in \( \Omega \).

As an example consider a sample of \( n \) from a distribution on \( R \) that has a finite mean \( \mu \) and finite variance \( \sigma^2 \). Then from the early discussions we have that (\( \bar{Y}, s^2_0 \)) is an unbiased estimator for (\( \mu, \sigma^2 \)).

**B COMPARISON OF UNBIASED ESTIMATORS**

The property of unbiasedness says that the distribution of an estimator is located in a certain sense at the value of the parameter being estimated. We now consider the comparison of unbiased estimators. For this we would think immediately of the scale of the distribution and examine, say, the variance or perhaps the reciprocal variance or precision.

Sometimes it is convenient to have a numerical comparison of estimators. For this we give a definition of the comparative efficiency of two unbiased estimators of a parameter.
DEFINITION 3

The efficiency of an unbiased estimator $t_1$ of $\beta(\theta)$ with respect to an unbiased estimator $t_2$ of $\beta(\theta)$ is

$$
\text{Eff}(t_1, t_2) = \frac{1/\text{var}(t_1|\theta)}{1/\text{var}(t_2|\theta)} = \frac{\text{precision of } t_1}{\text{precision of } t_2} = \frac{\text{var}(t_2|\theta)}{\text{var}(t_1|\theta)}.
$$

The efficiency is a function of $\theta$—but for a lot of common examples it turns out to be independent of $\theta$. Some other aspects of efficiency are examined in Section 9.4.

EXAMPLE 1

The normal$(\mu, \sigma)$: Let $y$ be a sample from the normal$(\mu, \sigma)$ with $(\mu, \sigma)$ in $R \times R^+$. We have examined the sample average $\bar{y}$ as a natural estimator of $\mu$: we have $E(\bar{y}|\mu, \sigma) = \mu, \quad \text{var}(\bar{y}|\mu, \sigma) = \sigma^2/n$.

The sample median $\tilde{y}$ is also a very simple function on the sample space, and we would naturally associate it with the center of the distribution (for an even sample let $\bar{y}$ be the average of the 2 central values); we have $E(\tilde{y}|\mu, \sigma) = \mu, \quad \text{var}(\tilde{y}|\mu, \sigma) \approx \pi \sigma^2/2n$; the mean follows from the symmetry of the distribution of $\tilde{y} - \mu$; the approximate variance is indicated by Problem 6.1.13. Thus the efficiency of $\tilde{y}$ with respect to $\bar{y}$ is

$$
\text{Eff}(\tilde{y}, \bar{y}) = \frac{2\pi/\sigma^2}{\pi \sigma^2/2n} = \frac{2}{\pi} \approx 0.64.
$$

The variance of a reasonable estimator of a parameter is often proportional to the reciprocal of the sample size; this happens commonly with models that have a fixed region of positive density. We then have a simple interpretation for efficiency. Suppose we compare the use of the median $\tilde{y}$ for a sample of size 100 with the use of the mean $\bar{y}$ for a sample of size 64. We then have

$$
E(\tilde{y}) = \mu, \quad \text{var} (\tilde{y}) = \pi \sigma^2/200 = \sigma^2/64,
$$

$$
E(\bar{y}) = \mu, \quad \text{var} (\bar{y}) = \sigma^2/64.
$$

Thus the two estimators have the same location and the same scale, and thus in a sense are equivalent. And we can then, in this context, say that $\tilde{y}$ uses 64 percent of a sample by comparison with $\bar{y}$.

The comparison of vector estimates is somewhat more complicated. Each coordinate of an estimator has a variance, and in addition each pair of coordinates has a covariance. Consider an unbiased estimator $t$ of the parameter $\beta(\theta)$, and let $\Sigma(\theta)$ be the variance matrix:

$$
\Sigma(\theta) = \begin{pmatrix}
\sigma_{11} & \cdots & \sigma_{1r} \\
\vdots & & \vdots \\
\sigma_{r1} & \cdots & \sigma_{rr}
\end{pmatrix}.$$

(4)
then $\sigma_{11}$ is the variance of the first coordinate $t_1$, ..., $\sigma_{rr}$ is the variance of the $r$th coordinate $t_r$, and, for example, $\sigma_{1r}$ is the covariance of $t_1$ and $t_r$.

Now suppose that we consider a real-valued parameter $l_1\beta_1(\theta) + \cdots + l_r\beta_r(\theta) = \beta(\theta)$, where $l_1, \ldots, l_r$ are given real numbers. Clearly the estimator $l_1t_1 + \cdots + l_rt_r = \beta'$ is an unbiased estimator of $\beta(\theta)$; its variance can be obtained from (5.3.18).

(5) $\text{var} (\beta') = \beta' \Sigma(\theta) \beta'$

which is a quadratic form with matrix $\Sigma(\theta)$.

Consider the comparison of two unbiased estimators $t_1$ and $t_2$ for the parameter $\beta(\theta)$. We will say that $t_1$ is better than $t_2$ if

(6) $\Sigma_{22}(\theta) - \Sigma_{11}(\theta)$

is positive semidefinite. If (6) holds, we will say that the matrix $\Sigma_{11}(\theta)$ is smaller than the matrix $\Sigma_{22}(\theta)$. For some background detail, see Supplement Section 2.5C.

If $t_1$ is better than $t_2$ then the relation (6) says that the first coordinate of $t_1$ has smaller ($\leq$) variance than the first coordinate of $t_2$, ..., and that the $r$th coordinate of $t_1$ has smaller ($\leq$) variance than the $r$th coordinate of $t_2$. In addition, the relation (6) together with (5) shows that the variance of $\beta't_1$ is smaller ($\leq$) than the variance of $\beta't_2$ for any given vector $\beta'$; note that these are unbiased estimates of $\beta(\theta)$ and use details from Section 2.5C.

### C COMBINING UNBIASED ESTIMATORS

Suppose that we have two unbiased estimators $t_1$ and $t_2$ for the real-valued parameter $\beta(\theta)$ and wish to examine linear combinations as possible estimators for $\beta(\theta)$.

**THEOREM 4**

If $t_1$ and $t_2$ are unbiased estimators of $\beta(\theta)$, then

(a) A linear combination of $t_1$ and $t_2$ that is unbiased for $\beta(\theta)$ has the form

(7) $at_1 + (1 - a)t_2$,

where the sum of the weights is 1.

(b) Among such linear unbiased estimators the variance at the parameter value $\theta_0$ is minimized by having

(8) $\frac{a}{1 - a} = \frac{1/\text{var} (t_1 | \theta_0) - \text{cov} (t_1, t_2 | \theta_0)}{1/\text{var} (t_2 | \theta_0) - \text{cov} (t_1, t_2 | \theta_0)}$

the weights are proportional to the reciprocal of excess of variance over covariance.

(c) If the estimators are independent or uncorrelated, then the variance at the parameter value $\theta_0$ is minimized by having

(9) $\frac{a}{1 - a} = \frac{1/\text{var} (t_1 | \theta_0)}{1/\text{var} (t_2 | \theta_0)}$

the weights are proportional to the precisions.
In the uncorrelated case the precision (at $\theta_0$) of this best linear unbiased estimator is

$$\frac{1}{\text{var}(t_1 | \theta_0)} + \frac{1}{\text{var}(t_2 | \theta_0)};$$

see Figure 9.6. We consider two examples before recording the proof.

**EXAMPLE 2**

Two instruments are available for measuring chlorophyll content. Five measurements with the first instrument gave an estimate $t_1 = 3.17$; the estimating procedure is unbiased and has $\text{var}(t_1) = (0.11)^2$; five measurements with the second instrument gave an estimate $t_2 = 3.11$; the estimating procedure is unbiased and has $\text{var}(t_2) = (0.07)^2$. By formula (9) the best linearly combined estimate is

$$t = \frac{1}{0.0121} 3.17 + \frac{1}{0.0049} 3.11 \frac{1}{0.0121 + 0.0049} = 3.127;$$

and by formula (10) it has variance

$$\text{var}(t) = \left(\frac{1}{0.0121} + \frac{1}{0.0049}\right)^{-1} = 0.0035 \approx (0.06)^2.$$
Proof of Theorem 4  (a) Consider a linear combination \( at_1 + bt_2 \) that is unbiased for \( \beta(\theta) \); then

\[
E(at_1(y) + bt_2(y)) = a\beta(\theta) + b\beta(\theta) = \beta(\theta),
\]

and hence \( b = 1 - a \), provided, of course, that we exclude the ultratrvivial \( \beta(\theta) \equiv 0 \).

(b) The variance of \( at_1 + (1 - a)t_2 \) at \( \theta_0 \) is

\[
a^2 \text{var}(t_1|\theta_0) + 2a(1 - a) \text{cov}(t_1, t_2|\theta_0) + (1 - a)^2 \text{var}(t_2|\theta_0).
\]

Setting the derivative with respect to \( a \) equal to zero gives

\[
2a(\text{var}(t_1|\theta_0) - \text{cov}(t_1, t_2|\theta_0)) = 2(1 - a)(\text{var}(t_2|\theta_0) - \text{cov}(t_1, t_2|\theta_0)),
\]

which is equivalent to the quoted condition. Alternatively, we can complete the square as suggested in Problem 9.2.13.

There is an analogous version of Theorem 4 which covers vector estimators. For simplicity we give here the special version for uncorrelated vector estimators:

THEOREM 5

If \( t_1 \) and \( t_2 \) are uncorrelated unbiased estimators for \( \beta(\theta) \) with variance matrices \( \Sigma_{11}(\theta) \) and \( \Sigma_{22}(\theta) \), then

(a) Any linear combination of \( t_1 \) and \( t_2 \) that is unbiased for \( \beta \) has the form

\[
At_1 + (l - A)t_2 \tag{11}
\]

where \( A \) is an \( r \times r \) matrix and \( l \) is the \( r \times r \) identity matrix [provided that \( \beta_i(\theta) \), \( \ldots, \beta_i(\theta) \) are linearly independent].

(b) Among such linear unbiased estimators the variance matrix at the parameter value \( \theta_0 \) is minimized by the estimator

\[
(\Sigma_{11}^{-1}(\theta_0) + \Sigma_{22}^{-1}(\theta_0))^{-1}(\Sigma_{11}^{-1}(\theta_0)t_1 + \Sigma_{22}^{-1}(\theta_0)t_2), \tag{12}
\]

the weighting matrices are "proportional" to the inverse variance matrices or the precision matrices.

(c) The precision matrix (at \( \theta_0 \) of this best linear unbiased estimator is

\[
\Sigma_{11}^{-1}(\theta_0) + \Sigma_{22}^{-1}(\theta_0) \tag{13}
\]

the sum of the component precisions.

The proof follows as a special case of the generalized Gauss–Markov Theorem in Problem 11.

D THE GAUSS–MARKOV THEOREM

FOR LINEAR LOCATION

The Gauss–Markov theorem is concerned with unbiased estimation for models that have the linear location structure discussed in Section 9.2. For many applications it
provides theoretical support for the use of the least-squares and weighted-least-squares methods.

Consider a response vector $y = (y_1, \ldots, y_n)'$ with a linear location model $X\beta$ as discussed in Section 9.2A. Specifically, consider

$$
\begin{pmatrix}
  y_1 \\
  \vdots \\
  y_n
\end{pmatrix} =
\begin{pmatrix}
  x_{11} & \cdots & x_{1r} \\
  \vdots & \ddots & \vdots \\
  x_{n1} & \cdots & x_{nr}
\end{pmatrix}
\begin{pmatrix}
  \beta_1 \\
  \vdots \\
  \beta_r
\end{pmatrix} +
\begin{pmatrix}
  \nu_1 \\
  \vdots \\
  \nu_n
\end{pmatrix},
$$

(14) $y = X\beta + \nu$.

where $X$ is an $n \times r$ matrix of full rank $r$ and where the variation $\nu = (\nu_1, \ldots, \nu_n)'$ has mean value $0$. For the present we consider the case where the coordinates have common variance $\sigma^2$ and zero covariance

$$
\text{VAR}(\nu) = \begin{pmatrix}
  \sigma^2 & 0 & \cdots & 0 \\
  0 & \sigma^2 & \cdots & 0 \\
  \vdots & \ddots & \ddots & \vdots \\
  0 & \cdots & 0 & \sigma^2
\end{pmatrix} = \sigma^2 I.
$$

As an example, consider a response that has linear location in terms of an input variable $x$ and has common variance for all values of $x$. Then for $n$ observations, we have

$$
y_1 = \beta_1 + \beta_2 x_1 + \nu_1, \\
\vdots \\
y_n = \beta_1 + \beta_2 x_n + \nu_n,
$$

where the $\nu$'s have common variance $\sigma^2$ and zero covariance.

There is some interest in thinking of the $\nu$'s as providing indirect estimators for the parameters. We have that

$$
y_1 \quad \text{is unbiased estimator for } \beta_1 + \beta_2 x_1 \\
y_2 \quad \text{is unbiased estimator for } \beta_1 + \beta_2 x_2 \\
\frac{y_2 - y_1}{x_2 - x_1} \quad \text{is unbiased estimator for } \beta_2 \\
\frac{x_2 y_1 - x_1 y_2}{x_2 - x_1} \quad \text{is unbiased estimator for } \beta_1
$$

and similarly for any pair with $x_j - x_i \neq 0$. Thus in looking for best estimates of $\beta_1$ and $\beta_2$, we can think in terms of combining or taking linear combinations of the indirect estimators $y_1, \ldots, y_n$.

**THEOREM 6**

(a) A linear estimator $Ay$ based on an $r \times n$ matrix $A$ is an unbiased estimator of $\beta$ if $AX = I$.

(b) Among such linear unbiased estimators the variance matrix is smallest for the least-squares estimator

$$
b = (X'X)^{-1}X'y.
$$

(c) The variance matrix of this best linear unbiased estimator is

$$
(X'X)^{-1}\sigma^2.
$$
The proof is recorded in Section E. For the definition of smaller matrices see Supplement Section 2.5C.

**EXAMPLE 4**

Consider the example concerning the yield from a chemical process as discussed in Section 9.2C. We now suppose that the variation has independent coordinates with common variance $\sigma^2$. The observed response is $y = (2, 3, 3, 7, 10)'$ and the model is

$$
\begin{pmatrix}
  y_1 \\
  y_2 \\
  y_3 \\
  y_4 \\
  y_5
\end{pmatrix} = \beta_1 \begin{pmatrix}
  1 \\
  1 \\
  1 \\
  1 \\
  1
\end{pmatrix} + \beta_2 \begin{pmatrix}
  -2 \\
  -1 \\
  0 \\
  1 \\
  2
\end{pmatrix} + \begin{pmatrix}
  v_1 \\
  v_2 \\
  v_3 \\
  v_4 \\
  v_5
\end{pmatrix}.
$$

Theorem 5 is concerned with estimators for $(\beta_1, \beta_2)$ that are linearly constructed from the response vector; the theorem says that the least-squares estimator $b = (b_1, b_2)'$ has the minimum variance matrix. From Section 9.2C we have the estimate

$$
b = \begin{pmatrix}
  5 \\
  2
\end{pmatrix}.
$$

The variance matrix of this estimator is

$$(X'X)^{-1}\sigma^2 = \begin{pmatrix}
  1/5 & 0 \\
  0 & 1/10
\end{pmatrix}\sigma^2.$$

Thus $5$ is the calculated value of the unbiased estimator $b_1$ for $\beta_1$; this estimator has variance $\sigma^2/5$, which is the minimum among linear unbiased estimators of $\beta_1$. And $2$ is the calculated value of the unbiased estimator $b_2$ for $\beta_2$; this estimator has variance $\sigma^2/10$, which is the minimum among linear unbiased estimators of $\beta_2$.

We will see later that an unbiased estimator for $\sigma^2$ is given by

$$
s^2 = \frac{S_2^2}{n - r},
$$

where $S_2$ is the squared length of the residual; the value here of $s^2 = 6/3 = 2$.

**E PROOF OF THE GAUSS–MARKOV THEOREM**

First suppose that $Ay$ is unbiased estimator for $\beta$. Then

$$E(Ay) = AE(y) = AX\beta = \beta;$$

and it follows that $AX = I$, as asserted in part (a) of the theorem.

Now consider the variance matrix for the estimators $Ay$. From formula (5.3.17) we have

$$VAR(Ay) = A\sigma^2AA' = \sigma^2AA'.
$$

Thus our problem is to examine matrices $A$ that satisfy $AX = I$ in search of one that minimizes the matrix (18).

We follow the method of completing the square as suggested for a simpler
context in Problem 9.2.13. The matrix $AA'$ has already the form of a "completed square": we rearrange it, however, as a "completed square" in a way that makes essential use of the constraint $AX = I$ satisfied by $A$,

$$AA' = (A - \cdots)(A - \cdots)' + DD',$$

where $DD'$ is an inner product matrix and the entry in the parentheses needs to use $AX = I$. We try the entry $CX'$, which produces the products $AX$ and $X'AX'$ and has the yet undetermined matrix $C$:

$$AA' = (A - CX')(A - CX')' + DD'$$

$$= AA' - AXC' - CX'A' + CX'XC' + DD'$$

$$= AA' - C' - C + CX'XC' + DD'.$$

The middle three terms collapse to one term if we choose $C = (X'X)^{-1}$:

$$AA' = AA' - (X'X)^{-1} - (X'X)^{-1} + (X'X)^{-1} + DD'.$$

This determines $DD'$ and we obtain

$$AA' = (A - (X'X)^{-1}X')(A - (X'X)^{-1}X')' + (X'X)^{-1}.

Note that $(X'X)^{-1}$ is positive definite and that

$$(A - (X'X)^{-1}X')(A - (X'X)^{-1}X')'$$

is an inner product matrix and is thus positive semidefinite; for some background details see Supplement Section 2.5C. From formula (21) we then see that the matrix $AA'$ subject to $AX = I$ can be made smallest by taking the inner product matrix (22) to the zero matrix; this choice for $A = (X'X)^{-1}X'$ gives

$$Ay = (X'X)^{-1}X'y,$$

which is the least-squares estimate as quoted in part (b) of the theorem.

The least-squares estimate makes the first term on the right of (21) equal to the zero matrix and thus $A'A$ equal to $(X'X)^{-1}$; it follows that the variance matrix for the least-squares estimate is

$$\sigma^2 AA' = \sigma^2 (X'X)^{-1},$$

as quoted in part (c) of the theorem.

---

**F EXERCISES**

1. A first determination of a distance by a surveyor has standard deviation 0.021; the observed value is 257.341; an independent determination has standard deviation 0.037; the observed value 257.361. Determine the best linearly combined estimate.

2. Let $\bar{y}_1 = \Sigma y_1 / m$ be an estimate of $\mu$ based on a first sample of $m$ and $\bar{y}_2 = \Sigma y_2 / n$ be an estimate of $\mu$ based on a second sample of $n$ from a distribution with finite variance. Derive the linear combination of $\bar{y}_1$ and $\bar{y}_2$ that has minimum variance.

3. Let $t$ and $u$ be independent estimates of $\beta = (\beta_1, \beta_2)'$.

(a) If the variance matrices are

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \sigma^2 \text{ and } \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \sigma^2,$$

respectively, determine the best linearly combined estimate.
(b) If the variance matrices are
\[
\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \sigma^2 \quad \text{and} \quad \begin{pmatrix} 0.95 & 0 \\ 0 & 0.95 \end{pmatrix} \sigma^2,
\]
respectively, determine the best linearly combined estimate.

4. Consider the linear model \( y = \beta x + \nu \), where the \( \nu \)'s are independent and each has zero mean and variance \( \sigma^2 \). Record the best linear unbiased estimate of \( \beta \); determine its variance from Theorem 6. See Example 9.1.1.

5. Consider the linear model \( y = (y_1, \ldots, y_n; y_{1n}, \ldots, y_{2n}) = (\theta_1, \ldots, \theta_1; \theta_2, \ldots, \theta_2)' + \nu \), where the \( \nu \)'s are independent and each has zero mean and variance \( \sigma^2 \). Show that the best linear unbiased estimate of \( (\theta_1, \theta_2) \) is \( (\tilde{\gamma}_1, \tilde{\gamma}_2) \); determine its variance matrix from Theorem 6. See Problem 9.1.11.

6. Consider the linear model
\[
y = \alpha \left( \begin{array}{c} 1 \\ 1 \\ \vdots \\ 1 \end{array} \right) + \beta \left( \begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_n \end{array} \right) + \nu,
\]
where the \( \nu \)'s are independent and each has zero mean and variance \( \sigma^2 \). Determine the variance matrix of the best linearly unbiased estimate of \( (\alpha, \beta) \); use Theorem 6 and recall Problem 9.2.11.

**G PROBLEMS**

7. Consider the linear model
\[
y = X\beta + \nu,
\]
where \( \nu \) is a sample from a normal \( (0, \sigma^2) \) with \( (\beta, \sigma^2) \) in \( \mathbb{R}^r \times \mathbb{R}^+ \); assume that \( X \) has full rank \( r \). Use Theorem 9.1.4 and the results from Section 9.2 to show that the maximum-likelihood estimate of \( \beta \) is
\[
\hat{\beta} = (X'X)^{-1}X'y
\]
and the maximum-likelihood estimate of \( \sigma^2 \) is
\[
\hat{\sigma}^2 = \frac{1}{n} (y'y - y'X\hat{\beta}).
\]
Use Theorem 6 to show that \( \hat{\beta} \) has mean \( \beta \) and variance matrix \( (X'X)^{-1}\sigma^2 \).

8. Let \( t_1, \ldots, t_k \) be independent unbiased estimators for \( \theta \) with known variances \( \sigma_1^2, \ldots, \sigma_k^2 \), respectively. Use Theorem 6 to show that
\[
(\sigma_1^{-2} + \cdots + \sigma_k^{-2})^{-1}(\sigma_1^{-2}t_1 + \cdots + \sigma_k^{-2}t_k)
\]
is the linear unbiased estimate with minimum variance
\[
(\sigma_1^{-2} + \cdots + \sigma_k^{-2})^{-1}.
\]

9. A large-sample property closely related to unbiasedness is consistency. Consider an estimator \( t_n \) for \( \beta(\theta) \) based on a sample of \( n \). Then \( \{t_n\} \) is said to be a consistent estimator if
\[
\lim_{n\to\infty} P(\beta(\theta) - \epsilon \leq t_n(\gamma) \leq \beta(\theta) + \epsilon) = 1
\]
for all \( \epsilon > 0 \), that is, if \( t_n(\gamma) \overset{P}{\to} \beta(\theta) \). Show that if \( t_n \) is unbiased for each \( n \) and \( \lim_{n\to\infty} \text{var}(t_n|\theta) = 0 \) for each \( \theta \), then \( \{t_n\} \) is a consistent estimator.

10. A proof of Theorem 5: Let \( d = t_1 - t_2 \).
(a) Show that \( E(d) = 0 \).
Sec. 9.4: The Information Inequality for Unbiased Estimators

(b) Show that the covariance matrix of $\mathbf{u} = (\Sigma_{11}^{-1} \mathbf{t}_1 + \Sigma_{22}^{-1} \mathbf{t}_2)$ with $\mathbf{d}$ is the zero matrix.
(c) Show that $A\mathbf{u} + B\mathbf{d}$ generates all linear combinations of $\mathbf{t}_1$ and $\mathbf{t}_2$.
(d) Determine conditions so that $A\mathbf{u} + B\mathbf{d}$ is an unbiased estimator of $\beta$.
(e) Find the unbiased estimator with minimum variance matrix.

11 Generalized Gauss-Markov theorem: Consider a response $\mathbf{y} = X\beta + \mathbf{v}$, where $X$ is an $n \times r$ matrix of full rank $r$ and the variation $\mathbf{v}$ has mean $\mathbf{0}$ and variance matrix $\sigma^2\mathbf{M}$ with $(\beta, \sigma^2)$ in $R^r \times R^1$; the matrix $\mathbf{M}$ is given. Prove:

**THEOREM 7**

(a) A linear estimator $\mathbf{Ay}$ based on an $r \times n$ matrix $A$ is an unbiased estimator of $\beta$ if $AX = I$.
(b) Among such linear unbiased estimators the variance matrix is smallest for the weighted-least-squares estimator

$$\mathbf{b} = (X'M^{-1}X)^{-1}X'M^{-1}\mathbf{y}$$

as given in Problem 9.2.14.
(c) The variance matrix of this best linear unbiased estimator is

$$(X'M^{-1}X)^{-1}\sigma^2.$$ 

**Method:** Modify the pattern of proof in Section E.

12 Use Theorem 7 to prove Theorem 5.

9.4

The Information Inequality for Unbiased Estimators

Least squares and the linear combination of estimators are methods primarily concerned with location parameters for response variables. In this section we examine more general parameters, including those that refer to the shape or form of the distribution, and we obtain a lower bound for the variance of unbiased estimates. The inequality that expresses this lower bound was developed by Frechét, Fisher, Cramér, and Rao; it is sometimes called the information inequality, as it involves the Fisher information function defined in the large-sample Section 8.6.

The information inequality can sometimes be used in reverse to construct good unbiased estimates; but the success of this procedure is limited to the special exponential models discussed in Section 8.5.

The information inequality is essentially a local result on the parameter space. Accordingly, to develop the inequality, we suppose that we are in a situation where we know that the true parameter value is in some small interval $(\theta_0 \pm \delta)$ and we require that the estimate be unbiased only for this small interval. The application of the inequality, however, is not restricted to such special situations.
A PROPERTIES OF THE SCORE FUNCTION

In Chapter 8 we saw how the likelihood function was a basic ingredient to the process of statistical inference. In particular, the slope of the log-likelihood function, the score function, had several important properties. The score function is the basis for the information inequality; we first collect some properties of the score function.

For notation let $y$ be the full response under investigation with sample space $S$, and let $\theta$ be the parameter with real values in the parameter space $\Omega = \mathbb{R}$. We assume that the statistical model

$$\{f(\cdot | \theta) : \theta \in \Omega\}$$

is given by density functions and satisfies the regularity conditions as discussed in Section 8.1.

The score function $S(y|\cdot)$ was defined in Section 8.1D:

$$S(y|\theta) = \frac{\partial}{\partial \theta} \ln f(y|\theta) = \frac{\partial}{\partial \theta} l(y|\theta).$$

The following properties of the score function were developed in Section 8.6B on the basis of Assumption 8.6.1:

1. $E(S(y|\theta)|\theta) = 0$,
2. $\text{var}(S(y|\theta)|\theta) = J(\theta) = -E(S'(y|\theta)|\theta)$.

The first equation says that the mean slope of the log-likelihood function is zero at the true parameter value. The second says that the variance of the slope at the true parameter value is related to the mean curvature at the true parameter value. Assumption 8.6.1 requires the density function to depend on $\theta$ in a very respectable way. The function $J(\theta)$ is the Fisher information function for the full response $y$ and is defined by the left side of (2).

We now derive a third result closely connected with those just described. Let $t(y)$ be a function with mean value $\beta(\theta)$. Again we assume that differentiation with respect to $\theta$ can be carried through the sample-space summation or integration sign; we record the analysis for the integration case; the summation case is parallel. In the pattern of Section 8.6B we differentiate the integral

$$\int t(y)f(y|\theta) \, dy = \beta(\theta),$$

with respect to $\theta$:

$$\int t(y) \frac{\partial}{\partial \theta} f(y|\theta) \, dy = \int t(y)S(y|\theta)f(y|\theta) \, dy = \beta'(\theta).$$

Thus we obtain

$$(3) \quad E(t(y)S(y|\theta)|\theta) = \frac{d}{d\theta} E(t(y)|\theta).$$

Note that (3) reduces to (1) by setting $t(y) \equiv 1$. In summary: derivative outside $E$ gives score inside $E$. 


THE INFORMATION INEQUALITY

For the moment let us picture ourselves in an application where we know that the true parameter value \( \theta \) is very close to the value \( \theta_0 \). And let us suppose that we are prepared to use unbiased estimation even though an unbiased estimator may well give values far outside a small interval \( (\theta_0 \pm \delta) \).

An estimator \( t(y) \) that is unbiased generally has the properties

\[
E(t(y) | \theta) = \theta = \theta_0 + (\theta - \theta_0),
\]

where we have reexpressed the right-hand side relative to the value \( \theta_0 \). From this we are led to the following definition of unbiasedness for \( \theta \) close to \( \theta_0 \):

**DEFINITION 1**

\( t \) is locally unbiased\( (\theta_0) \) for the parameter \( \theta \) if

\[
E(t(y) | \theta) = \theta_0 + (\theta - \theta_0) + o(\theta - \theta_0)
\]

or, equivalently, if

\[
E(t(y) | \theta_0) = \theta_0.
\]

\[
\frac{\partial}{\partial \theta} E(t(y) | \theta) \bigg|_{\theta_0} = 1.
\]

This definition says that formula (4) holds to a first derivative approximation for \( \theta \) near \( \theta_0 \); see Figure 9.7. Note that \( o(\theta - \theta_0) \) represents a remainder that goes to zero as a proportion of \( \theta - \theta_0 \).

We have, of course, that if \( t \) is unbiased generally, then it is locally unbiased\( (\theta_0) \) for each \( \theta_0 \) and that if \( t \) is locally unbiased\( (\theta_0) \) for each \( \theta_0 \), then it is unbiased generally.

Now let \( t \) be a locally unbiased\( (\theta_0) \) estimator for \( \theta \). Then formulas (3) and (7) give

\[
E(t(y) S(y | \theta_0) | \theta_0) = 1.
\]

---

**FIGURE 9.7**

\( t_1 \) is unbiased for \( \theta \); \( t_2 \) is locally unbiased\( (\theta_0) \) for \( \theta \).
We recall that covariance is mean product minus product means. Thus using formula (1) we obtain

(8) \[ \text{cov} \left( t(y), S(y|\theta_0) | \theta_0 \right) = 1. \]

We now record the information inequality:

\underline{THEOREM 2}

Under the regularity assumptions in Section 8.6B, if \( t \) is locally unbiased(\( \theta_0 \)) for \( \theta \), then

(9) \[ \text{var} \left( t(y) | \theta_0 \right) \geq J^{-1}(\theta_0) \]

with equality if and only if \( t(y) \) and \( S(y|\theta_0) \) are affinely related; that is, \( S(y|\theta_0) = a + ct(y) \) for some \( a, c \).

**Proof** We apply the covariance inequality (5.3.24) to formula (8):

\[ \text{var} \left( t(y) | \theta_0 \right) \text{var} \left( S(y|\theta_0) | \theta_0 \right) \geq 1^2; \]

rearrangement gives formula (9).

The results in the theorem apply immediately, of course, to an estimate that is unbiased generally.

\underline{EXAMPLE 1}

The location normal: Let \((y_1, \ldots, y_n)\) be a sample from the normal(\(\theta, \sigma_0\)) with \(\theta\) in \(R\). Then we have

\[ S(y|\theta_0) = \frac{\bar{y} - \theta_0}{\sigma_0^2/n}, \quad J(\theta) = \frac{n}{\sigma_0^2} \]

as derived in Section 8.6B. Consider an unbiased estimator \( t \) of \( \theta \), or, more generally, a locally unbiased(\( \theta_0 \)) estimator of \( \theta \). Then from Theorem 2 we have

\[ \text{var} \left( t(y) | \theta_0 \right) \geq \frac{\sigma_0^2}{n} \]

with equality if and only if \( t(y) \) and \( S(y|\theta_0) \) are affinely related.

Our usual estimator of \( \theta \) is the sample average \( \bar{y} \). It is unbiased and has variance \( \sigma_0^2/n \), which is the lower bound in the inequality. The inequality then says that we cannot find a better estimate, better in the sense of smaller variance.

Thus in summary we have that \( \bar{y} \) is an unbiased estimator of \( \theta \) and it has smaller (\( \leq \)) variance than any other unbiased estimator of \( \theta \); accordingly, we call \( \bar{y} \) a uniformly minimum variance (UMV) unbiased estimator of \( \theta \).

The theorem also says that \( \bar{y} \) is the unique UMV unbiased estimator; for it is the only affine function of \( S(y|\theta_0) \) that is unbiased for \( \theta \).

The inequality (9) is sometimes used with Definition 9.3.1 to give an absolute measure of the efficiency of an estimator:
Sec. 9.4: The Information Inequality for Unbiased Estimators

\[ \text{Eff}(t|\theta) = \frac{1/\text{var}(t|\theta)}{J(\theta)}. \]

We will see cases where the UMV unbiased estimator has variance larger than the information lower bound (in Section 9.5). In such cases the above efficiency would not be 100 percent. This is to suggest caution in interpreting formula (10). Its usefulness is primarily in a large-sample context as discussed in Section D, which follows.

\textbf{C THE BEST LOCALLY UNBIASED ESTIMATOR}

We now use the information inequality in reverse to construct a locally unbiased(\(\theta_0\)) estimator that has minimum variance at the information lower bound. Theorem 2 says that a locally unbiased(\(\theta_0\)) estimator with variance at the lower bound must be an affine function of the score \(S(y|\theta_0)\). Thus we consider affine functions

\[ t(y) = a + cS(y|\theta_0) \]

of the score and see if \(a, c\) can be determined so that \(t(y)\) is locally unbiased(\(\theta_0\)).

Condition (6) for local unbiasedness gives

\[ E(a + cS(y|\theta_0)|\theta_0) = \theta_0 \]

\[ = a + c \cdot 0; \]

thus \(a = \theta_0\). Condition (7) gives

\[ E((a + cS(y|\theta_0))S(y|\theta_0)|\theta_0) = 1 \]

\[ = c \text{ var}(S(y|\theta_0)|\theta_0); \]

thus \(c = J^{-1}(\theta_0)\). It follows that the estimator

\[ (11) \quad t(y) = \theta_0 + J^{-1}(\theta_0)S(y|\theta_0) \]

is the unique locally unbiased(\(\theta_0\)) estimator that has minimum variance at the information lower bound.

\textbf{EXAMPLE 2}

\textbf{Poisson(\(\theta\))}: Consider a response \(y\) that is Poisson(\(\theta\)) with \(\theta\) in \(R^+\). We have immediately that

\[ S(y|\theta_0) = \frac{y}{\theta_0} - 1, \quad J(\theta_0) = \frac{1}{\theta_0}. \]

We now calculate the best unbiased estimator for \(\theta\) near \(\theta_0\):

\[ t(y) = \theta_0 + \theta_0 \left( \frac{y}{\theta_0} - 1 \right) = y. \]

This estimate is independent of \(\theta_0\). Thus it is unbiased generally and it has variance at the information lower bound; it is UMV unbiased.
D LARGE-SAMPLE ESTIMATION

We have examined some aspects of inference for large samples in Section 8.6. Now consider estimation for large samples in terms of the results in this section.

For the statistical model
\[ \{ f(\cdot | \theta) : \theta \in \Omega \}, \]
let \( S(y | \theta) \) be the score function and \( J(\theta) \) be the information. And let \((y_1, \ldots, y_n)\) be a sample from the model.

We know from Section 8.6 that a large sample determines almost conclusively a small interval in which the true parameter value must lie. If the parameter is believed to be in the neighborhood of \( \theta_0 \), then the best local estimate (11) for \( \theta \) is

\[
\theta_0 + \frac{\sum_{i=1}^{n} S(y_i | \theta_0)}{nj(\theta_0)}.
\tag{12}
\]

If this value, say \( \theta_1 \), is rather far from \( \theta_0 \), then an appropriate improved estimate based on \( \theta_1 \) would be

\[
\theta_1 + \frac{\sum_{i=1}^{n} S(y_i | \theta_1)}{nj(\theta_1)}.
\]

Typically, this procedure converges very rapidly (two or three iterations) to the maximum-likelihood estimate \( \hat{\theta} \). In particular, we can examine the best local estimate as calculated at \( \tilde{\theta} \):

\[
\tilde{\theta} + \frac{\sum_{i=1}^{n} S(y_i | \tilde{\theta})}{nj(\tilde{\theta})} = \hat{\theta},
\]

which is simplified by means of the likelihood equation (9.1.8). Thus the argument in terms of local estimates is tied closely to the maximum-likelihood method. The large-sample properties of the maximum-likelihood estimate were discussed in Section 8.6C. Note that the large-sample variance of the maximum-likelihood estimate is \( J^{-1}(\theta) = 1/nj(\theta) \), which is at the information lower bound.

It is of interest that the Newton–Raphson iterative procedure for the maximum-likelihood estimate has the form

\[
\theta_0 + \frac{\sum_{i=1}^{n} S(y_i | \theta_0)}{\sum_{i=1}^{n} S'(y_i | \theta_0)}.
\tag{13}
\]

The local estimate (12) has the mean value

\[
j(\theta) = E \left( -\sum_{i=1}^{n} S'(y_i | \theta) \right)
\]
in place of the denominator as given in (13). Very often the procedure using local estimates converges much faster than the Newton–Raphson procedure.
E WHEN DO THE BEST LOCAL ESTIMATORS GIVE THE BEST ESTIMATORS?

Consider an unbiased estimator $t$ that has variance at the information lower bound for all values of $\theta$. Then from formula (11) we have

$$ t(y) = \theta + J^{-1}(\theta)S(y|\theta) $$

for all values of $\theta$. We rearrange and then integrate with respect to $\theta$:

$$ S(y|\theta) = -\theta J(\theta) + J(\theta) t(y), $$

$$ \ln f(y|\theta) = \phi(\theta) + \psi(\theta) t(y) + a(y), $$

where $a(y)$ is the constant of integration and $\phi(\theta)$ and $\psi(\theta)$ are the obvious indefinite integrals. Thus we obtain

$$ f(y|\theta) = \gamma(\theta) \exp \{\psi(\theta) t(y)\} h(y), $$

which has exponential form with one $\psi$-function.

We thus see an example of a theoretical method whose effectiveness is limited to models having exponential form. Recall the remarks in Section 8.5D. Also, we have seen that the effectiveness of the method is limited to a model having exponential form with one $\psi$-function. But does it always work for such models?

Suppose that we have the score and information as calculated with respect to $\theta$ and yet we are interested in unbiased estimation of the real parameter $\beta(\theta)$. Then it is straightforward (Problem 6) to show that if $t$ is locally unbiased ($\beta(\theta_0)$) for $\beta(\theta)$, then

$$ \text{var} (t(y)|\theta_0) \geq (\beta'(\theta_0))^2 J^{-1}(\theta_0). $$

Now consider a statistical model having the exponential form (14). The parameter $\theta$ itself may not have an unbiased estimator with variance at the information lower bound. Problem 7 is to show that the only parameter having such an unbiased estimator is

$$ \frac{\phi'(\theta)}{\psi'(\theta)} $$

or a simple affine adjustment thereof. Thus not only is the effectiveness of the method limited to a model having exponential form with one $\psi$-function, but it is limited to essentially one parametrization of such a model.

F THE ROUTE TO UMV UNBIASED ESTIMATORS

In conclusion, we note that if an estimation problem has a UMV unbiased estimator with variance at the information lower bound, then the procedure of finding the best local estimate produces that UMV estimator.

Some problems, however, have UMV unbiased estimators with variance above the information lower bound. We will see that the methods in Section 9.5 can sometimes be successful in producing these estimators. We will also see that some estimation problems, in fact, have no UMV unbiased estimator.
G SUPPLEMENT: FOR VECTOR PARAMETERS

Consider briefly the case of a vector parameter \( \theta = (\theta_1, \ldots, \theta_p)' \) with values in \( \mathbb{R}^p \) or some connected open set of \( \mathbb{R}^p \). And suppose that we have the vector analogs of the regularity conditions in Section 8.6.

From Section 8.6G we have

\[
E(S(y|\theta)|\theta) = 0, \tag{17}
\]

\[
\text{VAR} \ (S(y|\theta)|\theta) = J(\theta) = -E \left( \frac{\partial S(y|\theta)}{\partial \theta'} | \theta \right). \tag{18}
\]

And by the procedure in Section A we have

\[
E(t_u(y)S_u(y|\theta)|\theta) = \frac{\partial}{\partial \theta_u} E(t_u(y)|\theta), \tag{19}
\]

or in matrix form,

\[
E(t(y)S'(y|\theta)|\theta) = \frac{d}{d\theta'} E(t(y)|\theta), \tag{20}
\]

where the prime on the \( S \) here denotes the transpose of the vector.

An estimator \( t \) is locally unbiased \((\theta_0)\) if

\[
E(t(y)|\theta_0) = \theta_0, \tag{21}
\]

\[
\left. \frac{\partial}{\partial \theta_u} E(t_v(y)|\theta) \right|_{\theta = \theta_0} = \delta_{uv}, \tag{22}
\]

where \( \delta_{uv} = 1 \) if \( u = v \) and 0 otherwise. And by formulas (17), (20), and (22) a locally unbiased \((\theta_0)\) estimator satisfies

\[
\text{COV} \ (t(y), S(y|\theta_0)|\theta_0) = I, \tag{23}
\]

the \( r \times r \) identity matrix.

The information inequality is then obtained from Problem 5.5.10. The matrix

\[
\text{VAR} \ (t(y)|\theta_0) - J^{-1}(\theta_0) \tag{24}
\]

is positive semidefinite. The best locally unbiased \((\theta_0)\) estimator is

\[
\theta_0 + J^{-1}(\theta_0)S(y|\theta_0). \tag{25}
\]

An unbiased estimator with variance matrix at the information lower bound has the form

\[
t(y) = \theta + J^{-1}(\theta)S(y|\theta),
\]

and this occurs only if the model has exponential form with \( r \psi \)-functions.

H EXERCISES

1. Let \( x_1, \ldots, x_n \) be a sample from the Bernoulli distribution with \( p \) in \( \Omega = (0, 1) \). Derive the minimum variance locally unbiased \((p_0)\) estimate of \( p \). Is a UMV unbiased estimate available?

2. Let \( y_1, \ldots, y_n \) be a sample from the exponential distribution \( f(y|\theta) = \theta^{-1} \exp \{-y/\theta\} \) on \( R^+ \) with \( \theta \) in \( \Omega = R^+ \). Derive the minimum variance locally unbiased \((\theta_0)\) estimate of \( \theta \). Is a UMV unbiased estimate available?
3 Let \((y_1, \ldots, y_n)\) be a sample from the normal\((\mu, \sigma^2)\) with \(\sigma^2\) in \(R^+\). Derive the minimum variance locally unbiased(\(\sigma^2\)) estimate of \(\sigma^2\). Is a UMV unbiased estimate available?

4 Let \((y_1, \ldots, y_n)\) be a sample from \(f(y|\theta) = \theta y^{k-1}\) on \((0, 1)\) and zero otherwise with \(\theta\) in \(\Omega = R^+\). Derive the minimum variance locally unbiased(\(\theta_0\)) estimate of \(\theta\). Is a UMV unbiased estimate for \(\theta\) with variance at the information bound available?

5 Consider independent models: \(f_1(y|\theta)\) with \(S_1(y|\theta)\) and \(J_1(\theta)\); and \(f_2(y|\theta)\) with \(S_2(y|\theta)\) and \(J_2(\theta)\). For local unbiased(\(\theta_0\)) estimation we have \(\theta_0 + J_1^{-1}(\theta_0)S_1(y|\theta_0)\) from the first model and \(\theta_0 + J_2^{-1}(\theta_0)S_2(y|\theta_0)\) from the second model. Combine these local estimates by the methods in Section 9.3C and show that the combined estimate is the best locally unbiased(\(\theta_0\)) estimate for the combined model.

I PROBLEMS

6 Let \(\beta(\theta)\) be a continuously differentiable monotone increasing function mapping \(R\) into \(R\) with \(\beta'(\theta) = d\beta(\theta)/d\theta\). Let \(S(y|\theta)\) and \(J(\theta)\) be the score and information relative to \(\theta\). Show that the score, information, and best locally unbiased(\(\beta(\theta)\)) estimate relative to \(\beta(\theta)\) are

\[
S(y|\theta)/\beta'(\theta), \quad J(\theta)/[\beta'(\theta)]^2, \quad \beta(\theta_0) + \beta'(\theta_0)S(y|\theta_0)/J(\theta_0).
\]

7 For the exponential model (14) show that the only parameter with an unbiased estimator at the information lower bound is \(\phi(\theta)/\psi'(\theta)\) or an affine function thereof; note that \(\phi(\theta) = \ln y(\theta)\). The formulas in Problem 6 are not needed.

8 For Exercise 4 determine the parameter that has an unbiased estimator with variance at the information lower bound.

9 Let \(y\) have a distribution with density

\[f(y|\theta) = \theta(\theta + 1)y(1 - y)^{k-1}\]

on \(0 < y < 1\) and zero otherwise. Determine the parameter that has an unbiased estimator at the information lower bound.

10 Consider a distribution on \(R^{m+n}\) for

\[
\left( \begin{array}{c}
  x \\
  y
\end{array} \right) = (x_1, \ldots, x_m, y_1, \ldots, y_n)
\]

with mean \(0\) and variance matrix

\[
\Sigma = \begin{pmatrix}
  \Sigma_{11} & \Sigma_{12} \\
  \Sigma_{21} & \Sigma_{22}
\end{pmatrix}
\]

(as correspondingly partitioned). Consider the variance matrix of \(y = \Sigma_{21}\Sigma_{11}^{-1}x\) and show that \(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\) is positive semidefinite and is zero if and only if \(y = Cx\), where \(C\) is \(m \times n\). Compare with Problem 5.5.10.

J PROBLEMS FOR VECTOR PARAMETERS

11 Consider independent models: \(f_1(y_1|\theta)\) with \(S_1(y_1|\theta)\) and \(J_1(\theta)\); and \(f_2(y_2|\theta)\) with \(S_2(y_2|\theta)\) and information matrix \(J_2(\theta)\). Combine the best locally unbiased(\(\theta_0\)) estimator by the methods in Section 9.3C and show that the combined estimator is the best locally unbiased(\(\theta_0\)) estimator from the combined model.

12 Let \((y_1, \ldots, y_n)\) be a sample from a normal\((\mu, \sigma^2)\) with \(\theta = (\mu, \sigma^2)\) in \(R \times R^+\). Calculate \(S(y|\theta)\) and \(J(\theta)\). Compare with Problem 8.6.12.

13 (continuation) Derive the minimum variance-matrix locally unbiased(\(\theta_0\)) estimator of \(\theta\). Is an unbiased estimator with variance matrix at the information lower bound available?
14 (continuation) Calculate the variance matrix of \((\bar{y}, s^2_0)\) and compare with the information lower bound.

15 Prove formula (20) for the derivative of the mean of a vector estimate.

16 Prove the information inequality (24) for vector parameters. Show that (24) is the zero matrix if and only if \(t(y)\) is an affine function of \(S(y|\theta_0)\). Use Problem 10.

17 (continuation) Show that the best locally unbiased \((\theta_0)\) estimator of \(\theta\) is \(t = \theta_0 + J^{-1}(\theta_0)S(y|\theta_0)\).

18 (continuation) Show that the information lower bound is attained for the unbiased estimator of the vector parameter \((\theta_1, \ldots, \theta_r)\) only if the model is exponential with \(r\) \(\psi\)-functions.

19 (continuation) Consider the reduced exponential model \(f(y|\theta) = \exp\{\phi(\theta) + \psi(\theta)s(y)h(y)\}\) with \(r\) \(\psi\)-functions and with \(\theta\) in \(\mathbb{R}^r\). Show that the only vector parameter that has an unbiased estimator with variance matrix at the information lower bound is \((\partial \psi(\theta)/\partial \theta)^{-1}\partial \phi(\theta)/\partial \theta\) or an affine equivalent thereof.

20 For the normal model in Problem 12 show that the only parameter with an unbiased estimator at the information lower bound is \((\mu, \mu^2 + \sigma^2)\) or an affine equivalent; use Problem 19.

9.5 LIKELIHOOD AND COMPLETENESS FOR UNBIASED ESTIMATORS

When the likelihood function was introduced in Section 8.1 we noted that a statistical model says the same thing about each sample point with a particular likelihood function. This led us in Section 8.4 to define the likelihood statistic, a simple statistic that indexes the possible likelihood functions.

In this section we will see that the choice of an unbiased estimator should be restricted to those that are based on the likelihood statistic; of course this agrees with the general views in Chapter 8 but is based here on minimum variance. And we will see that if the likelihood statistic has a property of completeness, then such an estimate based on the likelihood statistic is unique.

A ESTIMATORS BASED ON THE LIKELIHOOD STATISTIC

For notation let \(y\) be the full response under investigation with sample space \(S\), and let \(\theta\) be the parameter with values in a parameter space \(\Omega\). We assume that the statistical model,

\[\{f(\cdot|\theta) : \theta \in \Omega\}\],

is given in terms of density functions as discussed in Section 7.1. This makes available the likelihood statistic as defined in Section 8.4B; let \(s\) designate the likelihood statistic.

From the probability Section 5.5 we know that variance is reduced by taking a conditional mean. This is how we will improve an initial estimator to obtain an estimator based on the likelihood statistic. For a real-valued parameter \(\beta(\theta)\) we have the following Rao–Blackwell theorem:
THEOREM 1

If $t(y)$ is an unbiased estimator of $\beta(\theta)$ and $s(y)$ is the likelihood statistic, then
(a) The regression of $t(y)$ on $s(y)$, which is the function

$$E(t(y) : s(y) = s) = r(s)$$

is an unbiased estimator of $\beta(\theta)$ and is based on the likelihood statistic $s(y)$.
(b) The variance is reduced in going from the estimator $t$ to the estimator $r$,

$$\text{var}(t(y)|\theta) \geq \text{var}(r(s(y))|\theta)$$

with equality if and only if the estimator $t(y) = r(s(y))$ is already a function of
the likelihood statistic $s(y)$ (with probability 1).

The theorem shows that an unbiased estimator not based on the likelihood statistic
can be improved by taking its mean value given the likelihood statistic. The theorem
remains true if $s(y)$ is any sufficient statistic; the likelihood statistic provides the largest
reduction and would thus be the preferred sufficient statistic. The proof is recorded in
Section B.

Consider some simple examples. The results are rather elementary and indeed
familiar, but they do illustrate the ideas involved.

EXAMPLE 1

Sample from the Bernoulli: Let $(x_1, \ldots, x_n)$ be a sample from the Bernoulli($\rho$)
distribution with $\rho$ in $[0, 1]$. From Example 8.4.2 we know that $y = \Sigma^n x_i$ is the
likelihood statistic.

We examine unbiased estimation of the parameter $\rho$. A very simple and
elementary estimator is given by the first coordinate $x_1$:

$$E(x_1|\rho) = 1 \cdot \rho + 0 \cdot q = \rho.$$

Of course, this estimator is rather ridiculous in that it just takes two values, 0 and
1, but technically it is unbiased. We now use Theorem 1 to derive an improved
estimator $r(y)$:

$$r(y) = E(x_1|y) = 1 \cdot \frac{y}{n} + 0 \cdot \frac{n - y}{n} = \frac{y}{n};$$

for this note from symmetry that the conditional distribution of $(x_1, \ldots, x_n)$ given the
sum $y = \Sigma x_i$ has equal probability $1/{n \choose y}$ for each sequence consisting of
$y$ 1's and $n - y$ 0's, and thus that $x_1$ is equal to 0 and 1 with probabilities $(n - y)/n$ and
$y/n$, respectively. Thus we see that the improved estimator is just the proportion
$\hat{\rho} = y/n$, which we have used frequently before. Also we see that

$$\text{var}(x_1|\rho) = \rho q, \quad \text{var}(y/n|\rho) = \rho q/n.$$
EXAMPLE 2

Sample from the location normal: Let \( \{y_1, \ldots, y_n\} \) be a sample from the normal\((\theta, \sigma^2)\) distribution with \( \theta \) in \( R \). From Section 8.4A we know that \( \bar{y} \) or \( u_1 = \sqrt{n} \bar{y} \) is the likelihood statistic.

We examine unbiased estimation of the parameter \( \theta \). A very simple and immediate estimator is given by the first coordinate \( y_1 \):

\[
E(y_1 | \theta) = \theta.
\]

Again we are starting with a rather ridiculous initial estimator: the first coordinate—ignoring all other coordinates. We now use Theorem 1 to derive an improved estimator. For this we recall the distribution results obtained from the transformation in formula (8.2.10). This transformation can be solved, giving

\[ y_1 = \frac{1}{\sqrt{n}} u_1 + \frac{1}{\sqrt{2}} u_2 + \cdots + \frac{1}{\sqrt{n(n - 1)}} u_n, \]

where the \( u \)'s are statistically independent with common variance \( \sigma^2 \) and with means all equal to 0 except for \( u_1 \). Thus

\[
E(y_1 : u_1) = \frac{1}{\sqrt{n}} \cdot u_1 + \frac{1}{\sqrt{2}} \cdot 0 + \cdots + \frac{1}{\sqrt{n(n - 1)}} \cdot 0
\]

\[ = \bar{y}. \]

Note that the improved estimator is the familiar sample average. Also note that

\[
\text{var } (y_1 | \theta) = \sigma^2 \quad \text{var } (\bar{y} | \theta) = \frac{\sigma^2}{n}.
\]

---

B THE RAO–BLACKWELL THEOREM: ITS PROOF AND A VECTOR EXTENSION

Consider the proof of Theorem 1. We know that the conditional distribution given the likelihood statistic is independent of \( \theta \); see Theorem 8.4.3. Thus the conditional mean

\[
E(t(y) : s(y) = s) = r(s)
\]

does not depend on \( \theta \), and correspondingly \( r(s(y)) \) is a function on the sample space that does not involve \( \theta \). We now use formula (5.5.3) for taking a mean in two stages:

\[
(3) \quad \beta(\theta) = E(t(y) | \theta) = E(E(t(y) : s(y) = s) | \theta) = E(r(s) | \theta).
\]

Thus \( r(s(y)) \) is an unbiased estimate of \( \beta(\theta) \), proving part (a) of the theorem.

We now consider the variance reduction formula (2). For this we use formula (5.5.4), expressing variance in terms of variance of regression and variance about regression:

\[
(4) \quad \text{var } (t | \theta) = E(\text{var } (t : s | \theta) + \text{var } (r | \theta).
\]

Thus the inequality (2) holds. Also we have equality if and only if \( E(\text{var } (t : s | \theta) = 0, \) or, equivalently, \( \text{var } (t : s) = 0 \) (with probability 1), or, equivalently, \( t \) is a constant given \( s \) (with probability 1). This proves part (b) of the theorem.
For the case of a vector parameter $\beta(\theta)$, we record the generalized Rao–Blackwell theorem for improving estimators:

**THEOREM 2**

If $t(y)$ is an unbiased estimator of $\beta(\theta)$ and $s(y)$ is the likelihood statistic, then

(a) The regression of $t(y)$ on $s(y)$,

$$E(t(y) : s(y) = s) = r(s),$$

is an unbiased estimator of $\beta(\theta)$ based on the likelihood statistic $s(y)$.

(b) The variance matrix is reduced in going from the estimator $t$ to the estimator $r$; thus

$$\text{VAR}(t|\theta) \rightarrow \text{VAR}(r|\theta)$$

is positive semidefinite and is the zero matrix if and only if the estimator $t(y) = r(s(y))$ is already a function of the likelihood statistic $s(y)$ (with probability 1).

The proof follows by using (5.5.9) and (5.5.10) in place of (5.5.3) and (5.5.4). Note that the theorem says to improve each coordinate by taking a mean as described in Theorem 1. This reduces the variance matrix; it also reduces the variance of any coordinate that is not already a function of the likelihood statistic; and it reduces the variance of any linear combination of coordinates that is not already a function of the likelihood statistic.

### C  COMPLETENESS, AND UNIQUENESS
### OF UNBIASED ESTIMATORS

Theorem 1 says that in looking for an unbiased estimator of a parameter we need examine only those based on the likelihood statistic. How many such estimators are there? We will see that for some statistical models there is only one; it must then be the UMV-unbiased estimator.

First a convenient definition:

**DEFINITION 3**

A parameter $\beta(\theta)$ is estimable if there exists an unbiased estimator for it.

Parameters can be found that do not have unbiased estimates. For example, with a sample of 2 from the Bernoulli($p$) only the polynomials $a + bp + cp^2$ up to order 2 have unbiased estimates. And, for example, with a sample from an arbitrary continuous distribution on $R$, the median does not have an unbiased estimate.

We use the notation in Section A and we let $g(s|\theta)$ be the density for the likelihood statistic; we record the proof for the integration case. Suppose that $\beta(\theta)$ is estimable. Then by Theorem 1 there is an unbiased estimator based on the likelihood statistic.
Suppose there are two, \( r_1(s|\theta) \) and \( r_2(s|\theta) \). Then
\[
\beta(\theta) = \int r_1(s)g(s|\theta)\,ds \\
\beta(\theta) = \int r_2(s)g(s|\theta)\,ds,
\]
and hence
\[
(7) \quad 0 = \int r(s)g(s|\theta)\,ds,
\]
where \( r(s) = r_1(s) - r_2(s) \) is the difference between the two unbiased estimators. We call \( r \) an unbiased estimator of zero.

**DEFINITION 4**

A statistical model \( \{g(\cdot|\theta) : \theta \in \Omega \} \) is complete if
\[
(8) \quad \int r(s)g(s|\theta)\,ds = 0 \quad \text{for} \ \theta \in \Omega
\]
implies that \( r(s) = 0 \) (with probability 1), that is, if there are no nontrivial unbiased estimators of zero.

If the statistical model for the likelihood statistic is complete, then equation (7) implies that \( r(s) = 0 \) and any two unbiased estimators based on the likelihood statistic must be equal. This proves the following Lehmann–Scheffé theorem.

**THEOREM 5**

If the statistical model for the likelihood statistic is complete, then an estimable parameter has exactly one unbiased estimator based on the likelihood statistic.

Rather than say that a statistical model for the likelihood statistic is complete, we will, for convenience, say that the *likelihood statistic is complete*, provided that the context makes clear what statistical model is being used.

By Theorems 1 and 2 we know that variance (or \textsc{VARIance} matrix) is reduced by going to estimators based on the likelihood statistic. If the likelihood statistic is complete, then we know there is only one improved estimator and, accordingly, it is UMV unbiased. This gives the following:

**COROLLARY 6**

If the likelihood statistic is complete, then an unbiased estimator based on the likelihood statistic is the UMV unbiased estimator.

**D EXAMPLES OF COMPLETENESS**

Consider two examples of completeness, one discrete and one absolutely continuous.
EXAMPLE 3

Poisson($\lambda$): Let $s$ have a Poisson($\lambda$) distribution with $\lambda$ in $R^+$. We show that $s$ is complete. For this let $r(s)$ be an unbiased estimator of zero:

$$\sum_0^\infty r(s)\lambda^s e^{-\lambda}/s! = 0 \quad \text{for } \lambda \in R^+,$$

$$\sum_0^\infty \frac{r(s)}{s!} \lambda^s = 0 \quad \lambda \in R^+.$$

This is a power series that is convergent to zero for all values of $\lambda$ in $R^+$. The coefficients of a power-series expansion are unique; hence $r(s)/s! = 0$, or, equivalently, $r(s) = 0$. This establishes completeness.

EXAMPLE 4

The location normal: Let $s$ have a normal($\theta$, $\tau_0$) distribution with $\theta$ in $R$. We show that $s$ is complete. For this, let $r(s)$ be an unbiased estimator of zero:

$$\int_{-\infty}^{\infty} r(s)\frac{1}{\sqrt{2\pi} \tau_0} \exp \left\{ -\frac{1}{2\tau_0^2} (s - \theta)^2 \right\} ds = 0,$$

$$\int_{-\infty}^{\infty} \left[ r(s) \exp \left\{ -\frac{s^2}{2\tau_0^2} \right\} \right] \exp \left\{ s \frac{\theta}{\tau_0^2} \right\} ds = 0,$$

$$\int_{-\infty}^{\infty} r^+(s) \exp \left\{ -\frac{s^2}{2\tau_0^2} \right\} \exp \{ st \} ds = \int_{-\infty}^{\infty} r^-(s) \exp \left\{ -\frac{s^2}{2\tau_0^2} \right\} \exp \{ st \} ds$$

for all $\theta$ in $R$ or all $t = \theta/\tau_0^2$ in $R$, where $r^+$ and $r^-$ are the positive and negative components of $r$: $r^+(s) = \max (0, r(s))$, $r^-(s) = \max (0, -r(s))$.

Let $D$ be the value of the two expressions in (9) with $t = 0$. If $D \neq 0$, then we can write

$$\int_{-\infty}^{\infty} p_1(s) \exp \{ st \} ds = \int_{-\infty}^{\infty} p_2(s) \exp \{ st \} ds$$

for all $t$ in $R$ where

$$p_1(s) = \frac{r^+(s)}{D} \exp \left\{ -\frac{s}{2\tau_0^2} \right\}, \quad p_2(s) = \frac{r^-(s)}{D} \exp \left\{ -\frac{s}{2\tau_0^2} \right\}$$

are probability densities. Equation (9) says that $p_1$ and $p_2$ have the same moment-generating function and thus are identical. But from their definition $p_1$ and $p_2$ are nonzero on different sets, thus providing a contradiction. It follows that $D = 0$ and thus that $r^+(s) = 0$ and $r^-(s) = 0$, and thus $r(s) = 0$ with probability 1. This establishes completeness.

Statistical models can be not complete; see, for example, Exercise 5. Nevertheless, a UMV unbiased estimate may still exist.
E SOME EXAMPLES OF UMV UNBIASED ESTIMATORS

Consider a statistical model with a complete likelihood statistic. To construct the UMV unbiased estimator for a parameter $\beta(\theta)$ we have essentially two approaches. We can look for some unbiased estimator $t(y)$ and then take the regression on $s(y)$, obtaining the UMV unbiased estimator $r(s(y))$. This is based on Theorems 1 and 2 and as illustrated by Examples 1 and 2 in Section A. Alternatively, we can try to construct directly an unbiased estimator based on the likelihood statistic; if we succeed, then it is the UMV unbiased estimator. The first approach is amply illustrated in Section A; we discuss the second approach now.

Let $g(s|\theta)$ be the density function for the likelihood statistic. An estimable parameter $\beta(\theta)$ has an estimator $r(s)$ such that

\begin{equation}
(11) \quad \beta(\theta) = \int r(s)g(s|\theta)\,ds.
\end{equation}

From Corollary 8.4.4 we know that the likelihood function from $s = s(y)$ with the marginal model $g(s|\theta)$ is the same as the likelihood function from the antecedent $y$ using the full model $f(y|\theta)$. Accordingly, we can write $g(s|\theta) = k(s)L(s|\theta)$, where $k(s)$ selects the appropriate representative from the likelihood function; hence

\begin{equation}
(12) \quad \beta(\theta) = \int r(s)k(s)L(s|\theta)\,ds.
\end{equation}

Thus we can think of an estimable parameter $\beta(\cdot)$ as a weighted combination (11) of functions $g(s|\cdot)$ or as a weighted combination (12) of the possible likelihood functions $L(s|\cdot)$. More positively, we try to expand the parameter (11) in terms of densities $g(s|\cdot)$ and thus obtain the coefficients $r(s)$ that give the UMV unbiased estimator.

EXAMPLE 5

Sample from the Poisson($\theta$): Let $(y_1, \ldots, y_n)$ be a sample from a Poisson($\theta$) distribution with $\theta$ in $R^+$. The likelihood statistic $s = \sum_i^n y_i$ is complete: Example 3 with $\lambda = n\theta$. The function $\bar{y}$ is an unbiased estimator of $\theta$ based on $s$; thus it is UMV unbiased.

Now suppose that we are interested in the parameter $e^{-\theta}$; this is the probability that a response $y$ is zero. We could start with an initial estimator based on the indicator function $h$ from the first coordinate,

\[ h(y_i) = 1, 0 \quad \text{according as } y_1 = 0, > 0; \]

and we could improve it using Theorem 1. Alternatively, we can try to expand $e^{-\theta}$ in terms of the likelihood functions:

\[ e^{-\theta} = \sum_{s=0}^{\infty} r(s) \frac{(n\theta)^s e^{-n\theta}}{s!}, \]

\[ e^{n-n\theta} = \sum_{s=0}^{\infty} r(s) \frac{\theta^s}{s!}. \]

Then identifying the coefficient of $\theta^s/s!$ on each side, we obtain
\[(n - 1)^2 = r(s) = r(s)n^d,\]

\[r(s) = \left(1 - \frac{1}{n}\right)^d.\]

Thus \((1 - 1/n)^4\) is the UMV unbiased estimator of \(e^{-\theta}\).

**EXAMPLE 6**

**Sample from the location normal:** Let \((y_1, \ldots, y_n)\) be a sample from a normal(\(\theta, \sigma_0^2\)) distribution with \(\theta\) in \(R\). The likelihood statistic \(s = \bar{y}\) is complete: Example 4 with \(\tau_{0}^{2} = \sigma_{0}^{2}/n\).

The function \(\bar{y}\) is an unbiased estimator of \(\theta\) based on \(s\); thus it is UMV unbiased.

In a similar way we see that \(\bar{y}^2\) is the UMV unbiased estimator for \(E(\bar{y}^2) = \theta^2 + \sigma_0^2/n\), and then that \(\bar{y}^2 + (n - 1)\sigma_0^2/n\) is the UMV unbiased estimator for \(\mu_2 = \theta^2 + \sigma_0^2\). In fact, any function of \(\bar{y}\) with finite variance is the UMV unbiased estimator of its mean.

For each of these examples there is really only one parameter—however one chooses to express it. Thus, if we take \(s\) as the natural estimate of \(\theta\), we would use \(\beta(s)\) as the appropriate estimate of \(\beta(\theta)\). The property of unbiasedness is typically spoiled by a nonlinear function \(\beta\). The fact that we would need a new estimator on the basis of UMV unbiasedness is really an indication of the arbitrariness of the unbiased property.

**F WHEN DO THE METHODS WORK?**

The procedures in this section give a UMV unbiased estimator for a model that has a complete likelihood statistic. The exercises and problems will illustrate the kinds of models that have this property: the exponential model with the number of \(\psi\)-functions equal to the dimension of the parameter; the variable carrier model with the number of boundary types equal to the dimension of the parameter; and fairly straightforward mixtures of these. For the models with a fixed region of positive density we are restricted here, as with the methods in Section 9.4, to models with simple exponential form. We have noted earlier in Section 8.5D that the common models having exponential form are in the continuous case essentially just the location normal, the scale normal, and the location-scale normal. And these indeed are the examples we are using here.

**G EXERCISES**

1. **The scale exponential:** Let \((y_1, \ldots, y_n)\) be a sample from the exponential distribution \(f(y|\theta) = \theta^{-1} \exp\{-y/\theta\}\) on \(R^+\) with \(\theta\) in \(R^+\). Show that the likelihood statistic \(\Sigma y_i\) is complete; use Problem 17. Obtain the UMV unbiased estimator for \(\theta\).

2. **The scale normal:** Let \((y_1, \ldots, y_n)\) be a sample from the normal(\(\mu_0, \sigma^2\)) with \(\sigma^2\) in \(\Omega = R^+\). Show that the likelihood statistic \(\Sigma (y_i - \mu_0)^2\) is complete; use Problem 17. Obtain the UMV unbiased estimator for \(\sigma^2\).

3. **The uniform\((0, \theta)\):** Let \((y_1, \ldots, y_n)\) be a sample from the uniform\((0, \theta)\) with \(\theta\) in \(R^+\). Show that the likelihood statistic \(\max y_i\) is complete. Obtain the UMV unbiased estimator for \(\theta\).
Show that \( 2\bar{Y} \) is an unbiased estimate of \( \theta \). How do the variances of these two unbiased estimators compare for \( n \) large?

4 The Bernoulli(\( p \)): Let \((x_1, \ldots, x_n)\) be a sample from the Bernoulli(\( p \)) distribution with \( p \in [0, 1] \). Show that the likelihood statistic \( y = \Sigma x_i \) is complete. Obtain the UMV unbiased estimator for \( p \).

5 Let \((y_1, y_2)\) be a sample from the uniform(\( \theta, \theta + 1 \)) with \( \theta \) in \( R \). Show that the likelihood statistic \((y_{11}, y_{12})\) is not complete; that is, find a nontrivial function of \((y_{11}, y_{12})\) that has mean value 0 for each \( \theta \) value. In fact, there is no UMV unbiased estimate; Lehmann and Scheffé (1950).

**H PROBLEMS**

6 For a sample \((y_1, \ldots, y_n)\) from the Poisson(\( \theta \)) in Example 5, find the UMV unbiased estimator of \( e^{-\theta} \) by determining the regression of \( h(x_i) \) on \( s = \Sigma y_i \).

7 Let \((x_1, \ldots, x_n)\) be a sample from the Bernoulli(\( p \)) distribution with \( p \) in \([0, 1]\). The likelihood statistic \( y = \Sigma x_i \) is complete; Exercise 4. Obtain the UMV unbiased estimator for \( p^2 \); for \( pq \).

8 Let \((y_1, \ldots, y_n)\) be a sample from the binomial(\( n, p \)) distribution with \( p \) in \([0, 1]\). Determine the UMV unbiased estimate of \( q^a \); the probability of a zero count; compare with Problem 6.

9 Let \( y \) have the hypergeometric(\( N, n, D \)) distribution with \( D \) in \([0, 1, \ldots, N]\). Show that \( y \) is complete. Obtain the UMV unbiased estimator for \( D/N \).

10 Let \((y_1, \ldots, y_n)\) be a sample from the model \( f(y|\theta) = k(\theta)^s(\phi/y)h(y) \), where \( \phi \) is the indicator function for the interval \((0, 1]\). Show that \( \max y_i \) is complete.

**I PROBLEMS FOR VECTOR PARAMETERS**

11 **The location-scale normal**: Let \((y_1, \ldots, y_n)\) be a sample from the normal(\( \mu, \sigma \)) with \((\mu, \sigma^2)\) in \( R \times R^+ \). Show that the likelihood statistic \((\Sigma y_i, \Sigma y_i^2)\) is complete; use Problem 17. Obtain the UMV-matrix unbiased estimator for \((\mu, \sigma^2)\).

12 Let \((y_1, \ldots, y_n)\) be a sample from the normal(\( \mu, \sigma \)) with \((\mu, \sigma^2)\) in \( R \times R^+ \). Show that \( s^2(n - 1)^{-1}(n - 1)/2^{1/2}(n - 2)^{-1}(n - 1)/2^{1/2}(n - 1)/2^{1/2} \approx s^2(1 + 1/4(n - 1)) \) is the UMV unbiased estimate of \( \sigma \); use Problem 11.

13 Let \((y_1, \ldots, y_n)\) be a sample from the uniform(\( \theta_1, \theta_2 \)) with \( \theta_1 < \theta_2 \) and \( \theta_1 \) in \( R \). Show that the likelihood statistic \((\min y_i, \max y_i)\) is complete. Obtain the UMV-matrix unbiased estimator for \((\theta_1, \theta_2)\).

14 **The location-scale exponential**: Let \((y_1, \ldots, y_n)\) be a sample from \( f(y|\theta, \tau) = \tau^{-1}c(y - \theta) \exp \{ -\frac{1}{\tau}(y - \theta)/\tau \} \), where \( c \) is the indicator function for \((0, \infty) \). Show that the likelihood statistic is \((\min y_i, \Sigma y_i)\). Show that the likelihood statistic is complete; compare with Exercises 1 and 3.

15 Let \((y_1, \ldots, y_r)\) be multinomial \((n; p_1, \ldots, p_r)\) with \( p_1 \geq 0, \Sigma p_i = 1 \). Show that \((y_1, \ldots, y_r)\) is complete. Obtain the UMV matrix unbiased estimator for \((p_1, \ldots, p_r)\).

16 Let \((y_1, \ldots, y_r)\) be general hypergeometric \((N, n, D_1/N, \ldots, D_r/N)\) with \( D_i = 0, 1, \ldots, \) and \( \Sigma D_i = N \). Show that \((y_1, \ldots, y_r)\) is complete. Determine the UMV matrix unbiased estimator for \((D_1/N, \ldots, D_r/N)\).

17 **Exponential form**: Let \( y \) have density function

\[
 f(y|\theta) = \gamma(\theta) \exp \left\{ \sum_{i=1}^{r} s_i(y)\theta_{si} \right\} h(y)
\]

with respect to volume or a counting measure. Then by formula (B.4.14) \( s = (s_1(y), \ldots, s_r(y)) \) has density

\[
 g(s|\theta) = \gamma(\theta) \exp \left\{ \sum_{i=1}^{r} s_i\theta_{si} \right\} H(s).
\]
Then for the model with \( \theta \) in \( \mathbb{R}' \) (or in a subset of \( \mathbb{R}' \) containing an open rectangle), show that \( \mathbf{s} \) is complete. **Method:** If the rectangle does not contain the origin, then set \( \theta = \phi + a \) and redistribute terms so that the new parameter has a rectangle containing the origin. Follow Example 4 and use the uniqueness theorem for multivariate moment-generating functions in Supplement Section 5.8E.

18 Consider the normal linear model \( \mathbf{y} = X\beta + \mathbf{v} \), where \( \mathbf{v} \) is a sample from the normal(0, \( \sigma^2 \)) with \( \beta \) in \( \mathbb{R}' \). Show that the likelihood statistic \( X'\mathbf{y} \) is complete; use Problem 17. For future reference note that \( b(\mathbf{y}) = (X'X)^{-1}X'\mathbf{y} \) is an equivalent function.

19 Consider the normal linear model \( \mathbf{y} = X\beta + a\mathbf{z} \), where \( \mathbf{z} \) is a sample from the normal(0, 1) with \( \theta = (\beta, \sigma^2) \) in \( \mathbb{R}' \times \mathbb{R}^+ \). Show that the likelihood statistic \( (X'\mathbf{y}, y'y) \) is complete; use Problem 17 and recall Problem 8.5.7. For future reference note that \( (b(\mathbf{y}), s^2(\mathbf{y})) \) is an equivalent function, where \( s^2(\mathbf{y}) \) is the minimized sum of squares in Section 9.2B.

20 (continuation) Show that the least-squares estimates \( b_1, \ldots, b_k \) are UMV unbiased; recall Theorem 9.1.4.

21 **Measuring wavelength:** Measurements are made to determine a period of oscillation \( \beta \). Let \( \gamma_0 \) be measured time at which the system is in a certain phase, and let \( \gamma_1, \ldots, \gamma_n \) be the measured times of the next \( n \) recurrences to that phase. If the measurement error is normally distributed without bias and with common variance, then the linear model \( \mathbf{y} = \alpha + \beta_1 \mathbf{x} + a \mathbf{z} \) is appropriate where \( \mathbf{z} \) is a sample from the standard normal and \( \mathbf{x} = (0, 1, 2, \ldots, n)' \). Find the UMV unbiased estimator \( b \) of \( \beta \); use Problem 20 together with Problem 9.2.11.

22 (continuation) Consider the following estimators:

\[
\begin{align*}
   b_2 &= \frac{\bar{y}_n - \gamma_0}{n} \\
   b_1 &= \begin{cases} \\
   \frac{\sum_{i=1}^{n} y_i - \sum_{i=0}^{n/2-1} y_i}{(n^2 + 2n)/4} & (n \text{ even}), \\
   \frac{\sum_{i=1}^{n} y_i - \sum_{i=0}^{(n-1)/2} y_i}{(n + 1)/4} & (n \text{ odd}).
   \end{cases}
\end{align*}
\]

(a) Show that \( b_1 \) and \( b_2 \) are unbiased estimators of \( \beta \). Calculate the variances of \( b, b_1, \) and \( b_2 \).

(b) Suppose that \( b_2 \) is available for \( n = 300 \). Find values of \( n \) for which \( b \) and \( b_1 \) would have the same precision.

23 Let \( (\gamma_1, \ldots, \gamma_n) \) be a sample from some density function on \( \mathbb{R} \) (\( \Omega \) would be an index set for the class of density functions or the class of piecewise continuous density functions). The statistic \( (\gamma_1, \ldots, \gamma_n) \) is sufficient; see Problem 8.4.8. An equivalent form of the statistic is \( (\Sigma y_i, \Sigma y_i^2, \ldots, \Sigma y_i^n) \); see Problem 3.1.9. The density functions as described above include those of the model \( f(y) = \phi \theta \exp \left\{ -\gamma^2 \theta_1 \gamma_1 + \cdots + \theta_n \gamma^n \right\} \) having exponential form. Deduce that the statistic is complete; use Problem 17.

24 Consider a sample of \( n > 1 \) from the multivariate normal \( (\mu; \Sigma) \) with \( \mu \) in \( \mathbb{R}^k \) and \( \Sigma \) in the space of positive definite symmetric matrices (an open set of \( \mathbb{R}^{k \times (k+1)/2} \)); see Problem 8.5.8.

Let

\[
Y = \begin{pmatrix}
   y_{11} & \cdots & y_{1k} \\
   \vdots & \ddots & \vdots \\
   y_{k1} & \cdots & y_{kk}
\end{pmatrix},
\]

where the \( n \) rows record the \( n \) elements in the sample. Show that

\[
\begin{pmatrix}
   \bar{y}_{11} & \cdots & \bar{y}_{1k} \\
   \vdots & \ddots & \vdots \\
   \bar{y}_{k1} & \cdots & \bar{y}_{kk}
\end{pmatrix} = n^{-1}1'Y,
\]

\[
\begin{pmatrix}
   s_{11} & \cdots & s_{1k} \\
   \vdots & \ddots & \vdots \\
   s_{k1} & \cdots & s_{kk}
\end{pmatrix} = (Y - 1n^{-1}1'Y)(Y - 1n^{-1}1'Y)/(n - 1),
\]

where \( 1 \) is the 1-vector with \( n \) coordinates and the \( s_{ij} \) are sample covariances; Section 5.4D.
Show that the sample averages and covariances are UMV unbiased estimates of the means and covariances of the distribution.

Theorem 5 shows that a UMV unbiased estimate is unique under the assumption of a complete likelihood statistic. Show that the uniqueness holds without this assumption. For this, let \( \sigma^2(\theta) \) be the variance of two UMV unbiased estimates \( t_1 \) and \( t_2 \) of a parameter \( \beta(\theta) \), let \( \sigma_{12}(\theta) \) be the covariance, and consider the variance of \( at_1 + (1 - a)t_2 \).

SUPPLEMENTARY MATERIAL

Supplement Section 9.6 is on pages 572–579.

Occasional cases arise in practice where the parameter value \( \theta \) of interest has itself come from an earlier random system. If the probability characteristics of this earlier system are known, then ordinary probability theory prescribes that we examine the combined system: the earlier system together with the present system. It also prescribes that we obtain the conditional distribution of what we have not observed (\( \theta \)) given what we know (the observed response \( y \)). Compare with the discussion in Sections 4.1F and 4.1G. The analysis of such combined systems, called Bayes analysis, is discussed in Supplement Section 9.6.

In other cases where there is no earlier system, just an unknown \( \theta \) of a current system, some statisticians suggest that we use a probability distribution to present our feelings and prejudices about \( \theta \) and then use the Bayes analysis just mentioned. Some discussion of this subjective Bayes analysis is included in the supplementary section.

NOTES AND REFERENCES

Following our pattern of approaching statistics from an open and general viewpoint, we have introduced this chapter on estimation by presenting some numerical procedures applicable to data without reference to a formal statistical model. We then introduced unbiased estimation and used the method of combining estimates to motivate the general Gauss–Markov theorem. The remainder of the chapter was then involved with the core material of traditional estimation theory: the Cramér–Rao information inequality developed in its essential, local form, and the Rao–Blackwell and Lehmann–Scheffé results on minimum variance unbiased estimation.

On least squares and the linear model:


On the information inequality:


On sufficiency and completeness:


On Bayes estimation:


In Chapters 7 and 8 we introduced estimation, tests of significance, and confidence intervals. Then in Chapter 9 we examined estimation in detail.

In Chapter 10 we shall examine in detail a formalized version of tests of significance. In Section 10.1 we investigate the formal framework for hypothesis testing and derive a fundamental lemma for obtaining a best test for a very simple problem. In Section 10.2 we use the lemma to obtain best tests for more general problems involving composite hypotheses and we use the likelihood statistic (or sufficiency) to obtain improved tests. In Section 10.3 we examine some general methods for obtaining tests: generalized likelihood ratio, unbiasedness, large sample, and sequential analysis. Finally, in Section 10.4 we examine tests for the validity of a statistical model itself, the chi-square tests of fit.

In Supplement Section 10.5 we examine confidence intervals and regions, specifically the formal ties with hypothesis testing.
10
TESTING STATISTICAL HYPOTHESES

10.1
THE FUNDAMENTAL LEMMA

Consider the biochemical example examined in Section 7.2B. Certain theoretical
calculations had suggested the value \( \mu = 14.8 \) for the mean response, and we were
interested in whether the observed data in Section 7.1E were in accord with this
suggested value. We calculated the standardized deviation of the sample average \( \bar{Y} \)
from the suggested mean 14.8, obtaining

\[
    z = \frac{16.8 - 14.8}{0.7} = 2.86.
\]

This observed value was moderately extreme as assessed against the standard normal
distribution. Accordingly, the value provided moderately strong evidence against the
suggested or hypothesized value 14.8 for the mean \( \mu \).

We noted that a value of \( z \) in the central range of the standard normal would be in
accord with the hypothesis; that a moderately extreme value would provide some
evidence against the hypothesis; and that a value very far out on the tails would
effectively contradict the hypothesis. Only in this last and very extreme case are we in
the position of having grounds for a near-categorical statement or decision concerning
the truth of the hypothesis.

In Chapters 7 and 8 we did not consider formally the question of how to choose a
function for assessing the departure from a hypothesis. One way of obtaining criteria
for such a choice is to introduce cutoff or critical values on the range of a function:
"accept" the hypothesis on one side of a critical value; "reject" the hypothesis on the
other side. We could then examine the probability of "accepting" the hypothesis and
the probability of "rejecting" the hypothesis under the various possibilities for the true
parameter. We would then have a basis for comparing various functions.

In Section 7.2G we mentioned some very serious aspects of making decisions in
this formal way—when typically there is not the evidence to support a decision. We
can, however, use the theory connected with choosing such a decision procedure to
guide us in our choice of function for making the tests of significance we have
discussed in preceding chapters. It is with this goal in mind that we now examine the
very attractive mathematical theory of hypothesis testing.
A EXAMPLE

Consider a response $y$ that has a normal($\theta$, $\sigma_0$) distribution with $\theta$ in $R$. And suppose that an investigator is concerned with the hypothesis $H_0: \theta \leq \theta_0$, where $\theta_0$ is a value indicated by some earlier investigations. The alternative to this hypothesis is given by the alternative hypothesis, $H_1: \theta > \theta_0$; typically this alternative is something the investigator is hoping or searching for. Let $(y_1, \ldots, y_n)$ be a sample from the response distribution just described.

The test-of-significance approach is to see if the data provide evidence against the hypothesis. Our earlier discussions suggest the use of the sample average and comparing it with the normal($\theta_0$, $\sigma_0/\sqrt{n}$) distribution.

In this section we consider this example from the point of view of making decisions. As part of this we examine the following test or decision procedure:

(1) accept $H_0$: if $\bar{y} < \theta_0 + 1.64\sigma_0/\sqrt{n}$;

(2) reject $H_0$: if $\bar{y} \geq \theta_0 + 1.64\sigma_0/\sqrt{n}$;

the value 1.64 is the value exceeded with probability 5 percent by the standard normal.

We can easily examine the performance characteristics of the preceding test procedure. Let $\Phi(\theta)$ be the probability that the test procedure leads to the decision "reject $H_0$":

$$
\Phi(\theta) = P(''\text{reject } H_0'')|\theta) = P(\bar{y} \geq \theta_0 + 1.64\sigma_0/\sqrt{n}|\theta).
$$

We call $\Phi(\theta)$ the power function of the test procedure; see Figure 10.1.

The power function $\Phi(\theta)$ in formula (2) is easily calculated numerically. Note that the decision "reject $H_0$" has preimage

$$
D = \{\bar{y}: \bar{y} \geq \theta_0 + 1.64\sigma_0/\sqrt{n}\}
$$
on the space of values for the function $\bar{y}$, and has preimage

$$
C = \left\{ y: \frac{\sum_{i=1}^{n} y_i}{n} \geq \theta_0 + 1.64\sigma_0/\sqrt{n} \right\}
$$
on the space for the sample $(y_1, \ldots, y_n)$; see Figure 10.2. The power function $\Phi(\theta)$ can most easily be calculated from the distribution of the function $\bar{y}$:

$$
\Phi(\theta) = P\left(\frac{\bar{y} - \theta}{\sigma_0/\sqrt{n}} \geq 1.64 + \frac{\theta_0 - \theta}{\sigma_0/\sqrt{n}} | \theta\right)
$$

$$
= P\left(z \geq 1.64 + \frac{\theta_0 - \theta}{\sigma_0/\sqrt{n}}\right)
$$

$$
= 1 - G\left(1.64 + \frac{\theta_0 - \theta}{\sigma_0/\sqrt{n}}\right),
$$

where $G$ is the distribution function for the standard normal. Note the normal-distribution-function shape in Figure 10.1.

If $H_0$ is true and $\theta \leq \theta_0$, then the decision "reject $H_0$" is incorrect; it is called a Type I error. For $\theta \leq \theta_0$ the function $\Phi(\theta)$ gives the probability for such a Type I error.

If $H_1$ is true and $\theta > \theta_0$, then the decision "accept $H_0$" is incorrect; it is called a Type II error. For $\theta > \theta_0$ the function $1 - \Phi(\theta)$ gives the probability for an incorrect decision. The probability of an incorrect decision is plotted in Figure 10.3.
FIGURE 10.1
Power function of the test procedure (1).

FIGURE 10.2
Set $D$ of values for $\bar{y}$ that lead to "reject $H_0";$ and corresponding set $C$ of values for $y$ that lead to "reject $H_0.""

B THE GENERAL FORMULATION

Consider a random system with response $y$. For the statistical model, let $\mathcal{S}$ be the sample space and $P(\cdot | \theta)$ be the probability measure for $y$, with parameter $\theta$ taking values in the parameter space $\Omega$. And suppose that the investigator has some hypothesis concerning the random system described by the preceding model. The investigator can consider each value $\theta$ and see whether the hypothesis holds for the corresponding description $P(\cdot | \theta)$. He can then assemble the values for which the hypothesis holds as the set $H_0$ on $\Omega$ and the values for which the hypothesis does not hold as the complementary set $H_1$. Thus the hypothesis induces a partition of $\Omega$ into two sets: $H_0$, called the null hypothesis, and $H_1$, called the alternative hypothesis.

The term "null hypothesis" derives from experimental applications concerned with cause-effect relationships. As part of determining whether there is evidence of a cause-effect relationship, an investigator considers a null hypothesis of no such
relationship; he then assesses whether the model under the null hypothesis provides an explanation for any apparent effects in the data. As part of this assessment certain principles of experimentation are necessary. We return to this experimental situation in Chapter 12.

For notation let \( d_0 \) designate the decision "accept \( H_0 \)" and \( d_1 \) designate the decision "reject \( H_0 \)." The possible decisions then are given by the decision space \( \{d_0, d_1\} \).

Now suppose that the investigator has selected a test procedure: then for each possible response value \( y \) he must have an associated decision, either \( d_0 \) or \( d_1 \). In effect, then, he has a decision function \( \delta \) which maps the sample space \( S \) into the decision space \( \{d_0, d_1\} \). It is convenient sometimes to think in terms of the preimage partition induced by \( \delta \) on the sample space \( S \). For this let \( C = \delta^{-1}(d_1) \) be the preimage of \( d_1 \), the set of sample points \( y \) that lead to the decision \( d_1 \). A test procedure \( \delta \) can thus be presented in terms of a partition of the sample space into two sets: \( C \), called the critical region, and \( C^c \), called the acceptance region.

We can formalize this in terms of the following definitions:

**DEFINITION 1**

A decision function is a (measurable) function \( \delta \) from the sample space \( S \) into the decision space \( \{d_0, d_1\} \).

**DEFINITION 2**

A critical region is a (measurable) set \( C = \delta^{-1}(d_1) \) on the sample space \( S \).

Now consider some theoretical aspects involved in choosing a decision function \( \delta \) or equivalently in choosing a critical region \( C \). For a test procedure, \( \delta \) or \( C \), the investigator will be interested in how it performs. This can be recorded in terms of the power function \( \Phi_{\delta} \):

\[
(3) \quad \Phi_{\delta} (\theta) = P(\text{"reject } H_0 \text{"} | \theta) = P(\delta(y) = d_1 | \theta) = P(C | \theta).
\]

The investigator would want a test procedure for which \( \Phi_{\delta}(\theta) \) is small for \( \theta \) in \( H_0 \) and large for \( \theta \) in \( H_1 \).
The performance characteristics of a test procedure can also be recorded in terms of the operating characteristic function \( OC_\delta \):

\[
OC_\delta(\theta) = P(\text{"accept } H_0 \text{"} | \theta) = P(\delta(y) = d_0 | \theta) = P(C^c | \theta).
\]

Clearly either function can be calculated from the other: as \( C \cup C^c = \mathbb{S} \), we have

\[
P_\delta(\theta) + OC_\delta(\theta) = 1.
\]

### C RELATED FUNCTIONS THAT DESCRIBE PERFORMANCE

If the hypothesis \( H_0 \) is true, then the decision \( d_1 \) is an incorrect decision; it is a Type I error or an error of the first kind. The probability of a Type I error with a decision function \( \delta \) is

\[
\alpha_\delta(\theta) = \Phi_\delta(\theta) \quad \text{for } \theta \text{ in } H_0,
\]

\[
= 0 \quad \text{for } \theta \text{ in } H_1.
\]

If the hypothesis \( H_1 \) is true, then the decision \( d_0 \) is an incorrect decision; it is a Type II error or an error of the second kind. The probability of a Type II error with a decision function \( \delta \) is

\[
\beta_\delta(\theta) = 0 \quad \text{for } \theta \text{ in } H_0,
\]

\[
= 1 - \Phi_\delta(\theta) \quad \text{for } \theta \text{ in } H_1.
\]

The probability for an incorrect decision is then given by

\[
\alpha_\delta(\theta) + \beta_\delta(\theta).
\]

This is plotted in Figure 10.3 for the test function (1).

In certain industrial applications it may be possible to determine the financial loss \( l(d_1, \theta) \) entailed by the decision \( d_1 \) when the parameter has the value \( \theta \). In such cases the investigator would be interested in the mean loss, or risk,

\[
\mathcal{R}_\delta(\theta) = E(l(\delta(y), \theta) | \theta) = l(d_1, \theta)\Phi_\delta(\theta) + l(d_0, \theta)(1 - \Phi_\delta(\theta)).
\]

The loss function may have the reasonable property that there is no loss with a correct decision; then

\[
l(d_0, \theta) = 0 \quad \text{for } \theta \text{ in } H_0,
\]

\[
= b(\theta) \quad \text{for } \theta \text{ in } H_1,
\]

\[
l(d_1, \theta) = a(\theta) \quad \text{for } \theta \text{ in } H_0,
\]

\[
= 0 \quad \text{for } \theta \text{ in } H_1.
\]

The risk function then has the form

\[
\mathcal{R}_\delta(\theta) = a(\theta)\alpha_\delta(\theta) + b(\theta)\beta_\delta(\theta).
\]

From a general point of view the investigator will want a test procedure for which the probability

\[
\alpha_\delta(\theta) + \beta_\delta(\theta)
\]

for incorrect decisions is small. Ideally he would want this to be small for each value of \( \theta \). In practice, however, he will usually find that the procedure that is best for certain \( \theta \) values is not best for other \( \theta \) values; he must then compromise in some way.
If a valid loss function is available, the investigator will want a test procedure for which the risk
\[(11) \quad R_\theta(\phi)\]
is small. Ideally he would want this small for each value of $\theta$. In practice he will usually find that the procedure that is best for certain $\theta$ values is not best for other $\theta$ values; he must then compromise in some way.

**D CONVENTIONAL CRITERIA FOR CHOOSING A TEST PROCEDURE**

Consider the problem of choosing a good test procedure on the basis of performance characteristics. The usual approach involves restricting attention to those test procedures that have a certain bound $\alpha$, say 5 percent or 1 percent, on the probability of a Type I error. The rationale is that deciding that a treatment effect exists when, in fact, it does not is serious and should be especially guarded against. Thus we restrict our attention to size $\alpha$ test procedures:

**DEFINITION 3**

A test procedure is a **size $\alpha$ test procedure** if
\[(12) \quad \mathcal{P}_\theta(\phi) = P(C|\theta) \leq \alpha \quad \text{for $\theta$ in $H_0$};\]
it is an **exact size $\alpha$ test procedure** if
\[(13) \quad \sup_{\theta \in H_0} \mathcal{P}_\theta(\phi) = \sup_{\theta \in H_0} P(C|\theta) = \alpha.\]

The test in Section A is an exact 5 percent test. It is also a 10 percent test, but as we shall see, a rather poor 10 percent test in comparison with other 10 percent tests.

The usual approach then is to search among the size $\alpha$ test procedures for one that maximizes the power,
\[(14) \quad \mathcal{P}_\theta(\phi) = P(C|\theta) \quad \text{for $\theta$ in $H_1$}.\]

Of course, the test that maximizes the power $P(C|\theta)$ for one $\theta$ value in $H_1$ may not be the test that maximizes it for another $\theta$ value. Some sort of a compromise is then needed. The criteria (12) and (14) are illustrated in Figure 10.4.

**E THE FUNDAMENTAL LEMMA**

A hypothesis-testing problem is particularly simple if the hypothesis consists of a single parameter value and the alternative consists of a single parameter value. We might, for example, be testing the hypothesis that a response is normal(100, 5) against the alternative that it is normal(110, 5).

We say that a null or alternative hypothesis is **simple** if it has one parameter value, and is **composite** if it has more than one parameter value. Accordingly, we now consider the problem of testing a simple hypothesis against a simple alternative. Of course, most practical problems involve testing a composite hypothesis against a
composite alternative. Fortunately, the solution of the special problem gives us the mechanics for obtaining solutions for many of the more general problems.

For the statistical model let $f(\cdot|\theta)$ be the density function with parameter $\theta$ in $\Omega = \{\theta_0, \theta_1\}$. We consider the problem of finding the best size $\alpha$ test of the hypothesis $H_0 = \{\theta_0\}$ against the alternative $H_1 = \{\theta_1\}$.

**LEMMA 4**

*Fundamental Lemma:* A test with critical region $C$ having

$$
\frac{L(y|\theta_1)}{L(y|\theta_0)} = \frac{f(y|\theta_1)}{f(y|\theta_0)} \geq k \quad \text{for } y \text{ in } C,
$$

$$
\leq k \quad \text{in } C^c
$$

is a most powerful test of $H_0 = \{\theta_0\}$ against $H_1 = \{\theta_1\}$ at some size $\alpha$. If the value $k$ and the set $C$ can be chosen to satisfy (exact size $\alpha$)

$$
\int_C f(y|\theta_0) \, dy = \alpha.
$$

then $C$ is most powerful among tests having size $\alpha$.

The hypothesis-testing problem is one of finding a set $C$ containing $\alpha$ of the $f(y|\theta_0)$ probability and then as much as possible of the $f(y|\theta_1)$ probability. Clearly points $y$ with large values of $f(y|\theta_1)/f(y|\theta_0)$ should go in $C$ first.

Also note that (15) does not define $C$ for points having the likelihood ratio equal to $k$; for these points any assignment to $C$ or $C'$ is permissible subject to (16).

The proof is given after the following simple example.

**EXAMPLE 1**

*The location normal:* Consider a sample $y = (y_1, \ldots, y_n)'$ from a normal($\theta$, $\sigma_0^2$) distribution with $\theta$ in $R$. To illustrate the lemma we take $\Omega = \{\theta_0, \theta_1\}$ and consider testing $H_0: \theta = \theta_0$ against $H_1: \theta = \theta_1$, where $\theta_0 < \theta_1$. By the lemma a most powerful test is to reject for large values of $L(y|\theta_1)/L(y|\theta_0)$ or, equivalently, to reject for large values of...
(17) \[ l(y|\theta_1) - l(y|\theta_0) = -\frac{1}{2\sigma_0^2} \sum (y_i - \theta_1)^2 + \frac{1}{2\sigma_0^2} \sum (y_i - \theta_0)^2 \]
\[ = \frac{\theta_1 - \theta_0}{\sigma_0^2} \sum \left( y_i - \frac{\theta_1 + \theta_0}{2} \right), \]
or equivalently \( \theta_1 > \theta_0 \) to reject for large values of the likelihood statistic \( \sum y_i \) or \( \bar{y} \).

The most powerful test can then be obtained by calculating the critical value exceeded with probability \( \alpha \) under the \( \theta_0 \) distribution. The hypothesis distribution of \( \bar{y} \) is normal \( \theta_0, \sigma_0/\sqrt{n} \). Hence the size \( \alpha \) test has critical region
\[ C = \{ y : \bar{y} \geq \theta_0 + z_\alpha \sigma_0/\sqrt{n} \}, \]
where \( z_\alpha \) is the value exceeded with probability \( \alpha \) under the standard normal distribution. Note that the test does not depend on \( \theta_1 \) and hence is most powerful for each \( \theta_1 \) value \( > \theta_0 \); it is thus uniformly most powerful UMP size \( \alpha \) against the alternative \( H_1 : \theta > \theta_0 \).

**Proof of Lemma 4** We present the proof for the absolutely continuous case; the discrete case is analogous. Consider a set \( C \) satisfying (15) and (16) for some \( \alpha \); and let \( D \) be any other size \( \alpha \) test; then
\[ \int_C f(y|\theta_0) \, dy \geq \int_D f(y|\theta_0) \, dy. \]
We remove the set \( C_0 = C \cap D \); let \( C^* = C - C_0 \) and \( D^* = D - C_0 \); hence
\[ \int_{C^*} f(y|\theta_0) \, dy \geq \int_{D^*} f(y|\theta_0) \, dy. \]
The points of \( C^* \) are inside \( C \); thus by (15),
\[ k f(y|\theta_0) \leq f(y|\theta_1). \]
The points of \( D^* \) are outside \( C \); thus, by (15),
\[ k f(y|\theta_0) \geq f(y|\theta_1). \]
We apply these inequalities to (18); we multiply through by \( k \), then substitute in accord with the inequalities and obtain
\[ \int_{C^*} f(y|\theta_1) \, dy \geq \int_{D^*} f(y|\theta_1) \, dy. \]
We now add back the integral over \( C_0 \), obtaining
\[ \int_C f(y|\theta_1) \, dy \geq \int_D f(y|\theta_1) \, dy. \]
This says that the power of \( D \) is less than or equal to the power of \( C \). Hence \( C \) has maximum power among tests of size \( \alpha \).

**F** A COMPLICATION WITH DISCRETE DISTRIBUTIONS

With a discrete distribution we can expect to have difficulty satisfying the exact size \( \alpha \) condition (16). For suppose that we are to reject for large values of a discrete variable
and find that under $H_0$ the discrete variable has, say, a Poisson(1) distribution. Then under $H_0$ we have

$$P(y \geq 3) = 0.080, \quad P(y \geq 4) = 0.019.$$  

Thus if we want a 5 percent test we must split the probability at the point $y = 3$. In effect, this can be accomplished by using a randomized test function.

**DEFINITION 5**

A randomized test function $\phi$ is a (measurable) real-valued function satisfying $0 \leq \phi(y) \leq 1$.

A test function $\phi$ is used as follows. If $\phi(y) = 1$, the decision is "reject $H_0$." If $\phi(y) = 0$, the decision is "accept $H_0." If $\phi(y) = a$, then use some random system and with probability $a$ choose the decision "reject $H_0" and with probability $1 - a$ choose the decision "accept $H_0." For a test with critical region $C$, note that the randomized test function $\phi$ is the indicator function of $C$.

The power function for a randomized test has the form

$$P_d(\theta) = E(\phi(y)|\theta) = \int \phi(y)f(y|\theta)\,dy.$$

We now present the more general lemma in which the best size $\alpha$ test always exists.

**LEMMA 6**

For testing $\theta_0$ against $\theta_1$, the test

$$\phi(y) = 1 \quad \text{for } \frac{L(y|\theta_1)}{L(y|\theta_0)} > k,$$

$$= a \quad = k,$$

$$= 0 \quad < k,$$

where $a$ and $k$ are chosen to satisfy (exact size $\alpha$)

$$\int \phi(y)f(y|\theta_0)\,dy = \alpha,$$

is most powerful among tests having size $\alpha$.

The proof is a straightforward extension of that for Lemma 4; see Problems 16 and 17.

The essential part of the lemma is that $a$ and $k$ can be chosen to satisfy the exact size $\alpha$ condition (21). For this, the Poisson example preceding the lemma suggests the method: define $\phi$ by

$$\phi(y) = 1 \quad \text{for } y \geq 4,$$

$$= \frac{0.050 - 0.019}{0.080 - 0.019} \quad y = 3,$$

$$= 0 \quad y \leq 2.$$
This appropriately splits the probability at \( y = 3 \) so that

\[
E(\phi(y) | \theta_0) = 1 \times 0.019 + \frac{0.050 - 0.019}{0.080 - 0.019}(0.080 - 0.019) = 0.050;
\]

thus the exact size 5 percent condition is satisfied.

**G EXERCISES**

1. The resistance was determined for five pieces of wire samples from production:
   \[0.138, \ 0.140, \ 0.136, \ 0.142, \ 0.140\]
   On the assumption that the variation is normal(0, 0.003), test at the 5 percent level the hypothesis \( H_0: \theta \leq 0.135 \), where \( \theta \) is the mean response.

2. The yield of a certain variety of corn was determined on seven plots of land:
   \[87.7, \ 87.2, \ 92.3, \ 94.2, \ 93.6, \ 85.6, \ 93.5\]
   On the assumption that the variation is normal(0, 4.2), test at the 5 percent level the hypothesis \( H_0: \theta \leq 85.0 \), where \( \theta \) is the mean response.

3. The breaking load was determined for a sample of six from the production run of a certain plastic:
   \[18.3, \ 19.3, \ 21.2, \ 22.4, \ 17.1, \ 27.7\]
   On the assumption that the variation is normal(0, 1.9), test at the 5 percent level the hypothesis \( H_0: \theta \geq 21.0 \), where \( \theta \) is the mean response.

4. Sixteen measurements in a chemical response yield \( \bar{y} = 285.3 \). On the assumption that the variation is normal(0, 3.1), test at the 5 percent level the hypothesis \( H_0: \theta \leq 280 \), where \( \theta \) is the mean response.

5. An investigator knows that a response is approximately normal with variance 1.44. Previously the mean had been 75, but a new treatment may have increased it. For testing the hypothesis that the mean \( \theta \) is 75 against the alternative that it is, say, 76, determine the most powerful size 1 percent test for a sample size of 10. Plot the power function.

6. (continuation) For a 1 percent test, the power at \( \theta = 76 \) depends on the sample size \( n \). Give an expression for the power and determine the minimum \( n \) so that there is a 95 percent chance of detecting a mean equal to 76.

**H PROBLEMS**

7. Let \( \{y_1, \ldots, y_n\} \) be a sample from the exponential distribution \( f(y|\theta) = \theta^{-1} \exp(-y/\theta) \) on \( R^+ \) (say, the interoccurrence times of Section 4.3A). Find the form of a most powerful test of \( H_0: \theta = \theta_0 \) against \( H_1: \theta = \theta_1 > \theta_0 \); the critical value can be presented in terms of a percentage point of a familiar distribution. Does the test depend on \( \theta_0 \)?

8. Let \( \{y_1, \ldots, y_n\} \) be a sample from the normal(\( \mu, \sigma \)) with \( \sigma \) in \( R^+ \).
   (a) Find the most powerful size \( \alpha \) test for \( H_0: \sigma = \sigma_0 \) against \( H_1: \sigma = \sigma_1 (> \sigma_0) \). Does the test depend on \( \sigma_0 \)?
   (b) Give an expression for the power function of the test in terms of a chi-square distribution function.

9. (continuation)
   (a) Find the most powerful size \( \alpha \) test for \( H_0: \sigma = \sigma_0 \) against \( H_1: \sigma = \sigma_1 (< \sigma_0) \). Does the test depend on \( \sigma_0 \)?
   (b) Give an expression for the power function of the test in terms of a chi-square distribution function.
Let \((y_1, \ldots, y_n)\) be a sample from the Poisson\((\theta)\) distribution. Find the form of the most powerful size \(\alpha\) test of \(H_0: \theta = \theta_0\) against \(H_1: \theta = \theta_1\) (<\(\theta_0\)). Does the test depend on \(\theta_1\)?

(a continuation) Find the form of the most powerful size \(\alpha\) test of \(H_0: \theta = \theta_0\) against \(H_1: \theta = \theta_1\) (<\(\theta_0\)). Does the test depend on \(\theta_1\)?

Let \((x_1, \ldots, x_n)\) be a sample from the Bernoulli\((p)\) distribution. Find the form of a most powerful size \(\alpha\) test of \(H_0: p = p_0\) against \(H_1: p = p_1\) (>\(p_0\)). Does the test depend on \(p_1\)?

(a continuation) Find the form of the most powerful size \(\alpha\) test of \(H_0: p = p_0\) against \(H_1: p = p_1\) (<\(p_0\)). Does the test depend on \(p_1\)?

Let \((y_1, \ldots, y_n)\) be a sample from the uniform\((0, \theta)\) distribution. Find the form of the most powerful size \(\alpha\) test of \(H_0: \theta = \theta_0\) against \(H_1: \theta = \theta_1\) (>\(\theta_0\)). Does the test depend on \(\theta_1\)?

(a continuation) Find the form of the most powerful size \(\alpha\) test of \(H_0: \theta = \theta_0\) against \(H_1: \theta = \theta_1\) (<\(\theta_0\)). Does the test depend on \(\theta_1\)?

### PROBLEMS: EXTENSIONS OF THE THEORY

16. **Part of Lemma 6:** Show that \(a\) and \(k\) can be chosen so that a test function \(\psi\) given by (20) satisfies the exact size \(\alpha\) condition (21). **Method:** Let \(G\) be the distribution function for

\[
\psi = \frac{f(y | \theta_1)}{f(y | \theta_0)}
\]

under the \(\theta_0\)-distribution for \(y\). Note that the \(\theta_0\)-probability for a zero denominator for \(u\) is zero; thus we need consider only the set where \(f(y | \theta_0) > 0\). Show that \(a\) and \(k\) can be determined in terms of characteristics of \(G\).

17. **Part of Lemma 6:** Let \(\psi\) satisfy (20) and (21) and let \(\psi\) be any other test of size \(\alpha\). Show that

\[
\Theta_\psi(\theta_1) = \Theta_\psi(\theta_1) = \int (\phi(y) - \psi(y)) f(y | \theta_1) dy \geq \int (\phi(y) - \psi(y)) k f(y | \theta_0) dy = k (\Theta_\phi(\theta_0) - \Theta_\psi(\theta_0)).
\]

**Method:** After dividing up the region of integration, use the likelihood ratio inequalities (20) together with the corresponding sign of \(\phi(y) - \psi(y)\).

18. **Extension of Lemma 6:** Let \(\psi\) be a most powerful size \(\alpha\) test of \(f(y | \theta_0)\) against \(f(y | \theta_1)\). Show that \(\psi\) has the likelihood ratio form;

\[
\psi(y) = \begin{cases} 
1, & f(y | \theta_1)/f(y | \theta_0) > k, \\
0, & k < f(y | \theta_1)/f(y | \theta_0) < k
\end{cases}
\]

and thus differs from the test \(\phi\) in (20) only at points where the likelihood ratio equals \(k\). **Method:** From size and power properties show that

\[
\int (\phi(y) - \psi(y))(f(y | \theta_0) - k f(y | \theta_0)) dy \leq 0,
\]

and from properties of \(\phi\) show that the integrand is \(\geq 0\).

19. **(continuation)** If \(E(\psi(y) | \theta_0) < \alpha\), then \(k = 0\) and the maximum power at \(\theta_1\) is unity.

20. **Generalization of Lemma 6:** Let \(f_0\) and \(f_1\) be real-valued integrable functions relative to a measure \(\mu\). Among test functions \(\phi\) for which

\[
\int \phi(y) f_0(y) d\mu(y) = c
\]

a necessary and sufficient condition that \(\phi\) maximize

\[
\int \phi(y) f_1(y) d\mu(y)
\]

is that \(\phi\) satisfy (23) and have the form...
\[ (24) \quad \phi(y) = 1, \quad f_i(y) > k f_i(y), \]
\[ = 0, \quad k f_i(y), \]

for some \( k. \)

21 Consider responses \( y_1, y_2, \ldots \) such that \((y_1, \ldots, y_k)\) has density function \( f_k(y_1, \ldots, y_k | \theta) \). If

\[ u_k = \frac{f_k(y_1, \ldots, y_k | \theta_0)}{f_k(y_1, \ldots, y_k | \theta)}, \]

show that \( E(u_{k+1}: y_1, \ldots, y_k | \theta_0) = u_k \). Thus the likelihood ratio is an example of a Martingale.

10.2

COMPOSITE HYPOTHESES

We are now in a position to examine some realistic hypothesis-testing problems. From Section 10.1 we have the solution for the problem of testing the hypothesis \( \theta = \theta_0 \) against the alternative \( \theta = \theta_1 \). As a problem in itself this is somewhat unrealistic, for it allows just two possible distributions for the response. We can, however, use the solution from the simple problem as a tool for solving more realistic problems, such as testing the following: the mean breaking strength \( \theta \) for cable from a new production process does not exceed 12,500 lb (available with other processes) against the alternative that it is greater than 12,500 lb, that is, \( H_0: \theta \leq 12,500 \) versus \( H_1: \theta > 12,500 \).

A THE LEAST FAVORABLE VALUE IN \( H_0 \)

We first consider the problem of testing a composite hypothesis against a simple alternative:

\[ (1) \quad H_0: \theta \text{ in } \omega \quad \text{vs.} \quad H_1: \theta = \theta_1. \]

We use \( \omega \) to designate a subset of the initial parameter space \( \Omega \); the problem as just presented has an effective parameter space \( \omega \cup \{ \theta_1 \} \).

The conventional approach to a hypothesis-testing problem was formulated in Section 10.1D. For the present problem we would search through the class of test functions \( \phi \) that satisfy the size \( \alpha \) condition,

\[ (2) \quad E(\phi(y) | \theta) = \int_{\omega} f(y | \theta) \, dy \leq \alpha \quad \text{for } \theta \text{ in } \omega, \]

and try to find the particular test that maximizes the power

\[ (3) \quad E(\phi(y) | \theta_1) = \int_{\omega} f(y | \theta_1) \, dy. \]

Our mathematical method with this problem is first to relax the condition (2), obtaining

\[ (4) \quad E(\phi(y) | \theta_0) = \int_{\omega} f(y | \theta_0) \, dy \leq \alpha, \]
where \( \theta_0 \) is some particular value in \( \omega \) deemed most difficult to distinguish from the alternative value \( \theta_1 \), a test function \( \phi \) satisfying condition (2) will satisfy the relaxed condition (4); thus the class of tests satisfying (4) is larger (\( \supset \)) than the original class satisfying (2). We look for the most powerful \((\theta_1)\) test in the larger class hoping that it will belong to the smaller class.

The mathematical method is to use the fundamental lemma 10.1.4 or 10.1.6 to determine the most powerful size \( \alpha \) for the modified problem

(5) \[ H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta = \theta_1. \]

If this most powerful test satisfies (2) rather than just (4) and is thus a size \( \alpha \) test for the original hypothesis \( \theta \in \omega \), then it is the most powerful size \( \alpha \) test of the original problem (1).

In such a case, in which our choice of \( \theta_0 \) value is successful, we call that \( \theta_0 \) value the least favorable value in \( \omega \) against the alternative \( \theta_1 \). The success of the method depends essentially on making a suitable choice for \( \theta_0 \). As a guiding principle we would choose the \( \theta_0 \) value that gives a distribution \( f(\cdot | \theta_0) \) "closest" to the alternative \( f(\cdot | \theta_1) \), or most difficult to distinguish from the alternative \( f(\cdot | \theta_1) \).

The hypothesis-testing problem (1) may, in fact, only be of secondary interest. Rather we may be interested in the more general problem

(6) \[ H_0 : \theta \in \omega \quad \text{vs.} \quad H_1 : \theta \in \Omega - \omega, \]

where the hypothesis and alternative are composite. If the most powerful (at \( \theta_1 \)) size \( \alpha \) test for the problem (1) does not depend on \( \theta_1 \) as it ranges over \( \Omega - \omega \), then it is the uniformly most powerful UMP size \( \alpha \) test for the more general problem (6).

### EXAMPLE 1

**The location normal**: Consider a sample \( y = (y_1, \ldots, y_n)' \) from a normal \((\theta, \sigma_0)\) distribution with \( \theta \) in \( R \). We consider the hypothesis-testing problem

(7) \[ H_0 : \theta \leq \theta_0 \quad \text{vs.} \quad H_1 : \theta > \theta_0. \]

We first determine the size \( \alpha \) test that has maximum power at the parameter value \( \theta_1 (>\theta_0) \). In effect, we are then considering the hypothesis-testing problem

(8) \[ H_0 : \theta \leq \theta_0 \quad \text{vs.} \quad H_1 : \theta = \theta_1. \]

Our approach is to choose a value in \( H_0 \) that is most difficult to distinguish from the alternative \( \theta_1 \); a natural first choice is the value \( \theta_0 \). We thus temporarily examine

(9) \[ H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta = \theta_1. \]

The most powerful test for this is available from Example 10.1.1:

(10) \[ C = \{ y : \bar{y} \geq \theta_0 + z_\alpha \sigma_0 / \sqrt{n} \}. \]

We must now check whether this test has size \( \alpha \) for the testing problem (8). The power function for \( \theta \leq \theta_0 \) is

\[ P_C(\theta) = P(\bar{y} \geq \theta_0 + z_\alpha \sigma_0 / \sqrt{n} | \theta) \]

\[ = P \left( \frac{\bar{y} - \theta}{\sigma_0 / \sqrt{n}} \geq \frac{\theta_0 - \theta}{\sigma_0 / \sqrt{n}} + z_\alpha | \theta \right), \]
which has maximum value \( \alpha \) when \( \theta_0 - \theta \) has its minimum value 0. Thus the test (10) has size \( \alpha \) for the testing problem (8) and is thus the most powerful size \( \alpha \) test for that problem (8).

Note that the test (10) does not depend on \( \theta_1 \) and hence is most powerful for each \( \theta_1 \) value \( > \theta_0 \); it is thus uniformly most powerful UMP size \( \alpha \) test for the original hypothesis-testing problem (7). See Figure 10.1 and note that we have found a test that has maximum power at each value of \( \theta \) larger than \( \theta_0 \). Thus we have accomplished what each of the arrows in Figure 10.4 indicates.

\section*{B The Least Favorable Distribution}

The method described in Section \( A \) may not work: there may be no \( \theta_0 \) value for which the corresponding test satisfies the size \( \alpha \) condition (2) for the full hypothesis: \( \theta \) in \( \omega \). The method can, however, be extended in a simple way. Let \( \lambda \) be a probability measure defined over the hypothesis \( \omega \): such a measure can produce a mixture or average density

\begin{equation}
(11) \quad f(y|\lambda) = \int_\omega f(y|\theta) \, d\lambda(\theta).
\end{equation}

We would like to choose \( \lambda \) so that the mixture is most difficult to distinguish from the particular alternative.

Our extended method is first to relax condition (2) obtaining

\begin{equation}
(12) \quad E(\phi(y)|\lambda) = \int_\mathcal{C} f(y|\lambda) \, dy \leq \alpha.
\end{equation}

A test function \( \phi \) satisfying (2) will satisfy the relaxed condition (12); use the Fubini Theorem 5.8.10. Thus the class of tests satisfying (12) is larger (\( \supset \)) than the original class satisfying (2). We look for the most powerful \( (\theta_1) \) test in the larger class hoping that it will belong to the smaller class.

The mathematical method is to use the fundamental lemma 10.1.4 or 10.1.6 to determine the most powerful size \( \alpha \) test of \( f(y|\lambda) \) against \( f(y|\theta_1) \). If this most powerful test satisfies (2) rather than just (12) and is thus of size \( \alpha \) for the original hypothesis \( \theta \in \omega \), then it is the most powerful size \( \alpha \) test of the original problem (1). In such a case the distribution \( \lambda \) is called least favorable over \( \omega \) against the alternative \( \theta_1 \).

If the most powerful \( (\theta_1) \) size \( \alpha \) test does not depend on \( \theta_1 \) in \( \Omega - \omega \), then it is UMP size \( \alpha \) for the more general problem (6).

Nontrivial illustrations can be rather involved and complicated; for an interesting illustration, see Problem 9. We present here a rather simple example.

\section*{Example 2}

For a sample \( y = (y_1, \ldots, y_n)' \) from the normal(\( \theta, \sigma^2 \)), consider the problem of testing the hypothesis \( H_0: \theta = \pm \delta_0 \) against the alternative \( H_1: \theta = 0 \). An intuitively reasonable choice for \( \lambda \) has probability \( 1/2 \) at each of \( +\delta_0, -\delta_0 \). The likelihood ratio test of \( f(y|\lambda) \) against \( f(y|0) \) is to reject for large values of
\[
\frac{f(y|\theta_0)}{f(y|\lambda)} = \frac{\exp\left\{-(1/2) \sum y_i^2\right\}}{(1/2) \exp\left\{-(1/2) \sum (y_i - \delta_0)^2\right\} + (1/2) \exp\left\{-(1/2) \sum (y_i + \delta_0)^2\right\}}
\]

or, equivalently, for small values of \(\exp\{n\delta_0 \bar{y}\} + \exp\{-n\delta_0 \bar{y}\}\) or, equivalently, for small values of \(|\bar{y}|\) (seen by plotting the sum of a positive and negative exponential). The size \(\alpha\) test of the mixture \(\lambda\) thus has critical region \(C = \{y: |\sqrt{n} \bar{y}| < k\}\), where \((\sqrt{n} \delta_0 - k, \sqrt{n} \delta_0 + k)\) has probability \(\alpha\) for the standard normal; note that \(\sqrt{n} \bar{y}\) has a distribution that is a 50:50 mixture of the normal \((\pm \sqrt{n} \delta_0, \pm 1)\) distributions, and note that the distribution of \(|\bar{y}|\) is the same from the normal(\(\delta_0, 1\) as from the normal(\(-\delta_0, 1\) distribution (and hence the same as from the 50:50 mixture). It follows that the test has size \(\alpha\) under the hypothesis \(H_0\), and hence that the test is most powerful size \(\alpha\) for \(H_0\).

**C MONOTONE LIKELIHOOD RATIO FORM**

Consider the case of a real-valued parameter \(\theta\) and the hypothesis-testing problem

(13) \(H_0: \theta \leq \theta_0\) vs. \(H_1: \theta > \theta_0\).

The method in Section A involves taking a value in the hypothesis, say \(\theta_0\), and a value in the alternative, say \(\theta_1\), and using the fundamental lemma to determine the most powerful test of \(\theta_0\) against \(\theta_1\). The resulting test is to reject for large values of the likelihood ratio,

(14) \(\frac{f(y|\theta_1)}{f(y|\theta_0)}\).

Now suppose that the method produces a uniformly most powerful test for (13) at each size \(\alpha\). Then it follows generally that large values of the likelihood ratio (14) will correspond to large values of a function \(t(y)\) for each value of \(\theta_1 > \theta_0\). For this we present the following definition.

**DEFINITION 1**

The statistical model \(\{f(\cdot|\theta): \theta \in R\}\) has **monotone likelihood ratio form** if there is a real-valued function \(t\) such that

(15) \(\frac{f(y|\theta_1)}{f(y|\theta_0)}\)

is a nondecreasing function of \(t(y)\) for each choice of \(\theta_0\) and \(\theta_1\) with \(\theta_1 > \theta_0\).

Note that Definition 1 implies that the likelihood statistic is a function of \(t(y)\) and thus that \(t(y)\) is a sufficient statistic. We recall from Section 8.5D that a real-valued likelihood statistic is intimately tied to models having exponential form with one
Thus a monotone likelihood ratio model itself is closely tied to models having exponential form with one $\psi$-function. As the prime example, consider the model
\[
f(y|\theta) = \gamma(\theta) \exp \left\{ t(y)\psi(\theta) \right\} h(y),
\]
where $\psi(\theta)$ is a nondecreasing function of $\theta$; this model has exponential form with one $\psi$-function. As specializations of this we have the usual location normal, scale normal, binomial, and Poisson examples.

The monotone likelihood ratio property gives immediately the following lemma.

**LEMMA 2**

For a model with monotone likelihood ratio form, a most powerful test of $\theta_0$ against $\theta_1$ has the form
\[
\phi(y) = \begin{cases} 
1 & \text{for } t(y) \geq t, \\
0 & \text{for } t(y) < t;
\end{cases}
\]
that is, reject for large values of $t(y)$ with possible randomization at the critical value $t$.

**Proof** The fundamental lemma says that a most powerful test is obtained by rejecting for large values of the likelihood ratio (15) or by Definition 1, for large values of the function $t(y)$; Problem 10.1.18 says a most powerful test necessarily has this form.

A monotone likelihood ratio model has monotone power for a likelihood ratio test:

**LEMMA 3**

For a model with monotone likelihood ratio form, the probability
\[
P(t(y) \geq t|\theta)
\]
is a nondecreasing function of $\theta$.

**Proof** From the monotone property in Definition 1, let $k$ be the value such that
\[
t(y) \geq t \quad \text{implies that} \quad \frac{f(y|\theta_2)}{f(y|\theta_1)} \geq k.
\]
\[
\leq k
\]
where $\theta_1 < \theta_2$. Consider the following sum:
\[
P(t(y) \leq t|\theta_1) + P(t(y) > t|\theta_1) = 1,
\]
\[
\int_{t(y) \leq t} kf(y|\theta_1) \, dy + \int_{t(y) > t} kf(y|\theta_1) \, dy = k.
\]
The integrand on the left is larger ($\geq$) than $f(y|\theta_2)$; and the integrand on the right is smaller ($\leq$) than $f(y|\theta_2)$. If we make these substitutions, we decrease the left integral and increase the right integral and obtain
FIGURE 10.5
As the distribution shifts to the right under the monotone likelihood ratio property,
the probability to the right of $t$ increases.

$$
\int_{t(y) \leq t} f(y \mid \theta_2) \, dy + \int_{t(y) > t} f(y \mid \theta_2) \, dy = 1.
$$

The ratio of right integral to left integral in (18) is altered by the substitution; the
ratio is increased ($\leq$). A comparison of (18) and (19) then shows that

$$
\int_{t(y) > t} f(y \mid \theta_1) \, dy \leq \int_{t(y) > t} f(y \mid \theta_2) \, dy
$$

and establishes the lemma. See Figure 10.5.

We now have the basic result for a monotone likelihood ratio model.

THEOREM 4

For a model with monotone likelihood ratio form the test

$$
\phi(y) = 1 \quad \text{if} \quad t(y) > k,
$$

$$
= a \quad \text{if} \quad k \leq t(y) \leq k,
$$

$$
= 0 \quad \text{if} \quad t(y) < k,
$$

where $k$ and $a$ satisfy

$$
E(\phi(y) \mid \theta_0) = \alpha,
$$

is a uniformly most powerful size $\alpha$ test of $H_0: \theta \leq \theta_0$ against $H_1: \theta > \theta_0$.

Proof. By Lemma 2 the most powerful size $\alpha$ test of $\theta_0$ against $\theta_1$ is given by (20)
with (21). Lemma 3 shows that the test has size $\alpha$ for the full hypothesis $\theta \leq \theta_0$; it
is thus the most powerful size $\alpha$ test for $\theta \leq \theta_0$ against $\theta_1$. The test does not
depend on $\theta_1$; it is thus UMP size $\alpha$ for $\theta \leq \theta_0$ against $\theta > \theta_0$. 
EXAMPLE 3

The location normal: Let \( y = (y_1, \ldots, y_n)' \) be normal(\( \theta, \sigma_0^2 \)) with \( \theta \) in \( R \). We show that the model has monotone likelihood ratio form:

\[
\frac{f(y|\theta_1)}{f(y|\theta_0)} = \exp \left\{ -\frac{1}{2\sigma_0^2} \sum (y_i - \theta)^2 \right\} \exp \left\{ -\frac{1}{2\sigma_0^2} \sum (y_i - \theta_0)^2 \right\} = \exp \left\{ \sum y_i \frac{\theta_1 - \theta_0}{\sigma_0^2} \right\} \exp \left\{ -n \frac{\theta_1^2 - \theta_0^2}{2\sigma_0^2} \right\},
\]

and this is a monotone increasing function of \( \Sigma y_i \) or of \( \bar{y} \). Then by Theorem 4 the UMP size \( \alpha \) test of \( \theta < \theta_0 \) against \( \theta > \theta_0 \) is to reject for large values of \( \bar{y} \) with critical value determined by the \( \theta_0 \) distribution. But this is just the test as determined in Example 1.

D ALTERNATIVES TO UMP

Our general approach has led us to look for UMP tests for hypothesis-testing problems. In Section C we have seen that UMP tests are closely tied to models having monotone likelihood ratio form. And these in turn are closely tied to models having exponential form with one \( \psi \)-function. Thus from Section 8.5D we see that the success of our approach is essentially limited to the location normal, the scale normal, the common binomial, and Poisson discrete models. What alternatives are available to us in other cases?

Consider the hypothesis-testing problem \( H_0: \theta \leq \theta_0 \) against \( H_1: \theta > \theta_0 \) or perhaps the simpler problem \( H_0: \theta = \theta_0 \) against \( H_1: \theta > \theta_0 \). If a UMP size \( \alpha \) test does not exist, then there are several directions in which we can proceed.

One possibility is to seek an important or distinctive value in the alternative and choose the most powerful test of \( \theta_0 \) against \( \theta_1 \):

\[
\phi(y) = 1 \quad \text{if} \frac{f(y|\theta_1)}{f(y|\theta_0)} > k,
\]

\[
= a \quad \text{if} \frac{f(y|\theta_1)}{f(y|\theta_0)} = k,
\]

\[
= 0 \quad \text{if} \frac{f(y|\theta_1)}{f(y|\theta_0)} < k,
\]

where \( k \) and \( a \) are chosen to give exact size \( \alpha \) under \( \theta_0 \). With a strategically chosen value \( \theta_1 \) the hope would be that the test had reasonable power for other \( \theta \) values in \( H_1 \).

Another possibility is to determine the size \( \alpha \) test satisfying

\[
\int \phi(y)f(y|\theta_0) \, dy \leq \alpha
\]

and maximizing the power for the \( \theta \) values just larger than \( \theta_0 \), that is, maximizing the slope of the power function at \( \theta_0 \). Such a test is called locally most powerful. The slope of the power function is available from (9.4.3) with appropriate regularity:

\[
\frac{d}{d\theta} \int \phi(y)f(y|\theta) \, dy = \int \phi(y)S(y|\theta)f(y|\theta) \, dy.
\]
The possibility then is to determine the size $\alpha$ test satisfying (23) and maximizing
\begin{equation}
\int \phi(y) S(y|\theta_0) f(y|\theta_0) \, dy.
\end{equation}

An obvious extension of the fundamental lemma is recorded in Problem 10.1.20; this extension says that the size $\alpha$ test of $H_0: \theta = \theta_0$ that maximizes the power slope is given by
\begin{align*}
\phi(y) &= 1 \quad \text{if } S(y|\theta_0) > k, \\
&= a \quad \text{if } S(y|\theta_0) = k, \\
&= 0 \quad \text{if } S(y|\theta_0) < k,
\end{align*}
where $k$ and $a$ are chosen to satisfy $E(\phi(y)|\theta_0) = \alpha$. With such a test the hope would be that good power for $\theta$ close to $\theta_0$ would mean reasonable power for other presumably more easily detected alternative values.

**EXAMPLE 4**

The location normal: Consider a sample $y = (y_1, \ldots, y_n)'$ from a normal$(\theta, \sigma_0^2)$ with $\theta$ in $R$. And suppose that we want the size $\alpha$ test of $\theta = \theta_0$ which has maximum power slope at $\theta_0$. The locally most powerful test is to reject for large values of
\begin{equation}
S(y|\theta_0) = \frac{\bar{y} - \theta_0}{\sigma_0^2/n}
\end{equation}
or equivalently for large values of $\bar{y}$. Thus we obtain, as we should expect, the test described in Examples 1 and 2.

**E TESTS BASED ON THE LIKELIHOOD STATISTIC**

The most powerful test of a simple hypothesis against a simple alternative is based on the likelihood ratio and accordingly is based on the likelihood statistic. Indeed, from Chapter 8 we know quite generally that a statistical procedure can reasonably be based completely on the likelihood statistic. We now see how this holds for general hypothesis-testing problems; we show that any test function $\phi$ has a corresponding test function $\psi$ based on the likelihood statistic that has the same power function.

Consider the power function of a test $\phi$ and let $s$ be the likelihood statistic; then by (5.5.3) we have
\begin{equation}
\Phi_\phi(\theta) = E(\phi(y)|\theta) = E(E(\phi(y):s)|\theta) = \Phi_\psi(\theta),
\end{equation}
where we let
\begin{equation}
\psi(s) = E(\phi(y):s).
\end{equation}
Thus the tests $\phi$ and $\psi$ have the same power function for all $\theta$ in $\Omega$. And, accordingly, we see that averaging over the conditional distribution given the likelihood statistic produces a test based on the likelihood statistic and with the same power function.
EXAMPLE 5

The location-scale normal: Consider a sample \( y = (y_1, \ldots, y_n)' \) from the normal(\( \mu, \sigma \)) with \( \theta = (\mu, \sigma) \) in \( R \times R^+ \). The preceding says that any test \( \phi \) based on \( (y_1, \ldots, y_n) \) has an equivalent test \( \psi \) based on the likelihood statistic \( (\bar{y}, s_y) \), equivalent in the sense of having the same power function. Thus we can note that it suffices to consider only the test based on the likelihood statistic \( (\bar{y}, s_y) \).

F EXERCISES

1. Location normal: Let \( (y_1, \ldots, y_n) \) be a sample from the normal(\( \theta, \sigma_0 \)) with \( \theta \) in \( R \). Determine the UMP size \( \alpha \) test of \( H_0: \theta \geq \theta_0 \) against \( H_1: \theta < \theta_0 \).

2. The scale exponential: Let \( (y_1, \ldots, y_n) \) be a sample from the exponential(\( \theta \)) distribution; Problem 10.1.7. Find the UMP size \( \alpha \) test of \( H_0: \theta \leq \theta_0 \) against \( H_1: \theta > \theta_0 \).

3. The scale normal: Let \( (y_1, \ldots, y_n) \) be a sample from the normal(\( \mu, \sigma \)) with \( \sigma \) in \( R^+ \); Problem 10.1.8. Find the UMP size \( \alpha \) test of \( H_0: \sigma \leq \sigma_0 \) against \( H_1: \sigma > \sigma_0 \).

4. (Continuation) Find the UMP size \( \alpha \) test of \( H_0: \sigma \geq \sigma_0 \) against \( H_1: \sigma < \sigma_0 \).

5. The Poisson(\( \theta \)) distribution: Let \( (y_1, \ldots, y_n) \) be a sample from the Poisson(\( \theta \)) with \( \theta \) in \( R^+ \); Problems 10.1.10 and 10.1.11. Find the UMP size \( \alpha \) test of

(a) \( H_0: \theta \leq \theta_0 \) against \( H_1: \theta > \theta_0 \).

(b) \( H_0: \theta \geq \theta_0 \) against \( H_1: \theta < \theta_0 \).

6. Binomial(\( n, p \)): Let \( (x_1, \ldots, x_n) \) be a sample from the Bernoulli(\( p \)) with \( p \) in \( [0, 1] \); Problem 10.1.12 and 10.1.13. Find the form of the UMP size \( \alpha \) test of

(a) \( H_0: p \leq p_0 \) against \( H_1: p > p_0 \).

(b) \( H_0: p \geq p_0 \) against \( H_1: p < p_0 \).

G PROBLEMS

7. Uniform(\( 0, \theta \)): Let \( (y_1, \ldots, y_n) \) be a sample from the uniform(\( 0, \theta \)); Problems 10.1.14 and 10.1.15. Find the UMP size \( \alpha \) test for \( H_0: \theta = \theta_0 \) against \( H_1: \theta \neq \theta_0 \); there is such an \( \alpha \) test in contrast to Exercises 1–6.

8. The general normal: scale: Let \( (y_1, \ldots, y_n) \) be a sample from a normal(\( \mu, \sigma \)) with \( (\mu, \sigma) \) in \( R \times R^+ \). Consider the composite hypothesis \( H_0: \sigma \geq \sigma_0, \mu \in R \) against the simple alternative \( H_1: \sigma = \sigma_1 (\sigma_0) \), \( \mu = \mu_1 \). As hypothesis distribution closest to the alternative, consider \( \sigma = \sigma_0, \mu = \mu_1 \).

(a) Show that the most powerful size \( \alpha \) test of \( (\mu_1, \sigma_0) \) against \( (\mu_1, \sigma_0) \) is to reject if

\[ \sum (y_i - \mu_1)^2 \leq \sigma_0^2 \chi^2_{1, \sigma_0} \],

where \( \chi^2_{1, \sigma_0} \) is the point exceeded with probability \( 1 - \alpha \) for the chi-square(\( n \)) distribution. Some straightforward analysis shows that the preceding test has correct size \( \alpha \) for the full hypothesis \( H_0: \sigma \geq \sigma_0, \mu \in R \).

(b) Deduce that a UMP size \( \alpha \) test does not exist for \( H_0: \sigma \geq \sigma_0 \) against \( H_1: \sigma < \sigma_0 \).

9. (Continuation) Consider the composite hypothesis \( H_0: \sigma \leq \sigma_0, \mu \in R \) against the simple alternative \( H_1: \sigma = \sigma_0 (\sigma_0) \), \( \mu = \mu_1 \). By Section E it suffices to examine tests based on \( (\bar{y}, \sum (y_i - \bar{y})^2) \). As a hypothesis mixture closest to the alternative, let \( \sigma^2 = \sigma_0^2 \) and let \( \mu \) be normal(\( \mu_1, (\sigma_0^2 - \sigma^2)/(n - 1) \)); this describes a probability measure \( \lambda \).

(a) For \( n \geq 2 \), show that the most powerful test of \( f(y | \lambda) \) against \( f(y | \mu_1,\sigma_0^2) \) is to reject if

\[ \sum (y_i - \bar{y})^2 \geq \sigma_0^2 \chi^2_{2, \sigma_0} \],

where \( \chi^2_{2, \sigma_0} \) is the \( \alpha \) point of the chi-square(\( n \)) distribution.

(b) Argue that the test is UMP size \( \alpha \) for \( H_0: \sigma \leq \sigma_0 \) against \( H_1: \sigma > \sigma_0 \).
10  The general normal; location: Let \( (y_1, \ldots, y_n) \) be a sample from the normal(\( \mu, \sigma \)) with \( (\mu, \sigma) \) in \( R \times R^+ \), and consider the problem \( H_0: \mu = 0, \sigma \in R^+ \) against \( H_1: \mu > 0, \sigma \in R^+ \). Some detailed analysis shows that for \( \alpha < 1/2 \), there does not exist a UMP test. Nevertheless, the test

\[
\phi(y) = 1(0) \quad \text{according as} \quad \sqrt{n} \bar{y} / s_y \geq (\leq) \ t_\alpha
\]

is a very appropriate test; \( t_\alpha \) is the value exceeded with probability \( \alpha \) by the Student(\( n - 1 \)) distribution. Use Problem 3.6.2 to obtain an expression for the power function of this test.

11  Let \( (y_{11}, \ldots, y_{1m}) \) be a sample from the normal(\( \mu_1, \sigma_0 \)) and \( (y_{21}, \ldots, y_{2n}) \) be a sample from the normal(\( \mu_2, \sigma_0 \)) with \( (\mu_1, \mu_2) \) in \( R^2 \).

(a) Find the most powerful size \( \alpha \) test of \( H_0: \mu_1 = \mu_2 \) against \( H_1: \mu_1 = \mu_0, \mu_2 = \mu_0 \) with \( \mu_0 > \mu_0 \). Method: As parameter value closest to \( H_1 \), try \( \mu_0 = (m\mu_1 + n\mu_2) / (m + n) \); the test is to reject for large values of \( \bar{y}_1 - \bar{y}_2 \).

(b) Find the UMP size \( \alpha \) test of \( H_0: \mu_1 = \mu_2 \) against \( H_1: \mu_1 > \mu_2 \).

12  Let \( (y_{11}, \ldots, y_{1m}) \) be a sample from the normal(0, \( \sigma_0 \)) and \( (y_{21}, \ldots, y_{2n}) \) be a sample from the normal(0, \( \sigma_2 \)) with \( (\sigma_1, \sigma_2) \) in \( (0, \infty)^2 \). Find the form of the most powerful test of \( H_0: \sigma_1 = \sigma_2 = \sigma \) against \( H_1: \sigma_1 = \sigma_1, \sigma_2 = \sigma_2 \), where \( \sigma_1 > \sigma_2 \). Show that \( \sigma_0 \) can be chosen so that the test is to reject for large values of

\[
F = \frac{\sum y_{1i}^2 / m}{\sum y_{2i}^2 / n},
\]

that is, for values of \( F \) that exceed the \( \alpha \) point of the \( F \) distribution(m, \( n \)). Conclude that the test: reject if \( F > F_\alpha \) is a UMP size \( \alpha \) test for \( H_0: \sigma_1 = \sigma_2 \) against \( H_1: \sigma_1 > \sigma_2 \).

13  (continuation) Show that the power function of the preceding test is given by

\[
1 - H_1 = 1 - \frac{\alpha_1^2 \sigma_0^2}{\alpha_2^2 \sigma_0^2}
\]

where \( H \) designates the distribution function for the \( F \) distribution(m, \( n \)).

14  Let \( (y_1, \ldots, y_n) \) be a sample from a distribution having a density function on \( R \). Let \( \xi \) designate a median of the density \( f \).

(a) Find the most powerful test of \( H_0: \xi \leq \xi_0 \) against the alternative \( f_1 \). Method: Write \( f_1 = \rho_1 f^* + q_1 f^* \), where \( f^* \) and \( f^* \) are the conditional densities given \( y < \xi_0 \) and given \( y > \xi_0 \) consider the hypothesis density \( f_0 = (1/2)f^* + (1/2)f^* \) as a density \"closest\" to \( f_1 \).

(b) Show that the sign test is UMP size \( \alpha \) for the problem \( H_0: \xi \leq \xi_0 \) against \( H_1: \xi > \xi_0 \); recall Section 7.48.

10.3  SOME OTHER METHODS IN HYPOTHESIS TESTING

For testing a statistical hypothesis we would naturally be interested in having a uniformly most powerful (UMP) test. But we have seen that such tests are rather rare. Indeed, for models where the region of positive density does not vary with \( \theta \), the occurrence is essentially limited to models having exponential form with one \( \psi \)-function. In practice, this almost limits such tests to the normal, binomial, and Poisson.
For testing a statistical hypothesis with the normal, binomial, and Poisson, we have obtained UMP tests for one-sided problems. In this section, we see that such tests do not exist for two-sided problems and we examine reasonable alternatives. We also examine several convenient methods for obtaining suitable tests for more general hypothesis-testing problems.

A TWO-SIDED TESTS FOR THE NORMAL

EXAMPLE 1

The normal with known variance: Let \((y_1, \ldots, y_n)\) be a sample from the \(\text{normal}(\theta, \sigma_0)\) with \(\theta \in \mathbb{R}\), and consider the hypothesis-testing problem \(H_0: \theta = \theta_0\) against \(H_1: \theta \neq \theta_0\). From Example 10.1.1 we know that the size \(\alpha\) test having maximum power for \(\theta\) values larger than \(\theta_0\) has critical region
\[
C_1 = \{y: \bar{y} \geq \theta_0 + z_{\alpha/2}\sigma_0/\sqrt{n}\}.
\]

Correspondingly, from symmetry we know that the size \(\alpha\) test having maximum power for \(\theta\) values less than \(\theta_0\) has critical region
\[
C_2 = \{y: \bar{y} \leq \theta_0 - z_{\alpha/2}\sigma_0/\sqrt{n}\}.
\]

The power functions of these two tests are plotted in Figure 10.6. And from the essential uniqueness of \(C_1\) and \(C_2\) (Problem 10.1.18) we can conclude that there is no UMP size \(\alpha\) test for the original problem.

We note that the test \(C_1\) rejects the hypothesis with probability \(<\alpha\) for \(\theta\) values to the left of \(\theta_0\). And correspondingly, \(C_2\) has the same inadequacy for \(\theta\) values larger than \(\theta_0\). One reasonable requirement that we can make is that the test reject with probability at least \(\alpha\) for \(\theta\) values in the alternative; such a test is called unbiased. This would seem to restrict us to tests with a horizontal tangent at \(\theta_0\). The results recorded in Section D below then establish the very reasonable test

\[
1 \quad C = \{y: |\bar{y} - \theta_0| \geq z_{\alpha/2}\sigma_0/\sqrt{n}\}
\]

as the UMP unbiased size \(\alpha\) test; see Figure 10.6. The corresponding test of significance was examined in Section 8.3A.

![Figure 10.6](image)

**FIGURE 10.6**

\(C_1\) has maximum power for \(\theta\) values \(>\theta_0\); \(C_2\) has maximum power for \(\theta\) values \(<\theta_0\); \(C\) has maximum power for \(\theta \neq \theta_0\) among tests with a horizontal tangent to the power function at \(\theta_0\).
EXAMPLE 2

The normal with unknown variance: Let \((y_1, \ldots, y_n)\) be a sample from the normal\((\mu, \sigma)\) with \((\mu, \sigma)\) in \(R \times R^+\), and consider the hypothesis-testing problem 
\(H_0: \mu = \mu_0, \sigma \in R^+\) against \(H_1: \mu \neq \mu_0, \sigma \in R^+\). We have noted in Problem 10.2.10 that there does not exist in general a UMP size \(\alpha\) test for either of the one-sided problems. Thus there certainly does not exist one for the two-sided problem just presented. Again a reasonable requirement is that the test reject with probability at least \(\alpha\) for \(\theta\) values in the alternative. The continuity of the power function then implies that such a test reject the hypothesis with probability exactly \(\alpha\) for \(\theta\) values in the hypothesis \(H_0: \mu = \mu_0, \sigma \in R^+\); such a test is called a similar test. The results quoted in Section E below then establish the very reasonable test

\[
C = \{y : |\bar{y} - \mu_0| \geq t_{\alpha/2}\bar{y}/\sqrt{n}\}
\]

as the UMP similar size \(\alpha\) test and as the UMP unbiased size \(\alpha\) test. The corresponding test of significance was examined in Section 8.3C.

B THE GENERALIZED LIKELIHOOD RATIO

The most powerful test of a simple hypothesis against a simple alternative is to reject for large values of the likelihood ratio

\[
\frac{L(y|\theta_1)}{L(y|\theta_0)} = \frac{f(y|\theta_1)}{f(y|\theta_0)}.
\]

A generalized version of this ratio has been proposed for general hypothesis-testing problems.

Consider a statistical model with density function \(f(\cdot | \theta)\) with \(\theta\) in \(\Omega\) and suppose that we are concerned with testing the hypothesis \(H_0\) against the alternative \(H_1 = \Omega - H_0\). The generalized likelihood ratio is defined by

\[
L(y) = \sup_{\theta \in \Omega} \frac{f(y|\theta)}{\sup_{\theta \in H_0} f(y|\theta)} = \frac{L(y|\hat{\theta})}{L(y|\hat{\theta}_0)},
\]

where \(\hat{\theta}(y)\) is the maximum-likelihood estimator in the full model with parameter space \(\Omega\) and \(\hat{\theta}(y)\) is the maximum-likelihood estimator in the special model with parameter space \(H_0\). The generalized likelihood ratio test is to reject the hypothesis for large values of \(L(y)\), where the critical value is determined so that the test has exact size \(\alpha\) for the hypothesis \(H_0\). Formula (3) used here parallels the ordinary likelihood ratio; other authors often use the reciprocal of (3).

This generalized likelihood ratio method has been remarkably successful in producing reasonable test procedures. It does, for example, give the compromise tests proposed in Examples 1 and 2. We now verify this for the more general Example 2.

EXAMPLE 3

The normal with unknown variance: Let \((y_1, \ldots, y_n)\) be a sample from the normal\((\mu, \sigma)\) with \((\mu, \sigma)\) in \(R \times R^+\), and consider the hypothesis-testing problem 
\(H_0: \mu = \mu_0, \sigma \in R^+\) against \(H_1: \mu \neq \mu_0, \sigma \in R^+\). The density function is
\( f(y \mid \mu, \sigma) = (2\pi \sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum (y_i - \mu)^2 \right\}. \)

For the full model there are two parameters \( \mu \) and \( \sigma \) and the maximum-likelihood estimators are

\[ \hat{\mu} = \bar{y}, \quad \hat{\sigma}^2 = n^{-1} \sum (y_i - \bar{y})^2; \]

thus

\[ f(y \mid \hat{\mu}, \hat{\sigma}) = \left( \frac{1}{2\pi} \sum \frac{(y_i - \bar{y})^2}{n} \right)^{-n/2} e^{-n/2}. \]

And for the special model of the hypothesis there is effectively one parameter \( \sigma \) (as \( \mu = \mu_0 \)) and the maximum-likelihood estimators are

\[ \hat{\mu} = \mu_0, \quad \hat{\sigma}^2 = n^{-1} \sum (y_i - \mu_0)^2; \]

thus

\[ f(y \mid \hat{\mu}, \hat{\sigma}) = \left( \frac{1}{2\pi} \sum \frac{(y_i - \mu_0)^2}{n} \right)^{-n/2} e^{-n/2}. \]

The generalized likelihood ratio is then given by

\[ L(y) = \left( \frac{\sum (y_i - \mu_0)^2}{\sum (y_i - \bar{y})^2} \right)^{n/2} = (1 + t^2/(n - 1))^{n/2}; \]

and the test is to reject for large values of \( L(y) \) or for large values of \( |t| \), where

\[ t = \frac{\bar{y} - \mu_0}{s_y/\sqrt{n}} \]

is the usual \( t \) function. This gives the test discussed in Example 2.

One difficulty with the use of the generalized likelihood ratio method is that of determining the critical value. Fortunately, the large-sample distribution is often available using results from Sections 8.6D and 8.6G. For this we assume the regularity conditions used in Section 8.6. And, in addition, we assume that \( \Omega \) is an open set in \( \mathbb{R}^r \) and that \( H_0 \) is an \((r - s)\)-dimensional region or surface in \( \Omega \); thus \( s \) is the number of effective parameters being tested or constrained by the hypothesis. It then follows from Section 8.6 and from properties of the multivariate normal that

\[ 2 \ln L(y) \]

has a limiting chi-square(s) distribution. Note that the alternative definition for \( L(y) \) would introduce a negative sign in (4).

For the generalized likelihood ratio in Example 3 note that \( 2 \ln L(y) \) is approximately \( t^2 \) for large \( n \) and that \( t^2 \) has a limiting chi-square(1) distribution.
C LARGE-SAMPLE TESTS

The test function based on the likelihood ratio or generalized likelihood ratio is usually easy to calculate but the critical value may be difficult to derive. Fortunately, as indicated in the preceding section, the large-sample distributions are usually available.

Let \((y_1, \ldots, y_n)\) be a sample from a distribution \(f(y|\theta)\) with \(\theta \in \mathbb{R}\).

The likelihood ratio for testing \(\theta_0\) versus \(\theta_1\) has the following logarithmic form for a single response \(y\):

\[
u = \ln f(y|\theta_1) - \ln f(y|\theta_0) = \ln \frac{f(y|\theta_1)}{f(y|\theta_0)}.
\]

Suppose that \(\mu_0\) and \(\sigma_0^2\) are the mean and variance of \(u\) under the hypothesis \(\theta = \theta_0\). Then for a sample of \(n\), the hypothesis \((\theta_0)\) distribution of the log-likelihood ratio

\[
\sum_{i=1}^{n} u_i = \ln \frac{f(y|\theta_1)}{f(y|\theta_0)}
\]

has limiting normal\((n\mu_0, \sqrt{n} \sigma_0)\) form; and the size \(\alpha\) test has the approximate form

\[
\phi(y) = \begin{cases} 1 & \text{if } \frac{f(y|\theta_1)}{f(y|\theta_0)} \geq n\mu_0 + z_\alpha \sqrt{n} \sigma_0, \\ 0 & \text{if } n\mu_0 + z_\alpha \sqrt{n} \sigma_0 < \frac{f(y|\theta_1)}{f(y|\theta_0)}. \end{cases}
\]

The locally most powerful test of \(\theta_0\) against larger values of \(\theta\) is based on the score function. The score function at \(\theta_0\) for a single response \(y\) is given by

\[
S(y|\theta_0) = \frac{\partial}{\partial \theta} \ln f(y|\theta) \bigg|_{\theta_0}.
\]

Under the regularity assumptions in Section 8.6 we have

\[
E(S(y|\theta_0)|\theta_0) = 0 \quad \text{var} (S(y|\theta_0)|\theta_0) = n(\theta_0).
\]

Then for a sample of \(n\) the hypothesis \((\theta_0)\) distribution of the score function

\[
S(y|\theta_0) = \sum_{i=1}^{n} S(y_i|\theta_0)
\]

has limiting normal\((0, \sqrt{n}j(\theta_0))\) form and the approximate size \(\alpha\) test has the form

\[
\phi(y) = \begin{cases} 1 & \text{if } S(y|\theta_0) \geq z_\alpha \sqrt{n}j(\theta_0), \\ 0 & \text{if } z_\alpha \sqrt{n}j(\theta_0) < S(y|\theta_0). \end{cases}
\]

For more general testing problems we might turn to the generalized likelihood ratio. The large-sample hypothesis distribution of the generalized likelihood ratio was recorded in Section B and the approximate size \(\alpha\) test thus has the form

\[
\phi(y) = \begin{cases} 1 & \text{if } 2 \ln L(y) \geq \chi^2_\alpha(1), \\ 0 & \text{if } \chi^2_\alpha(1) < 2 \ln L(y). \end{cases}
\]

where \(\chi^2_\alpha(1)\) is the \(\alpha\)-point on the right tail of the chi-square(1) distribution. See Example 3 and the concluding remarks in Section B.
**D SUPPLEMENT: UNBIASED TESTS**

In Example 1 we were looking for a suitable test for the two-sided problem. We noted that a reasonable requirement for a test was that it reject with probability at least \( \alpha \) for \( \theta \) values in the alternative. Accordingly, we define unbiased tests:

**DEFINITION 1**

A test with critical region \( C \) or test function \( \phi \) is **unbiased of size** \( \alpha \) for \( H_0 \) against \( H_1 \) if

\[
P(C|\theta) = E(\phi(y)|\theta) \leq \alpha \quad \text{for } \theta \text{ in } H_0,
\]

\[
\geq \alpha \quad \theta \text{ in } H_1.
\]

Now consider a real-valued parameter \( \theta \) and the hypothesis-testing problem \( H_0: \theta = \theta_0 \) against \( H_1: \theta \neq \theta_0 \). In many problems the power function is continuously differentiable with respect to \( \theta \). It then follows easily that an unbiased test has a power function with a horizontal tangent at \( \theta_0 \), and is thus a locally unbiased size \( \alpha \) test:

**DEFINITION 2**

A test with critical region \( C \) or test function \( \phi \) is a **locally unbiased size** \( \alpha \) test of \( H_0: \theta = \theta_0 \) against \( H_1: \theta \neq \theta_0 \) if

\[
P(C|\theta_0) = E(\phi(y)|\theta_0) = \alpha,
\]

\[
\frac{d}{d\theta} P(C|\theta) \bigg|_{\theta_0} = \frac{d}{d\theta} E(\phi(y)|\theta) \bigg|_{\theta_0} = 0.
\]

With some regularity the tangent condition (9) can be given the alternative form

\[
E(\phi(y)S(y|\theta_0)|\theta_0) = 0
\]

on the basis of equation (9.4.3). Local unbiasedness provides a mathematically flexible way of finding best unbiased tests.

The generalized fundamental lemma in Problem 10.1.20 can be used to obtain UMP unbiased size \( \alpha \) tests for the model

\[
f(y|\theta) = \gamma(\theta) \exp \{ t(y)\psi(\theta) \} h(y)
\]

where \( \psi(\theta) \) is continuously differentiable. We quote without proof the following theorem; for a proof, see Lehmann (1959).

**THEOREM 3**

For a model having exponential form (11), the UMP-unbiased size \( \alpha \) test for \( H_0: \theta = \theta_0 \) against \( H_0: \theta \neq \theta_0 \) has the form

\[
\phi(y) = 1 \quad \text{if } t(y) < t_1 \text{ or } t > t_2,
\]

\[
= 0 \quad t_1 < t(y) < t_2
\]
with possible randomization at \( t_1 \) and \( t_2 \) in the discrete case; the constants of the test are determined so that

\[
E(\phi(y) \mid \theta_0) = \alpha, \quad \frac{d}{d\theta} \left. E(\phi(y) \mid \theta) \right|_{\theta_0} = 0.
\]

**EXAMPLE 4**

*The normal with known variance:* Let \((y_1, \ldots, y_n)\) be a sample from the normal \((\theta, \sigma_0)\) with \(\theta\) in \(\mathbb{R}\), and consider testing \(\theta = \theta_0\) against \(\theta \neq \theta_0\). The density function can be written as

\[
f(y \mid \theta) = (2\pi\sigma_0^{-2})^{-n/2} \exp \left(-\frac{n\theta^2}{2\sigma_0^2}\right) \exp \left(\sum_i y_i \frac{\theta}{\sigma_0^2}\right) \exp \left(-\frac{1}{2\sigma_0^2} \sum_i y_i^2\right)
\]

and thus has the required exponential form. We seek a test based on \(\sum y_i\) or on \(\bar{y}\) having the form (12) and satisfying (13). Clearly the test with critical region

\[C = \{y: |\bar{y} - \theta_0| \geq z_\alpha \sigma_0 / \sqrt{n}\}\]

given in Example 1 has the required size \(\alpha\). Symmetry easily gives the unbiasedness; see Figure 10.6.

---

**E SUPPLEMENT: UNBIASED TESTS AND SIMILARITY**

Consider testing the hypothesis \(H_0\) against the alternative \(H_1\), and suppose that the power function for any test is a continuous function of \(\theta\). An unbiased size \(\alpha\) test satisfies conditions (8) and (9): the power is \(\leq \alpha\) on \(H_0\) and \(\geq \alpha\) on \(H_1\). From continuity it then follows that it is equal to \(\alpha\) for points on the boundary between \(H_0\) and \(H_1\). Let \(\omega\) be the boundary set of points on \(\Omega\); see Figure 10.7. Then we can note that an unbiased size \(\alpha\) test for \(H_0\) against \(H_1\) is a similar size \(\alpha\) test for \(\omega\):

![Diagram](image)

**FIGURE 10.7**

Hypothesis \(H_0: \mu \leq \mu_0\); alternative \(H_1: \mu > \mu_0\). The boundary set \(\omega = \{(\mu_0, \sigma) : \sigma \in \mathbb{R}^+, \mu > \mu_0\}\).
DEFINITION 4

A test with critical region \( C \) or test function \( \phi \) is a similar size \( \alpha \) test for \( \omega \) if

\[
P(C | \theta) = E(\phi(y) | \theta) = \alpha \quad \text{for } \theta \text{ in } \omega.
\]

The concepts of sufficiency and completeness can then be used to obtain UMP unbiased size \( \alpha \) tests for a model

\[
f(y | \theta, \psi) = \gamma(\theta, \psi) \exp \left\{ t(y)\theta + \sum_{i=1}^{r} s_i(y)\psi_i \right\} h(y)
\]

with \((\theta, \psi')\) in an open set in \( \mathbb{R}^{r+1} \). We quote the following theorems:

THEOREM 5

For a model having exponential form (15), the UMP unbiased size \( \alpha \) test for \( H_0: \theta \leq \theta_0 \) against \( H_1: \theta > \theta_0 \) has the form

\[
\phi(y) = \begin{cases} 1 & \text{if } t(y) > k(s), \\ 0 & \text{if } t(y) < k(s) \end{cases}
\]

with possible randomization at \( t(y) = k(s) \) in the discrete case; the critical value \( k(s) \) is chosen so that the test has conditional size \( \alpha \).

\[
E(\phi(y) : s | \theta_0) = \alpha,
\]

for each value of \( s \).

THEOREM 6

For a model having exponential form (15), the UMP unbiased size \( \alpha \) test for \( H_0: \theta = \theta_0 \) against \( H_1: \theta \neq \theta_0 \) has the form

\[
\phi(y) = \begin{cases} 1 & \text{if } t(y) < k_1(s) \text{ or } > k_2(s), \\ 0 & \text{if } k_1(s) < t(y) < k_2(s) \end{cases}
\]

with possible randomization at \( t(y) = k_1(s) \) and \( t(y) = k_2(s) \) in the discrete case; the critical values \( k_1(s) \) and \( k_2(s) \) are chosen so that the test is locally unbiased size \( \alpha \).

\[
E(\phi(y) : s | \theta_0) = \alpha, \quad \frac{d}{d\theta} E(\phi(y) : s | \theta) \bigg|_{\theta_0} = 0
\]

conditionally given \( s \), for each value of \( s \).

EXAMPLE 5

The normal with unknown variance: Let \((y_1, \ldots, y_n)\) be a sample from the normal\((\mu, \sigma)\) with \((\mu, \sigma)\) in \( \mathbb{R} \times \mathbb{R}^+ \), and consider testing \( \mu \leq \mu_0 \) against \( \mu > \mu_0 \) or testing \( \mu = \mu_0 \) against \( \mu \neq \mu_0 \). For simplicity we relocate and in effect examine the case \( \mu_0 = 0 \). The density can be written as
\[ f(y | \mu, \sigma) = (2\pi\sigma^2)^{-n/2} \exp \left\{ - \frac{n\mu^2}{2\sigma^2} \right\} \exp \left\{ \sum y_i \frac{\mu}{\sigma^2} + \sum \frac{y_i^2 - 1}{2\sigma^2} \right\}, \]

and this has the exponential form with \( t = \Sigma y_i \) and \( s = \Sigma y_i^2 \). Under the hypothesis the conditional distribution given \( s \) is uniform on the sphere \( s = \Sigma y_i^2 \) with \( s \) given. The density generally given \( s \) has the exponential form(11). The ordinary \( t \)-function \( \sqrt{n} \frac{y}{s} \) can be shown to be a one-to-one function of \( \Sigma y_i \) given the value of \( s \). The conditional distribution of \( \sqrt{n} \frac{y}{s} \) under the hypothesis given the value of \( y_i \) can fairly easily be seen to be the same as the marginal, that is, the Student\( (n - 1) \) distribution. It follows, then, fairly easily from Theorems 5 and 6 that \( C_1 = \{ y : \frac{y}{s} \geq t_{\alpha/2}s_y / \sqrt{n} \} \) gives the UMP unbiased size \( \alpha \) test for the first problem and \( C = \{ y : \frac{y}{s} \geq t_{\alpha} s_y / \sqrt{n} \} \) gives the UMP unbiased size \( \alpha \) test for the second problem.

These methods can be applied to the normal linear model (Section 9.2), giving the tests as examined in Section 11.4.

**F SUPPLEMENT: SEQUENTIAL ANALYSIS**

In Exercises 10.1.5 and 10.1.6 we considered a normal response with known variance and we were concerned with testing \( H_0 : \theta = 75 \) against \( H_1 : \theta > 75 \) at the 1 percent level. The problem was to determine the sample size so that the power was 0.95 at the alternative \( \theta = 76 \).

In this section we consider a sequential test of \( \theta_0 \) against \( \theta_1 \) such that observations are taken one by one until the \( \theta_1 \) to \( \theta_0 \) likelihood ratio becomes extreme—either above an upper bound, suggesting "reject," or below a lower bound, suggesting "accept." The mathematics is very attractive and there are approximations that make the method surprisingly easy to use. For applications the method seems particularly appropriate to industrial acceptance sampling of incoming manufactured items and to the developmental screening of new materials, drugs, and procedures.

Consider a response \( y \) and density function \( f(\cdot | \theta) \) with \( \theta \in \mathbb{R} \). The sequential test is concerned nominally with a simple hypothesis \( H_0 : \theta = \theta_0 \) against a simple alternative \( H_1 : \theta = \theta_1 \), but effectively with the more reasonable \( H_0 : \theta \leq \theta_* \) against \( H_1 : \theta > \theta_* \), where \( \theta_* \) is some intermediate value. After \( n \) observations the test is based on the likelihood ratio

\[
L_n(y_1, \ldots, y_n) = \frac{f(y_1 | \theta_1)}{f(y_1 | \theta_0)} \frac{f(y_2 | \theta_1)}{f(y_2 | \theta_0)} \cdots \frac{f(y_n | \theta_1)}{f(y_n | \theta_0)}
\]

and is to

- reject \( H_0 \) if \( L_n(y) \geq B \),
- continue sampling if \( A < L_n(y) < B \),
- accept \( H_0 \) if \( L_n(y) \leq A \)

where "continue" means to take another sample value and repeat the procedure; see Figure 10.8. The values \( A \) and \( B \) are chosen to give the test the desired properties in terms of size and power.

Alternatively, the test can be expressed in logarithmic form with the advantages
that changes in log-likelihood are additive; then after \( n \) observations the test is based on the log-likelihood

\[
\ell_n(y_1, \ldots, y_n) = \ln \frac{f(y_1|\theta_1)}{f(y_1|\theta_0)} + \cdots + \ln \frac{f(y_n|\theta_1)}{f(y_n|\theta_0)}
= z_1 + \cdots + z_n,
\]

where

\[
z = \ln f(y|\theta_1) - \ln f(y|\theta_0)
\]

and is to

\[
\begin{align*}
\text{reject } H_0 & \quad \text{if } \ell_n(y) \geq \ln B, \\
\text{continue} & \quad \ln A < \ell_n(y) < \ln B, \\
\text{accept } H_0 & \quad \ell_n(y) \leq \ln A.
\end{align*}
\]

Let \( N \) be the sample size at which a terminal decision "reject \( H_0 \)" or "accept \( H_0 \)" is made. \( N \) can vary in value and is a function on the countable product sample space for the sequence \( (y_1, y_2, \ldots) \).

The sequential test has the following optimum property established by Wald and Wolfowitz.
THEOREM 7

Among fixed and sequential decision procedures having error probabilities
\[ P(\text{reject } H_0|\theta_0) \leq \alpha, \quad P(\text{accept } H_0|\theta_1) \leq \beta \]
and finite \( E(N|\theta_0) \) and \( E(N|\theta_1) \), the sequential test (20) with error probabilities exactly \( \alpha \) and \( \beta \) minimizes both \( E(N|\theta_0) \) and \( E(N|\theta_1) \).

Consider an example.

EXAMPLE 6

Let \( y \) be a response that is \( \text{normal}(\theta, \sigma_0) \) and consider the sequential test of \( H_0: \theta = \theta_0 \) against \( H_1: \theta = \theta_1 \).

\[
z = \ln \frac{f(y|\theta_1)}{f(y|\theta_0)} = \frac{1}{2\sigma_0^2} ((y - \theta_0)^2 - (y - \theta_1)^2) = \frac{\theta_1 - \theta_0}{\sigma_0^2} \left(y - \frac{\theta_0 + \theta_1}{2}\right).
\]

Thus the test after \( n \) observations is to

- **reject** if \( \sum_1^n \left(y_i - \frac{\theta_0 + \theta_1}{2}\right) \leq \frac{\sigma_0^2 \ln B}{\theta_1 - \theta_0} \),
- **continue** if \( \frac{\sigma_0^2 \ln A}{\theta_1 - \theta_0} < \sum_1^n \left(y_i - \frac{\theta_0 + \theta_1}{2}\right) < \frac{\sigma_0^2 \ln B}{\theta_1 - \theta_0} \),
- **accept** if \( \sum_1^n \left(y_i - \frac{\theta_0 + \theta_1}{2}\right) \geq \frac{\sigma_0^2 \ln A}{\theta_1 - \theta_0} \).

Approximate values for \( A \) and \( B \) are available from the following lemma.

The critical steps in the proof of the fundamental lemma 10.1.4 can be used to prove the following lemma:

LEMA 8

The Wald sequential test with size \( \alpha \) and power \( 1 - \beta \) has

\[
A \approx \frac{\beta}{1 - \alpha}, \quad B \approx \frac{1 - \beta}{\alpha}.
\]

the exact values \( A \) and \( B \) give a tighter interval,

\[
\frac{\beta}{1 - \alpha} \leq A, \quad B \leq \frac{1 - \beta}{\alpha}.
\]

G PROBLEMS

1. Let \( y = (y_1, \ldots, y_n)^t \) be a sample from a \( \text{normal}(\theta, \sigma) \) with \( \theta \) in \( R \), and consider the hypothesis-testing problem \( H_0: \theta = \theta_0 \) against \( H_1: \theta \neq \theta_0 \). Determine the generalized likeli-
hood ratio $L(y)$ in Section B. How does the exact distribution of $L(y)$ compare with the limiting distribution recorded in Section B?

2. Let $(y_{11}, \ldots, y_{1m})$ be a sample of $m$ from the normal$(\mu_1, \sigma)$ and $(y_{21}, \ldots, y_{2n})$ be a sample of $n$ from the normal$(\mu_2, \sigma)$ with $(\mu_1, \mu_2, \sigma)$ in $R^2 \times R^+$. Consider the problem of testing the hypothesis $H_0: \mu_1 = \mu_2, \sigma \in R^+$ against the alternative $H_1: \mu_1 \neq \mu_2, \sigma \in R^+$. Show that the generalized likelihood ratio test is to reject for large values of $|t|$, where

$$t = \frac{\overline{y}_2 - \overline{y}_1}{s(1/m + 1/n)^{1/2}}, \quad s^2 = \frac{1}{m+n-2} \left( \frac{\sum (y_{1i} - \overline{y}_1)^2}{m} + \frac{\sum (y_{2i} - \overline{y}_2)^2}{n} \right);$$

recall Section 7.21 and Problem 8.3.9. How does the large-sample distribution (Section B) of $2 \ln L(y)$ compare with the exact distribution recorded in Problem 8.3.9?

3. Let $(y_{11}, \ldots, y_{1m})$ be a sample of $m$ from the normal$(\mu_1, \sigma_1)$ and $(y_{21}, \ldots, y_{2n})$ be a sample of $n$ from the normal$(\mu_2, \sigma_2)$ with $(\mu_1, \mu_2, \sigma_1, \sigma_2)$ in $R^3 \times (R^+)^2$. Consider the problem of testing $H_0: \sigma_1 = \sigma_2$ against $H_1: \sigma_1 \neq \sigma_2$. Show that the generalized likelihood ratio is a function of

$$F = \frac{\sum (y_{1i} - \overline{y}_1)^2/(m-1)}{\sum (y_{2i} - \overline{y}_2)^2/(n-1)};$$

the usual test is to reject if $F \leq F_{1-\alpha/2}$ or $F \geq F_{\alpha/2}$, where $F_{a}$ is the $\alpha$ point on the right tail of the $F(m-1, n-1)$ distribution. See Problems 10.2.12 and 10.2.13.

4. Consider the large-sample test given by formula (5). Suppose that $\nu$ has mean $\mu$ and variance $\sigma^2$ under the alternative $\theta$. Show that the power for a sample of $n$ has the limiting form

$$1 - G \left( z_{a} \frac{\sigma_0}{\sigma} - \frac{\mu - \mu_0}{\sigma_0/\sqrt{n}} \right),$$

where $G$ is the standard normal distribution function; let $\mu = \mu_0 + \delta/\sqrt{n}$ with $\delta$ fixed.

5. Use Theorem 3 to show that $C$ in Example 1 is UMP unbiased size $a$ for the two-sided normal with known variance.

6. For the model (11) having exponential form with one $\psi$-function, use the score properties in Section 9.4 to show that the local unbiasedness conditions (B) and (9) can be replaced by

$$E(\psi(y)|\theta_0) = \alpha, \quad E(\psi(x)|\theta_0) = \alpha E(\psi(y)|\theta_0).$$

7. Let $y$ be a response with the normal$(\mu_0, \sigma)$ distribution with $\sigma$ in $R^+$. Describe the sequential test of $\sigma_0$ against $\sigma_1 (\sigma_1 > \sigma_0)$ using error probabilities $\alpha = 0.05$ and $\beta = 0.05$; use the approximate formulas for $A$ and $B$.

8. Let $y$ be a response with the exponential$(\theta)$ distribution with $\theta$ in $R^+$. Describe the sequential test of $\theta_0$ against $\theta_1 (\theta_1 > \theta_0)$ with error probabilities $\alpha = 0.05$ and $\beta = 0.01$; use the approximate formulas for $A$ and $B$.

9. Let $y$ have a Poisson$($\theta$)$ distribution. Describe the sequential test of $\theta_0$ against $\theta_1 (\theta_1 > \theta_0)$ with error probabilities $\alpha = 0.01$ and $\beta = 0.01$; use the approximate formulas for $A$ and $B$.

10. Let $(y_1, \ldots, y_n)$ be a sample from the normal$(\mu, \sigma)$ and let $(w_1, \ldots, w_n)$ be an additional sample from the same normal, where $n$ depends on the first sample variance $s_y^2$. An estimator for $\mu$ is required with variance $\delta^2$ regardless of $\sigma^2$. Let us aim for $\delta^2$, a value to be chosen later just less than $\delta^2$. Given $s_y^2$, the estimator $\hat{\mu} = a\overline{y} + b\overline{w}$ with $a + b = 1$ has variance $(a^2/m + b^2/n)\sigma^2$. Argue that $a, b$, and $n$ can be chosen so that the estimated variance $(a^2/m + b^2/n)s_y^2 = \delta^2$. [For example, choose $n \geq 1$ as small as possible so that $s_y^2/(m + n)$ is just less than $\delta^2$; then argue that values of $a$ and $b$ can be found to give equality.]

11. (continuation) Prove that

$$z = \frac{a\overline{y} + b\overline{w} - \mu}{(a^2/m + b^2/n)^{1/2}a}$$
has a standard normal distribution given $s^2_0$ and hence that

$$t = \frac{\bar{y} + b\bar{w} - \mu}{\delta_0}$$

has a Student$(m - 1)$ distribution, which has $E(t) = 0$ and $\text{var}(t) = (m - 1)/(m - 3)$. Thus the estimator $a\bar{y} + b\bar{w}$ has a distribution with Student form, and mean $\mu$ and variance $\delta^2$, where we now choose $\delta^2_0 = (m - 3)\delta^2/(m - 1)$.

12 (continuation) For $m = 10$ give an expression for the minimum second sample size $n$ to attain a 95 percent confidence interval for $\mu$ of prescribed length $2l$.

10.4

TESTING A STATISTICAL MODEL

We have been examining a lot of estimators and tests for the parameters of a statistical model. But we have not yet stopped to question the statistical model itself. The underlying assumption for the use of these estimators and tests is that the model is a valid model for the particular application. Sometimes, however, we will have applications where we want to check the validity of the model itself.

In Chapter 7 we examined a variety of inference procedures based on symmetry and large samples. The underlying assumption for those procedures was that the system being investigated was random, that the data formed a sample from a distribution. Supplement Section 7.5 was concerned with testing randomness, with testing the validity of the assumption.

Now, in a similar way, for the estimators and tests based on parametric models we are concerned with testing a statistical model, with testing the validity of the assumptions we have been using. The tests we examine are the chi-square tests of fit. There are many other tests of fit but most are concerned with particular applications or are designed to detect particular departures from the model. The chi-square tests cover a broad range of models and they are simple and accessible.

A TWO EXAMPLES OF TESTS OF FIT

In Section 2.1E we examined the Rutherford and Geiger data involving 2608 counts on the number of $\alpha$-particles in a 7.5-second interval. The Poisson model is a fairly standard model for such frequency counts; indeed the frequency counts for radioactive disintegration form the classic application for the Poisson model.

We were concerned with whether the observed frequencies $f_i$ are in reasonable agreement with the mean frequencies for some Poisson model. We determined a fitted Poisson model, we calculated the mean frequencies for the fitted model, and we then examined the deviations

$$d_i = 2(\sqrt{f_i} - \sqrt{e_i})$$

and then the sum of squares of these deviations,

$$\chi^2 = \sum d_i^2 = 4(\sqrt{f_i} - \sqrt{e_i})^2 = 14.60.$$
We noted that the distribution for $\chi^2$ under the hypothesis of some Poisson distribution is approximately chi-square(10). The observed value 14.60 is a reasonable value for this chi-square distribution, and in this way we saw that the data are reasonable for the Poisson model.

In Section 2.2D we examined the Grummel and Dunningham data involving 250 observations on a continuous response. Some background information suggested that a normal distribution might be appropriate. To apply a chi-square type of test to continuous data it is necessary to group the observations and form cells, and thus obtain frequencies for various cells. The Grummel and Dunningham data were already grouped; some further grouping was done to avoid too-small frequencies.

We were concerned with whether the observed frequencies $f_i$ were in reasonable agreement with the mean frequencies for a normal model. We determined a fitted normal model, we calculated the mean frequencies for the fitted model, and we then examined the deviations,

$$d_i = 2(\sqrt{f_i} - \sqrt{\bar{e}_i}),$$

and then the sum of squares of these deviations,

$$\chi^2 = \sum d_i^2 = \sum 4(\sqrt{f_i} - \sqrt{\bar{e}_i})^2 = 3.89.$$

We noted that the distribution for $\chi^2$ under the hypothesis of some normal distribution is approximately chi-square(7). The observed value 3.89 is a reasonable value for this chi-square distribution, and in this way we saw that the data are reasonable for the normal model.

In this section we present the theory for these chi-square tests of fit and we examine a range of different applications. As part of this we obtain some tests for the parameters of a parametric model.

**B THE CHI-SQUARE TESTS OF FIT**

Consider the problem of testing a statistical model. In the discrete case, a model with parameter $\theta$ will record probabilities $p_i(\theta)$ for the various discrete sample points $x_i$. In the absolutely continuous case a model will record a density function on the sample space. As part of the chi-square approach we suppose that intervals or cells have been formed on the sample space; let $p_i(\theta)$ be the probability for the $i$th cell.

For a sample of $n$ let $f_i$ designate the frequency count of values at the $i$th sample point or in the $i$th cell. By Section 3.5B the distribution of the frequencies $(f_1, \ldots, f_k)$ is multinomial$(n, p_1(\theta), \ldots, p_k(\theta))$.

The multinomial distribution for large $n$ can be rather unwieldy to calculate. Fortunately, we have a variety of limiting distributions, which we now record for convenient reference. Applications continue with Section C.

**LEMMA 1**

As $n \to \infty$ the limiting distribution of $(t_1, \ldots, t_k)$,

$$t_1 = \frac{f_1 - np_1(\theta)}{(np_1(\theta))^{1/2}} \quad \ldots \quad t_k = \frac{f_k - np_k(\theta)}{(np_k(\theta))^{1/2}},$$

is the same as the conditional distribution of a sample of $k$ from the standard normal given the condition $\sum t_i \sqrt{p_i(\theta)} = 0$. 

LEMMA 2

As \( n \to \infty \) the limiting distribution of \((z_1, \ldots, z_k)\),

\[
(2) \quad z_1 = 2(\sqrt{f_1} - \sqrt{np_1(\theta)}), \ldots, z_k = 2(\sqrt{f_k} - \sqrt{np_k(\theta)}),
\]

is the same as the conditional distribution of a sample of \( k \) from the standard normal given the condition \( \sum_1^k z_i \sqrt{p_i(\theta)} = 0 \).

Some details of the proof are discussed in Section 1.

The limiting normal distributions just described lead immediately to limiting chi-square distributions.

THEOREM 3

As \( n \to \infty \), the limiting distribution of

\[
(3) \quad \chi^2 = \sum_1^k t_i^2 = \sum_1^k \frac{(f_i - np_i(\theta))^2}{np_i(\theta)} = \sum_1^k \frac{(f_i - e_i)^2}{e_i}
\]

and the limiting distribution of

\[
(4) \quad \chi^2 = \sum_1^k z_i^2 = \sum_1^k 4(\sqrt{f_i} - \sqrt{np_i(\theta)})^2 = \sum_1^k 4(\sqrt{f_i} - \sqrt{e_i})^2
\]

is chi-square(\( k - 1 \)).

The actual distribution of the \( \chi^2 \) is discrete; but the limiting distribution is continuous. The proof follows by using Lemma 6.3.11 and Theorem 6.1.3 together with the normal-to-chi-square results from Section 8.2C.

Now consider the case in which some parameters of the model are estimated or fitted to the data. For this suppose that the true cell probability vector \( p(\theta) = (\rho_1(\theta), \ldots, \rho_k(\theta))^\prime \) is in a small neighborhood of the point \( p(\theta_0) = (\rho_1(\theta_0), \ldots, \rho_k(\theta_0))^\prime \) and that in this neighborhood the model \( p(\theta) = (\rho_1(\theta), \ldots, \rho_k(\theta))^\prime \) is a continuously differentiable surface that is approximately linear with dimension given by the number \( s \) of coordinates of the effective parameter \( \theta = (\theta_1, \ldots, \theta_s)^\prime \).

THEOREM 4

As \( n \to \infty \) the limiting distribution of \((z_1, \ldots, z_k)\),

\[
(5) \quad z_1 = 2(\sqrt{f_1} - \sqrt{np_1(\hat{\theta})}), \ldots, z_k = 2(\sqrt{f_k} - \sqrt{np_k(\hat{\theta})}),
\]

where \( \hat{\theta} \) is obtained by minimizing \( \chi^2 \) in formula \( (4) \), is the same as the conditional distribution of a sample of \( k \) from the standard normal given the conditions

\[
(6) \quad \sum_1^k z_i \sqrt{p_i(\theta)} = 0, \quad \sum_1^k z_i \frac{1}{\sqrt{p_i(\theta)}} \frac{\partial p_i(\theta)}{\partial \theta_\alpha} = 0, \quad \alpha = 1, \ldots, s.
\]
And the limiting distribution of the minimum

$$\chi^2_2 = \sum_{i=1}^{k} \frac{(f_i - np_i(\theta_0))^2}{np_i(\theta_0)}$$

is chi-square\((k - 1 - s)\).

Some details of the proof are discussed in Section I. Maximum likelihood (using the multinomial) can also be used to determine the fitted parameter \(\theta\).

The \(t\)’s and the \(z\)’s provide two approximately equivalent ways of expressing the limiting normal form of the multinomial distribution. The use of the \(z\)’s, however, provides a single reexpression \(2\sqrt{f_i}\) of a frequency \(f_i\) rather than a reexpression \(f_i/\sqrt{e_i}\) that varies with the parameter value being examined. We will see that this permits the intercomparison of a succession of hypotheses and provides a corresponding decomposition of the basic chi-square; see Section H below. Results analogous to (7) are, of course, available for \(\chi^2_2\) based on the \(t\)’s; but the estimation should be done by maximum likelihood, and the chi-square decomposition just mentioned is not generally available.

C TEST OF A PROBABILITY MODEL

Consider observed frequencies \(f_1, \ldots, f_k\) and a model that specifies the corresponding probabilities \(p_1(\theta_0), \ldots, p_k(\theta_0)\). This is a probability model or a special statistical model in which there are no free parameters.

This probability model can be tested by calculating

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - np_i(\theta_0))^2}{np_i(\theta_0)} \quad \text{or} \quad \chi^2_2 = \sum_{i=1}^{k} 4\left(\sqrt{f_i} - \sqrt{np_i(\theta_0)}\right)^2$$

and comparing the value with the chi-square\((k - 1)\) distribution. Large values of the \(\chi^2\) would indicate unusual deviations of the frequencies from the mean frequencies specified by the model, and in turn would provide evidence against the model.

D TEST OF A BINOMIAL MODEL

Perhaps the simplest situation for the chi-square method of Section C is the binomial. We examine briefly how the chi-square method works in this simple situation.

Consider the case with \(k = 2\) and let \(y\) be the frequency for the first cell and \(n - y\) the frequency for the second cell. And let \(p\) and \(q = 1 - p\) be the corresponding probabilities with values specified by some theory; then the theory says that \(y\) has the binomial\((n, p)\) distribution.

For the first chi-square we have

$$\chi^2 = \frac{(y - np)^2}{np} + \frac{(n - y - nq)^2}{nq} = \frac{(y - np)^2}{npq}$$
and by Theorem 3 this has a limiting chi-square(1) distribution under the theory. This corresponds of course to using \((y - np)/(npq)^{1/2}\) as a standard normal variable as indicated by the results in Example 6.2.3, and is in agreement with the large-sample methods suggested in Section 8.6E.

For the second chi-square we have

\[
\chi^2_2 = 4(\sqrt{y} - \sqrt{np})^2 + 4(\sqrt{n - y} - \sqrt{nq})^2,
\]

and by Theorem 3 this has a limiting chi-square(1) distribution under the theory. This chi-square can be pictured conveniently on binomial probability paper which has axes that are marked in square-root units. Thus the points \((\sqrt{np}, \sqrt{nq})\) and \((\sqrt{y}, \sqrt{n - y})\) can be plotted directly using the values \(np\) and \(nq\) for the first point, and the values \(y\) and \(n - y\) for the second point; see Figure 10.9. The difference vector is

\[
(z_1/2, z_2/2) = (\sqrt{y} - \sqrt{np}, \sqrt{n - y} - \sqrt{nq}).
\]

By Lemma 2 the difference vector has a limiting normal\((0, 1/2)\) distribution on the tangent to the circle at the point \((\sqrt{np}, \sqrt{nq})\). Thus the observed difference vector is easily assessed in units of length \(1/2\).

The second chi-square based on root frequencies can also be examined in terms of an angular transformation:

\[
a = \sin^{-1} \frac{y}{n} - \sin^{-1} \sqrt{p},
\]

which is the angle \(a\) as pictured in Figure 10.9. Under the theory the angle \(a\) has a limiting normal distribution \((0, 1/2 \sqrt{n})\). The angular transformation is available in most collections of statistical tables, for example, Fisher and Yates (1949). It provides a convenient analytic equivalent of the graphical use of binomial paper.

**FIGURE 10.9**

The difference vector \((z_1/2, z_2/2)\) from \((\sqrt{np}, \sqrt{nq})\) to \((\sqrt{y}, \sqrt{n - y})\) can be assessed in units of length \(1/2\). The difference angle from \((\sqrt{np}, \sqrt{nq})\) to \((\sqrt{y}, \sqrt{n - y})\) can be assessed in units of length \(1/2 \sqrt{n}\).
**E TEST OF A PROBABILITY MODEL: AN EXAMPLE**

The following example from genetics will be used repeatedly in this section.

**EXAMPLE 1**

The breeding of a certain type of flower can produce offspring that can have a magenta \( M \) or a red \( R \) flower, and can have a green \( g \) or a red \( r \) stigma; thus the possibilities are \( Mg, Mr, Rg, \) and \( Rr \). For 220 offspring the data (#4) are recorded in the following contingency table:

<table>
<thead>
<tr>
<th></th>
<th>( g )</th>
<th>( r )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D )</td>
<td>117</td>
<td>31</td>
<td>148</td>
</tr>
<tr>
<td>( R )</td>
<td>55</td>
<td>17</td>
<td>72</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>172</td>
<td>48</td>
<td>220</td>
</tr>
</tbody>
</table>

A strong theory \( H_0 \) specifies the probabilities for the four kinds of offspring as 9/16, 3/16, 3/16, and 1/16. In the first display we present the corresponding mean frequencies \( \bar{e}_i \) and root mean frequencies \( \sqrt{e}_i \); in the last display we present the observed frequencies \( f_i \) and root frequencies \( \sqrt{f}_i \); the intermediate display records the differences \( \sqrt{f}_i - \sqrt{e}_i \):

<table>
<thead>
<tr>
<th></th>
<th>( g )</th>
<th>( r )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_0 )</td>
<td>123.75</td>
<td>41.25</td>
<td>165</td>
</tr>
<tr>
<td>( R )</td>
<td>41.25</td>
<td>13.75</td>
<td>55</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>165</td>
<td>55</td>
<td>220</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>( g )</th>
<th>( r )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D )</td>
<td>117</td>
<td>31</td>
<td>148</td>
</tr>
<tr>
<td>( R )</td>
<td>55</td>
<td>17</td>
<td>72</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>172</td>
<td>48</td>
<td>220</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>( g )</th>
<th>( r )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_0 )</td>
<td>11.12</td>
<td>6.42</td>
<td></td>
</tr>
<tr>
<td>( R )</td>
<td>6.42</td>
<td>3.71</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>( g )</th>
<th>( r )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D )</td>
<td>117</td>
<td>31</td>
<td>148</td>
</tr>
<tr>
<td>( R )</td>
<td>55</td>
<td>17</td>
<td>72</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>172</td>
<td>48</td>
<td>220</td>
</tr>
</tbody>
</table>

The observed value of chi-square is

\[
\chi^2 = 4((-0.30)^2 + (-0.85)^2 + (1.00)^2 + (0.41)^2) = 7.92.
\]
The distribution for $\chi^2$ under the strong theory $H_0$ is chi-square(3): the 5 percent point is 7.81; the 1 percent point is 11.3. Thus the observed $\chi^2$ gives some moderate evidence against the hypothesis and the strong theory $H_0$ may not be applicable.

F TEST OF A STATISTICAL MODEL
WITH FITTED PROBABILITIES

Consider observed frequencies $f_1, \ldots, f_k$ and a statistical model that specifies the corresponding probabilities $p_1(\theta), \ldots, p_k(\theta)$, where the parameter $\theta$ has $s$-dimensional freedom as described in Section B; assume that $s < k - 1$.

This statistical model can be tested by calculating

$$
\chi^2 = \sum_{i=1}^{k} 4(\sqrt{f_i} - \sqrt{np_i(\hat{\theta})})^2
$$

and comparing the value with the chi-square($k - 1 - s$) distribution; the fitted parameter can be obtained by minimizing chi-square as described in Theorem 4 (or by maximum likelihood using the multinomial). Large values of $\chi^2$ would indicate unusual deviations from the statistical model and in turn would provide evidence against the model.

Consider the case where one set of marginal probabilities in a contingency table is estimated. We reexamine the data in Example 1.

EXAMPLE 2

A somewhat weaker theory, $H_1$, allows probabilities $p_1$ and $q_1$ for the colors $M$ and $R$ corresponding to rows but keeps the independence between flower and stigma color and keeps the probabilities $3/4$ and $1/4$ for the colors $g$ and $r$. Thus the cells $Mg$, $Mr$, $Rg$, and $Rr$ have probabilities $(3/4)p_1$, $(1/4)p_1$, $(3/4)q_1$, and $(1/4)q_1$, respectively, with $p_1$ in $[0, 1]$.

The observed proportions for $M$ and $R$ are 148/220 and 72/220 and they give estimates (MLE) for $p_1$ and $q_1$. In the first display we present the corresponding mean frequencies $\bar{r}_i$ and root-mean frequencies $\sqrt{\bar{r}_i}$; in the last display we present the observed frequencies $f_i$ and root frequencies $\sqrt{f_i}$; the intermediate display records the differences $\sqrt{f_i} - \sqrt{\bar{r}_i}$; note that our fitted frequencies have come closer to the data:

<table>
<thead>
<tr>
<th>Frequencies</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$g$</td>
<td>$r$</td>
<td>Total</td>
<td>$g$</td>
<td>$r$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M$</td>
<td>111</td>
<td>37</td>
<td>148</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R$</td>
<td>54</td>
<td>18</td>
<td>72</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>165</td>
<td>55</td>
<td>220</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Root frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g$</td>
</tr>
<tr>
<td>$M$</td>
</tr>
<tr>
<td>$R$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$g$</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M$</td>
<td>0.28</td>
</tr>
<tr>
<td>$R$</td>
<td>0.07</td>
</tr>
</tbody>
</table>
The observed value of chi-square is

\[ \chi^2 = 4((0.28)^2 + (-0.51)^2 + (0.07)^2 + (-0.12)^2) = 1.43. \]

The distribution for \( \chi^2 \) under the theory 3, is chi-square(2): the 5 percent point is 6.0; the 1 percent point is 9.2. Thus the observed \( \chi^2 \) is a very reasonable value for the distribution and there is no evidence against the theory 3.

\section*{TEST OF INDEPENDENCE WITH A CONTINGENCY TABLE}

The chi-square methods in this section can provide a test for the statistical independence of two response variables. For this suppose that the two variables are discrete or that they have been made discrete by grouping into cells.

For the first variable let \( A_1, \ldots, A_k \) designate the \( k \) distinct possibilities and for the second variable let \( B_1, \ldots, B_l \) designate the \( l \) distinct possibilities; thus any observation on the pair of variables corresponds to a combination \( A_iB_j \). Then for \( n \) observations we can record the frequencies \( f_{ij} \) as in the following array:

\[
\begin{array}{c|ccc|c}
\hline
& B_1 & \ldots & B_l & \text{Total} \\
\hline
A_1 & f_{11} & \ldots & f_{1l} & m_1 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
A_k & f_{k1} & \ldots & f_{kl} & m_k \\
\hline
\text{Total} & n_1 & \ldots & n_l & n \\
\hline
\end{array}
\]

Let \( p_1, \ldots, p_k \) designate probabilities for the first variable (\( \Sigma p_i = 1 \)) and \( p'_1, \ldots, p'_l \) designate probabilities for the second variable (\( \Sigma p'_j = 1 \)). Then the hypothesis of independence specifies a model with cell probabilities

\[ p_{ij} = p_i p'_j; \]

this model has \( k - 1 \) free parameters for the rows and \( l - 1 \) free parameters for the columns.

The row probabilities can be estimated (MLE) by row proportions

\[ \hat{p}_i = \frac{m_i}{n}; \]

and the column probabilities by the column proportions

\[ \hat{p}_j = \frac{n_j}{n}. \]
Thus the fitted model has mean frequencies \( \hat{\mu}_i \hat{\mu}_j = \frac{m_i n_j}{n} \), as recorded in the following array:

\[
\begin{array}{cccc}
 & B_1 & \cdots & B_l & \text{Total} \\
A_1 & \frac{m_1 n_1}{n} & \cdots & \frac{m_1 n_l}{n} & m_1 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
A_k & \frac{m_k n_1}{n} & \cdots & \frac{m_k n_l}{n} & m_k \\
\text{Total} & n_1 & \cdots & n_l & n
\end{array}
\]

This statistical model specifying independence between rows and columns can be tested by calculating

\[
\chi_1^2 = \sum_{ij} \frac{(f_{ij} - \frac{m_i n_j}{n})^2}{\frac{m_i n_j}{n}}
\]

or

\[
\chi_2^2 = \sum_{ij} 4(\sqrt{f_{ij}} - \sqrt{\frac{m_i n_j}{n}})^2,
\]

and comparing the value with the chi-square\((k - 1)(l - 1)\) distribution; note that \(kl - 1 - (k - 1)(l - 1) = (k - 1)(l - 1)\). Large values of \(\chi^2\) would indicate unusual deviations from the independence model. The limiting distribution is sometimes slightly better approximated by using \(\chi_1^2\) than \(\chi_2^2\), particularly with the possibility of zero frequencies.

We now reexamine the data in Examples 1 and 2.

EXAMPLE 3

A still weaker theory, 3_2, allows probabilities \(p_1\) and \(q_1\) for the colors \(M\) and \(R\) and probabilities \(p_2\) and \(q_2\) for the colors \(g\) and \(r\), but independence between the flower color and stigma color.

The observed proportions for \(M\) and \(R\) are 148/220 and 72/220 and the observed proportions for \(g\) and \(r\) are 172/220 and 48/220; these provide estimates (MLE) for the row and column probabilities. In the first display we present the corresponding mean frequencies \(e_i\) and root-mean frequencies \(\sqrt{e_i}\); in the last display the observed frequencies \(f_i\) and root frequencies \(\sqrt{f_i}\); the intermediate display records the difference \(\sqrt{f_i} - \sqrt{e_i}\); note that our fitted frequencies have come still closer to the data.

<table>
<thead>
<tr>
<th>Frequencies</th>
<th>Root frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>(3_2)</td>
<td>(g)</td>
</tr>
<tr>
<td>(M)</td>
<td>115.71</td>
</tr>
<tr>
<td>(R)</td>
<td>56.29</td>
</tr>
<tr>
<td>Total</td>
<td>172</td>
</tr>
</tbody>
</table>
The observed value of chi-square is

\[ \chi^2 = 4((0.06)^2 + (-0.11)^2 + (-0.08)^2 + (0.16)^2) = 0.19. \]

The distribution for \( \chi^2 \) under the theory \( J_2 \) is chi-square(1): the 5 percent point is 3.84; the 1 percent point is 6.63. Thus the observed \( \chi^2 \) is a very reasonable value for the distribution, and there is no evidence against the theory \( J_2 \) (if anything, the observed value is rather small).

### H A Sucession of Hypotheses or Theories

The use of the second chi-square based on root frequencies allows intercomparisons within a succession of statistical models and provides a decomposition of the basic chi-square. We illustrate this by considering Examples 1, 2, and 3 in greater detail.

In the left column we record the root-frequency matrices for the fitted \( J_0 \), the fitted \( J_1 \), the fitted \( J_2 \), and the data \( \mathcal{D} \). On the right we record the successive difference matrices.

\[ p_1 = \begin{bmatrix} -0.58 & -0.34 \\ 0.93 & 0.53 \end{bmatrix} \quad SS_1 = 1.5978 \]

\[ p_2 = \begin{bmatrix} 0.22 & -0.40 \\ 0.15 & -0.28 \end{bmatrix} \quad SS_2 = 0.3093 \]
Root frequencies

\[ \begin{array}{cc}
\mathcal{J}_2 & g \\
M & 10.76 \\
R & 7.50 \\
\hline
\mathcal{D} & g \\
M & 10.82 \\
R & 7.42 \\
\end{array} \]

\[ p_3 = \begin{pmatrix}
0.06 & -0.11 \\
-0.08 & 0.16
\end{pmatrix}, \quad SS_3 = 0.0477 \]

The components (as vectors in \( \mathbb{R}^6 \)) and 4 times the squared lengths can be tabulated as in Sections 8.2B, 8.2C, and 9.2B; see Table 10.1.

The various components are approximately orthogonal, as is easily checked. They differ slightly from orthogonality because: (a) the fitted points lie on a sphere \( \Sigma (\sqrt{t_i})^2 = 220 \) with squared radius equal to 220; (b) the component models are represented by a curve and a surface on that sphere. Note, however, that the chi-square in Example 2,

\[ \chi^2 = 1.43 \approx 1.24 + 0.19 = 1.43, \]

and the chi-square in Example 1,

\[ \chi^2 = 7.92 \approx 6.39 + 1.24 + 0.19 = 7.82. \]

Thus we see that the original chi-squares have been separated into component chi-squares that are approximately orthogonal and refer to particular characteristics of the statistical model.

We can test for independence by examining 0.19 and comparing it with chi-square(1); the 5 percent point is 3.84; the 1 percent point is 6.63.

Assuming independence, we can test for the column probability \( p_2 = 3/4 \) by examining 1.24 and comparing it with chi-square(1); the 5 percent point is 3.84; the 1 percent point is 6.63.

And assuming the independence and the column probability we can test the row probability \( p_1 = 3/4 \) by examining 6.39 and comparing with chi-square(1); the 5 percent point is 3.84; the 1 percent point is 6.63. This test for the row probability is much more sensitive than the blanket test for the model in Example 1.

\[ \begin{array}{cccc}
\text{Source} & \text{Dimension} & \text{Component} & 4 \cdot (\text{squared length}) \\
\hline
p_1 = 3/4 & 1 & p_1 & 6.39 \\
p_2 = 3/4 & 1 & p_2 & 1.24 \\
\text{Independence} & 1 & p_3 & 0.19 \\
\end{array} \]
SUPPLEMENT: SOME DETAILS ON
THE LIMITING DISTRIBUTIONS

Consider \( y \) with a Poisson(\( \theta \)) distribution. Then it follows from Section 6.1D and Problem 6.1.10 that the density function for

\[
t = \frac{y - \theta}{\theta^{1/2}} \quad \text{or} \quad z = 2(\sqrt{y} - \sqrt{\theta})
\]

approaches the standard normal density function.

Now consider independent \( y_1, \ldots, y_k \), where \( y_i \) has the Poisson(\( np_i \)) distribution. Then the preceding shows that the density function for \( (t_1, \ldots, t_k) \) or \( (z_1, \ldots, z_k) \) as given by formula (1) or (2) approaches the density function for a sample of \( k \) from the standard normal.

The multinomial \( (n, p_1, \ldots, p_k) \) is obtained by conditioning \( (\Sigma y_i = n) \) the preceding joint Poisson distribution. The sample lattice points are distributed uniformly along the plane \( \Sigma y_i = n \). It follows that the conditional density along the surface \( \Sigma y_i = n \) has the limiting normal form as described in Section B.

The limiting distribution is derived from the preceding by applying the normal sampling results from Section 8.2B.

PROBLEMS

1. A die was tossed 1600 times:

<table>
<thead>
<tr>
<th>Event</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>301</td>
<td>308</td>
<td>340</td>
<td>214</td>
<td>196</td>
<td>241</td>
</tr>
</tbody>
</table>

Test the fit of the symmetric model having equal probabilities.

2. The progeny from a certain type of corn can be starchy \( S \) or sugary \( s \) and can be green \( G \) or white \( g \). The data for \( n = 3839 \) progeny are

\[
\begin{array}{c|cc}
  & G & g \\ \hline
  S & 1997 & 906 \\ 
  s & 904 & 32 \\
\end{array}
\]

Test the hypothesis that the probabilities for \( SG, Sg, sG, \) and \( sg \) are 9/16, 3/16, 3/16, and 1/16, respectively.

3. (continuation) Test the hypothesis that the two attributes are statistically independent.

4. (continuation) Separate the chi-squares so that the hypothesis in Problem 2 can be tested, assuming the hypothesis in Problem 3; comment on the validity of such a test with the given data.

5. The following data (Freund) on blood types were obtained with random samples from three racial groups:

<table>
<thead>
<tr>
<th></th>
<th>O</th>
<th>A</th>
<th>B</th>
<th>AB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race 1</td>
<td>176</td>
<td>148</td>
<td>96</td>
<td>72</td>
</tr>
<tr>
<td>Race 2</td>
<td>78</td>
<td>50</td>
<td>45</td>
<td>12</td>
</tr>
<tr>
<td>Race 3</td>
<td>15</td>
<td>19</td>
<td>8</td>
<td>7</td>
</tr>
</tbody>
</table>
Test the hypothesis that blood-type distribution is independent of racial group.

6 The following frequencies $f_i$ were obtained (Morris) for $d_i$, the death month in months after the birth month for publicly prominent people:

<table>
<thead>
<tr>
<th>$d_i$</th>
<th>-6</th>
<th>-5</th>
<th>-4</th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_i$</td>
<td>24</td>
<td>31</td>
<td>20</td>
<td>23</td>
<td>34</td>
<td>16</td>
<td>26</td>
<td>36</td>
<td>37</td>
<td>41</td>
<td>26</td>
<td>34</td>
</tr>
</tbody>
</table>

Test the hypothesis that the probabilities are equal.

7 (continuation) Test the hypothesis that the months $-1$ and 0 have a probability $\rho_1$ and the remaining months have a probability $\rho_2$ (note that $2\rho_1 + 10\rho_2 = 1$).

8 (continuation) Separate the chi-squares and test the hypothesis in Problem 6 assuming the hypothesis in Problem 7.

9 One hundred plants were classified according to two attributes: large $L$ or small $l$; white $W$ or colored $w$. The observed frequencies are

<table>
<thead>
<tr>
<th></th>
<th>$W$</th>
<th>$w$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L$</td>
<td>40</td>
<td>20</td>
<td>60</td>
</tr>
<tr>
<td>$l$</td>
<td>15</td>
<td>25</td>
<td>40</td>
</tr>
<tr>
<td>Total</td>
<td>55</td>
<td>45</td>
<td>100</td>
</tr>
</tbody>
</table>

(a) Test the model that specifies equal probabilities in the four cells.
(b) Test the model that specifies equal $W - w$ probabilities and independence between the attributes.
(c) Test the model that specifies independence allowing arbitrary $W - w$ and arbitrary $L - l$ probabilities.
(d) Record the three-way decomposition of the chi-square; what tests are appropriate?

SUPPLEMENTARY MATERIAL

Supplement Section 10.5 is on pages 579–584.

Confidence intervals have been an integral part of the development in Chapters 7, 8, and 9; some theoretical aspects of confidence intervals are examined in the supplementary section in terms of the hypothesis-testing theory of this chapter. Some additional discussion on confidence intervals for the probability $p$ in sampling from the Bernoulli$(p)$ is included also.

NOTES AND REFERENCES

Some aspects of tests of significance have been formalized in the theory of hypothesis testing developed in this chapter. This theory originated in the work of Neyman and Pearson (1933), and has been thoroughly expounded in Lehmann (1959). Tests of models using the ordinary chi-square test have been surveyed in Lancaster (1969).


In Chapters 9 and 10 we examined in detail the basic methods of statistical inference.

In Chapter 11 we shall examine statistical inference for the very common and widely used statistical models that involve a linear structure for location, the linear models. Much of the development can be done not just for normal error form but for arbitrary error form. In Section 11.1 we investigate the simple location-scale model and obtain inference methods for normal error and inference methods for arbitrary error. (The computer program for this is available.) In Section 11.2 some details of least squares concerning projections and residuals are assembled. Statistical inference for the regression or linear model is examined in Sections 11.3 and 11.4.
11 LINEAR MODELS

11.1 THE LOCATION-SCALE MODELS

We now examine a large and very important class of statistical models, the linear models. In preceding chapters we have repeatedly examined the location and the location-scale models using a normal distribution for the variation or error. These models are special cases of the linear model discussed in this chapter. In this first section we examine the location-scale model using an arbitrary distribution for the variation.

In Chapter 9 we examined some estimation results for the linear model for a vector response. In Sections 11.2 to 11.4 we investigate this linear model using an arbitrary distribution for the variation: in Section 11.2, some vector results; in Section 11.3, the linear model itself; and in Section 11.4, an example.

The location-scale model covers applications where the input variables to the system are kept constant. The unknowns then are the general level of the response and the scaling of the variation. It is also possible to include a parameter that allows a range of distributions for the variation.

The linear model more generally covers applications with a random system under varied conditions; in particular, it includes the standard models of experimental design where the input variables of the system are changed by design.

A THE LOCATION-SCALE MODEL

Consider a real-valued response $y$. Suppose that earlier experience with the system under investigation has identified the distribution that describes the variation in the response. Let $z$ designate a variable presenting this variation, and $f$ be the known density function for $z$. And suppose that under present conditions the response scaling of the variation is unknown and the general level of the response is unknown. Let $\sigma$ designate the response scaling and $\mu$ designate the general level of the response; this gives $y = \mu + az$.

Thus for the variation we have a probability model as given by the known density function $f$. And for the response we have the statistical model

$\{y = \mu + az : (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+\}$,

giving the set of possible functions for the response $y$ in terms of the variation $z$.

Consider the question of a standardized presentation for the variation. If $f$ can be standardized so that $E(z) = 0$, $\text{var}(z) = 1$, then the general level $\mu$ is the mean.
response and the scaling \( \sigma \) is the response standard deviation. Perhaps more realistically, if \( f \) is standardized so that the median of \( z \) is zero and the interval \((-1, +1)\) contains 68.26 percent of the probability (as with the standard normal), then the general level \( \mu \) is the median response and the interval \((\mu - \sigma, \mu + \sigma)\) contains 68.26 percent of the response values; call \( \sigma \) in this case the response standard error. This second standardization seems particularly appropriate with a distribution \( f \) that allows substantially more probability in the tails than the normal; especially so with a distribution such as the Cauchy, for which the mean and standard deviation do not exist.

Now consider a sequence of independent repetitions of the system under investigation. Let \( z = (z_1, \ldots, z_n)' \) designate the sequence for the variation, and \( y = (y_1, \ldots, y_n)' \) designate the corresponding sequence for the response. Then for the variation \( z \) we have the probability model as given by the density function

\[(1) \quad f(z_1) \cdots f(z_n)\]

on \( R^n \). And for the response we have the statistical model

\[(2) \quad \{y = \mu 1 + \sigma z: (\mu, \sigma) \in R \times R^+\},\]

giving the set of possible functions for the response \( y \) in terms of the variation \( z \).

**B DATA FOR \( n \) INDEPENDENT REPETITIONS**

Now consider an observed sequence \( y = (y_1, \ldots, y_n)' \) from \( n \) independent repetitions of the system under investigation. The observed response sequence is a function

\[(3) \quad y = \mu 1 + \sigma z\]

of a realized sequence \( z \) from the distribution (1) for variation.

What information does the observed \( y \) give concerning the realized \( z \) from the distribution? With no information concerning the values of \( \mu \) and \( \sigma \), it follows that

\[z = -\sigma^{-1}\mu 1 + \sigma^{-1}y\]

can be any point on the half-plane

\[(4) \quad \mathcal{L}^+(1; y) = \{a 1 + c y: a \in R, c \in R^+\}\]

hinged on the line \( \mathcal{L}(1) \) and passing through the observed response \( y \); see Figure 11.1. Thus \( \mathcal{L}^+(1; z) = \mathcal{L}^+(1; y) \) and all that we effectively observe concerning the realized \( z \) is the half-plane passing through it; we have no differential information concerning the position of \( z \) on that half-plane.

From Sections 4.1 and 4.2 we know that the proper description of the realized \( z \) is given by the conditional distribution given the observed information that \( z \) lies on the half-plane (4). We derive this in Section D, but first we look for convenient coordinates to describe a half-plane and convenient coordinates for a point on such a half-plane.

**C SIMPLE COORDINATES**

Consider simple coordinates for a half-plane \( \mathcal{L}^+(1; z) \) and simple coordinates for a point \( z \) on such a half-plane; see Figure 11.2.
FIGURE 11.1
The possible values for $z$ form the half-plane $\mathcal{L}^+(1; y) = \{a1 + cy: a \in \mathbb{R}, c \in R^+\}$ hinged on the line $\mathcal{L}(1)$ and passing through the observed $y$.

FIGURE 11.2
Unit orthogonal residual $d$ identifies the half-plane $\mathcal{L}^+(1; y)$. The point $z$ has coordinates $\bar{z}$ and $s$ relative to the basis vectors $1$ and $d$. 
The half-plane is hinged on the line \( L(1) \); accordingly, it is natural to use \( 1 \) as a basis vector. The vector \( z \) has projection \( \overline{z}1 \) on the line \( L(1) \) and projection \( z - \overline{z}1 \) on the orthogonal complement \( L^\perp(1) \); thus

\[
(5) \quad z = \overline{z}1 + (z - \overline{z}1) = \overline{z}1 + s(z)d(z),
\]

where the residual \( z - \overline{z}1 \) produces the unit residual \( d(z) \) and the residual length \( s(z) \),

\[
(6) \quad s^2(z) = |z - \overline{z}1|^2 = \sum (z_i - \overline{z})^2, \quad d(z) = \left( \frac{z_1 - \overline{z}}{s(z)}, \ldots, \frac{z_n - \overline{z}}{s(z)} \right)'.
\]

This suggests using the vectors \( 1 \) and \( d(z) \) as orthogonal basis vectors for a point on a half-plane; thus

\[
(7) \quad z = \overline{z}1 + s(z)d(z)
\]

with coordinates \((\overline{z}, s(z))\) relative to the orthogonal vectors \( 1 \) and \( d(z) \).

The points \( d(z) \) are at unit distance from the origin and lie in the \((n - 1)\)-dimensional subspace \( L^\perp(1) \). Thus the points \( d(z) \) form the unit sphere \( Q_{n-1} \) in \( L^\perp(1) \); see Figure 11.2.

Now consider the possible points \( z \) given the information provided by the observed response \( y \). From Section B we see that \( z \) and \( y \) must lie on the same half-plane hinged on the line \( L(1) \). Thus \( d(z) = d(y) \). For convenience we abbreviate \( d(y) \) as \( d \). We then see that the information provided by the observed response \( y \) is that \( d(z) = d \); thus we have the "observed" value \( d \) for the function \( d(z) \).

Now consider the relation (3) between the observed response \( y \) and the realized \( z \):

\[
y = \overline{y}1 + s(y)d = \mu 1 + \sigma z = \mu 1 + \sigma(\overline{z}1 + s(z)d)
\]

\[
= (\mu + \sigma \overline{z})1 + \sigma s(z)d.
\]

By equating coefficients relative to the basis vectors \( 1 \) and \( d \), we obtain

\[
\overline{y} = \mu + \sigma \overline{z},
\]

\[
s(y) = \sigma s,
\]

where we abbreviate \((\overline{z}, s(z)) = (\overline{z}, s)\).

\[\textbf{D} \quad \text{THE CONDITIONAL DISTRIBUTION}
\]

\[\text{GIVEN THE DATA}
\]

We now determine the marginal probability for what has been observed, the half-plane through \( z \) as given by \( d(z) = d \), and we determine the conditional probability for what cannot be observed, the position \((\overline{z}, s)\) of \( z \) on the observed half-plane.

The probability differential for the distribution on \( R^n \) is given by

\[
(9) \quad \prod_{i=1}^{n} f(z_i) \prod_{i=1}^{n} dz_i.
\]

To obtain the marginal and conditional distributions we make the change of variable (7) to the new coordinates \( \overline{z}, s, \) and \( d \). For the density function the substitution is trivial. For the differential we argue as follows. At the point \( z \) we can measure distance in the direction \( 1 \) by \( \sqrt{n} \overline{z} \) (note that the vector \( 1 \) has length \( \sqrt{n} \)). At \( z \) we can
measure distance in the orthogonal direction \( d \) by \( ds \). Then in directions orthogonal to the half-plane we can measure area or volume on the sphere formed by the points \( s d(z) \) with \( s \) fixed. Let \( da \) be area or volume on the unit sphere formed by the points \( d(z) \); then \( s^{n-2} da \) is area or volume on the sphere of radius \( s \). We thus have length or volume measures in three orthogonal directions and obtain

\[
\prod dz_i = \sqrt{n} \ d\vec{z} \cdot ds \cdot s^{n-2} \ da.
\]

Accordingly, the probability differential for the distribution on \( R^n \) can be reexpressed as

\[
\prod f(\vec{z} + sd_i) \sqrt{n} \ d\vec{z} \cdot ds \cdot s^{n-2} \ da.
\]

By integrating out \( \vec{z} \) and \( s \) we obtain the marginal density for \( d \):

(10) \[
k(d) = \int_0^\infty \int_0^\infty \prod f(\vec{z} + sd_i)s^{n-2} ds \sqrt{n} \ d\vec{z}.
\]

This then gives the following factorization of the probability differential (9):

\[
k(d) \ da \cdot k^{-1}(d) \prod f(\vec{z} + sd_i)s^{n-2} ds \sqrt{n} \ d\vec{z}.
\]

The probability differential is factored into the marginal probability for \( d \) times the conditional probability for \( (\vec{z}, s) \) given \( d \). Note that the norming constant \( k(d) \) is obtained by the integration (10) over the half-plane \( R \times R^+ \).

We are in a position to be much more general than has been indicated so far. Suppose that we allow a parameter \( \lambda \) with values in \( \Lambda \) for the distribution \( f = f_\lambda \) describing the variation. This then permits some imprecision in the information available concerning the distribution for variation.

The probability differential for the distribution on \( R^n \) is then given by

\[
\prod f_\lambda(z_i) \prod dz_i.
\]

This can be factored as before:

(11) \[
k_\lambda(d) \ da \cdot k_\lambda^{-1}(d) \prod f_\lambda(\vec{z} + sd_i)s^{n-2} ds \sqrt{n} \ d\vec{z}.
\]

into the marginal probability for the observed \( d \) times the conditional probability for \( (\vec{z}, s) \) given \( d \).

\section*{E \ inference concerning the parameters}

Now consider statistical inference from an observed response \( y \). As just described, we suppose that the distribution for the variation is given by one of the densities in the class

(12) \[
\{f_\lambda(z_1) \cdots f_\lambda(z_n) : \lambda \in \Lambda \}.
\]

and that the observed response \( y \) is obtained from the realized \( z \) by one of the functions in the class

(13) \[
\{ y = \mu 1 + \sigma z : (\mu, \sigma) \in R \times R^+ \}.
\]

The only characteristic of the realized \( z \) that is observed is the value of \( d(z) = d = d(y) \). The probability for this observed event is

(14) \[
k_\lambda(d) \ da.
\]
Thus we have the observed likelihood function \( L(d | \lambda) \) for the parameter \( \lambda \), where
\[
L(d | \lambda) = c k_\lambda(d)
\]
and the various values of \( c \) give the similarly shaped functions of \( \lambda \). The likelihood function alone can sometimes provide sharp discriminations. In certain cases the likelihood statistic for the model \( k_\lambda(\cdot) \) \( da \) may be accessible and the methods from Chapters 9 and 10 may be available for inferences concerning \( \lambda \).

Now suppose that a value for \( \lambda \) is given—either from background information, or from sharp inferences based on (15), or as a tentative value in a range of values based on (15).

With \( \lambda \) given we have a distribution
\[
k_\lambda^{-1}(d) \prod f(z + sd_j)s^{n-2}ds \sqrt{n} \, d\bar{z}
\]
describing the unknown position \((z, s)\) of the realized variation \( z \); the relation to the observed position of the response is
\[
\bar{y} = \mu + s\bar{z}, \quad s(y) = \sigma s.
\]
where \( \mu \) and \( \sigma \) are the general level and scaling for the response.

**F Inference Concerning \( \mu \)**

For inferences concerning \( \mu \) we can appropriately simplify the equations (17):
\[
\frac{\bar{y} - \mu}{s(y)} = \frac{\bar{z}}{s} = T \quad \text{or} \quad \frac{\sqrt{n} (\bar{y} - \mu)}{s_y} = \frac{\sqrt{n} \bar{z}}{s/\sqrt{n-1}} = t = \sqrt{n(n-1)} \, T.
\]
The marginal distribution of the ordinary \( t \)-function can be obtained from (16) by expressing \( \bar{z} \) in terms of \( s \) and \( t \) and then integrating out \( s \):
\[
g_\lambda(t : d) \, dt = k_\lambda^{-1}(d) \int_0^\infty \prod f_\lambda(s(t/\sqrt{n^2 - n} + d_i))s^{n-1} \, ds \cdot (n - 1)^{-1/2} \, dt.
\]
A hypothesis \( \mu = \mu_0 \) can be tested by calculating
\[
t = \frac{\sqrt{n} (\bar{y} - \mu_0)}{s_y}
\]
and comparing the calculated value with the distribution (19). This is an ordinary test of significance or hypothesis test as discussed in Chapters 7, 8, and 10, but in addition it is a conditional test given the value \( d(y) \) from the response. Thus appropriately here it uses more information, relevant information from the data.

A \( 1 - \alpha \) confidence interval for \( \mu \) can be obtained from a \( 1 - \alpha \) interval for \( t \):
\[
\int_{t_1}^{t_2} g_\lambda(t : d) \, dt = 1 - \alpha;
\]
the \( 1 - \alpha \) confidence interval for \( \mu \) is
\[
(\bar{y} - t_2 s_y/\sqrt{n}, \bar{y} - t_1 s_y/\sqrt{n}).
\]
This is an ordinary \( 1 - \alpha \) confidence interval as discussed in Chapters 7 and 8, but in addition it is a conditional confidence interval given the value \( d(y) \) from the response; note that \( E(1 - \alpha : d) = 1 - \alpha \). Thus it uses more of the relevant information from the data.
G INFERENCES CONCERNING $\sigma$

For inferences concerning $\sigma$ we can rewrite the second equation in (17):

$$
\frac{s(y)}{\sigma} = s \quad \text{or} \quad \frac{s_x}{\sigma} = \frac{s}{\sqrt{n - 1}} = s_x.
$$

The marginal distribution of this standard deviation $s_x$ can be obtained from (16) by integrating out $\bar{z}$:

$$
\int_{0}^{\infty} f_{\lambda}(\bar{z}) \prod_{i=1}^{n} f_{\lambda}(z_i - \bar{z}) \frac{1}{\sqrt{n}} d\bar{z} = (n - 1)^{n-1/2}s_x^{n-2} ds_x.
$$

A hypothesis $\sigma = \sigma_0$ can be tested by calculating

$$
\frac{s_x}{\sigma_0}
$$

and comparing the calculated value with the distribution (23). This is an ordinary test of significance, but in addition is conditional given $d(y)$ and thus uses more information from the data.

A 1 $- \alpha$ confidence interval for $\sigma$ can be obtained from a 1 $- \alpha$ interval for $s_x$:

$$
\int_{s_1}^{s_2} h_{\lambda}(s_x : d) ds_x = 1 - \alpha;
$$

the 1 $- \alpha$ confidence interval for $\sigma$ is

$$
(s_x/s_2, s_x/s_1).
$$

This is an ordinary 1 $- \alpha$ confidence interval, but in addition is conditional and thus uses more information from the data.

H EXAMPLE

Serum potassium determinations have been found to have variation with substantially more probability in the tails than the normal; a Student($\lambda$) distribution with $\lambda$ in the range 2 to 6 is indicated as an appropriate approximation. For a particular serum the following potassium determinations were obtained:

$$
\begin{align*}
4.6 & \quad 4.9 & \quad 5.1 & \quad 5.2 & \quad 5.2 \\
5.2 & \quad 5.2 & \quad 5.3 & \quad 5.3 & \quad 5.4 \\
5.4 & \quad 5.4 & \quad 5.4 & \quad 5.4 & \quad 5.4 \\
5.4 & \quad 5.4 & \quad 5.4 & \quad 5.5 & \quad 5.5 \\
5.5 & \quad 5.5 & \quad 5.5 & \quad 5.6 & \quad 5.6 \\
5.6 & \quad 5.7 & \quad 5.7 & \quad 5.8 & \quad 5.8
\end{align*}
$$

and we have

$$
\bar{y} = 5.3967, \quad s_y = 1.3526, \quad s_y = 0.25118.
$$

Let $\mu$ designate the median response and $\sigma$ designate the standard error such that $(\mu \pm \sigma)$ contains 68.26 percent of the probability. For this we standardize the distribution $f_{\lambda}$ for variation so that the median is zero and the interval $(-1, +1)$ contains 68.26 percent of the probability. As a distribution for variation we use the Student($\lambda$)
distribution with \( \lambda \) in the range \([1, \infty)\), which goes from the Cauchy to the normal. The Student distribution has median 0 but it has more probability in the interval \((-1, +1)\) than the 68.26 percent of the standard normal. Accordingly, let \( f_\lambda \) be the standardized Student(\( \lambda \)) distribution, that is, the Student(\( \lambda \)) distribution rescaled to have 68.26 percent probability in the interval \((-1, +1)\).

For the 30 determinations we have

\[
5.3967 = \mu + \sigma \bar{z},
\]

\[
1.3526 = \sigma s,
\]

where \( \bar{z} \) and \( s \) give the position of the realized variation \( z \) on the half-plane \( \mathcal{L}^+(1; \gamma) \) in \( \mathbb{R}^{30} \).

The likelihood function (15) for \( \lambda \) is recorded in Figure 11.3; it was obtained by the computer integration (10) over the upper half-plane. The values of the likelihood function are recorded relative to the value 1 for \( \lambda = \infty \).

The likelihood function has maximum likelihood estimate near \( \lambda = 3 \) and is not sharply discriminating at that point. A related set of 35 determinations had a likelihood function very sharply peaked near \( \lambda = 2 \). The combined likelihood function has maximum likelihood between 2 and 3 and indicates a range from approximately 1 to 5.

The density (19) for the ordinary \( t \)-function is plotted in Figure 11.4 for the values of \( \lambda = 1, 3, 5, \) and \( \infty \).

Consider now a 95 percent confidence interval for the median response \( \mu \). With \( \lambda = 3 \) the central 95 percent probability interval for \( t \) as obtained by computer integration is

\[
(t_1, t_2) = (-2.209, 1.126).
\]

Thus the 95 percent confidence interval for \( \mu \) is

\[
(\bar{y} - t_2 s_y / \sqrt{n}, \bar{y} - t_1 s_y / \sqrt{n}) = (5.345, 5.498).
\]

Note that the asymmetry in the interval for \( t \) represents the expected bias for the particular spacing of the 30 observations (several extreme values on the left-hand tail); the calculated confidence interval has this bias removed from the raw estimate 5.40.

More generally, we record intervals for \( t \) and \( \mu \) for \( \lambda = 1, 3, 6, 9, \infty \), as shown in Table 11.1.
FIGURE 11.4
Density function \((19)\) for the ordinary \(t\)-function is plotted for \(\lambda = 1, 3, 5,\) and \(\infty\).

TABLE 11.1
Intervals for \(t\) and \(\mu\)

<table>
<thead>
<tr>
<th>(\lambda)</th>
<th>Percentage</th>
<th>(t_1)</th>
<th>(t_2)</th>
<th>(\mu_1)</th>
<th>(\mu_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>90</td>
<td>-1.966</td>
<td>0.236</td>
<td>5.386</td>
<td>5.487</td>
</tr>
<tr>
<td></td>
<td>95</td>
<td>-2.220</td>
<td>0.461</td>
<td>5.376</td>
<td>5.498</td>
</tr>
<tr>
<td></td>
<td>99</td>
<td>-2.750</td>
<td>0.920</td>
<td>5.354</td>
<td>5.523</td>
</tr>
<tr>
<td></td>
<td>99.9</td>
<td>-3.437</td>
<td>1.549</td>
<td>5.326</td>
<td>5.554</td>
</tr>
<tr>
<td>3</td>
<td>90</td>
<td>-1.930</td>
<td>0.837</td>
<td>5.358</td>
<td>5.485</td>
</tr>
<tr>
<td></td>
<td>95</td>
<td>-2.209</td>
<td>1.126</td>
<td>5.345</td>
<td>5.498</td>
</tr>
<tr>
<td></td>
<td>99</td>
<td>-2.780</td>
<td>1.722</td>
<td>5.318</td>
<td>5.524</td>
</tr>
<tr>
<td></td>
<td>99.9</td>
<td>-3.499</td>
<td>2.482</td>
<td>5.283</td>
<td>5.557</td>
</tr>
<tr>
<td>6</td>
<td>90</td>
<td>-1.921</td>
<td>1.064</td>
<td>5.348</td>
<td>5.485</td>
</tr>
<tr>
<td></td>
<td>95</td>
<td>-2.218</td>
<td>1.376</td>
<td>5.334</td>
<td>5.498</td>
</tr>
<tr>
<td></td>
<td>99</td>
<td>-2.824</td>
<td>2.021</td>
<td>5.304</td>
<td>5.526</td>
</tr>
<tr>
<td></td>
<td>99.9</td>
<td>-3.589</td>
<td>2.844</td>
<td>5.266</td>
<td>5.561</td>
</tr>
<tr>
<td>9</td>
<td>90</td>
<td>-1.893</td>
<td>1.194</td>
<td>5.342</td>
<td>5.483</td>
</tr>
<tr>
<td></td>
<td>95</td>
<td>-2.199</td>
<td>1.517</td>
<td>5.327</td>
<td>5.498</td>
</tr>
<tr>
<td></td>
<td>99</td>
<td>-2.827</td>
<td>2.184</td>
<td>5.297</td>
<td>5.526</td>
</tr>
<tr>
<td></td>
<td>99.9</td>
<td>-3.620</td>
<td>3.033</td>
<td>5.258</td>
<td>5.563</td>
</tr>
<tr>
<td>(\infty)</td>
<td>90</td>
<td>-1.699</td>
<td>1.699</td>
<td>5.319</td>
<td>5.475</td>
</tr>
<tr>
<td></td>
<td>95</td>
<td>-2.045</td>
<td>2.045</td>
<td>5.303</td>
<td>5.490</td>
</tr>
<tr>
<td></td>
<td>99</td>
<td>-2.756</td>
<td>2.756</td>
<td>5.270</td>
<td>5.523</td>
</tr>
<tr>
<td></td>
<td>99.9</td>
<td>-3.659</td>
<td>3.659</td>
<td>5.229</td>
<td>5.564</td>
</tr>
</tbody>
</table>

The density \((23)\) for \(s_x\) obtained by computer integration is plotted in Figure 11.5 for \(\lambda = 1, 3, 5,\) and \(\infty\). With \(\lambda = 3\) the central 95 percent confidence interval for \(s_x\) is \((0.845, 1.768)\); the corresponding 95 percent confidence interval for \(\sigma\) is \((0.142, 0.297)\).
FIGURE 11.5
Density function (23) for the standard deviation $s_x$ is plotted for $\lambda = 1, 3, 5,$ and $\infty$.

### ADDITIONAL EXAMPLES

The methods discussed in this section work for any distribution $f$ that can be presented analytically or numerically to a computer for integration. Consider, however, two examples involving simple functions $f$.

#### EXAMPLE 1

**Uniform variation:** Consider a distribution for variation that is uniform(0, 1). Note then that $y = \mu + \sigma z$ presents $\mu$ as the lower limit for the response and $\sigma$ as the range for the response. Let $y$ be an observed response sequence. The conditional distribution of $(\bar{z}, s)$ given $d(s) = d$ is

$$k^{-1}(d) \prod f(\bar{z} + sd_i)s^{n-2} ds \sqrt{n} \ d\bar{z}$$

$$= cs^{n-2} ds \ d\bar{z}$$

$$0 < \bar{z} + sd_i < 1 \text{ for each } i$$

$$= n(n - 1)(d_{(m)} - d_{(i)})^{n-1}s^{n-2} \ d\bar{z} \ ds$$

inside the triangle formed by $(\bar{z}, s) = (0, 0), \ (1, 0), \ (-d_{(i)}/(d_{(m)} - d_{(i)}), \ 1/(d_{(m)} - d_{(i)}))$, and zero elsewhere. The marginal distributions for $t$ and $s_x$ are readily available.

#### EXAMPLE 2

**Normal variation:** Consider a distribution for underlying variation that is standard normal. Let $y$ be a sample of $n$ response values obtained by a transformation $y = \mu 1 + \sigma z$

of a realized sample $z$ from the standard normal. The conditional distribution of $(\bar{z}, s)$ given the observed $d$ is
\[
\begin{align*}
k^{-1}(d)(2\pi)^{-n/2} \exp \left\{ -\frac{1}{2} \sum (\bar{z} + s d_i)^2 \right\} s^{n-2} ds \sqrt{n} d\bar{z} &= c \exp \left\{ -\frac{n \bar{z}^2}{2} \right\} \exp \left\{ -\frac{s^2}{2} \right\} s^{n-2} ds \sqrt{n} d\bar{z} \\
&= \frac{\sqrt{n}}{(2\pi)^{1/2}} \exp \left\{ -\frac{n \bar{z}^2}{2} \right\} d\bar{z} \cdot \frac{1}{\Gamma\left(\frac{n-1}{2}\right)} \exp \left\{ -\frac{s^2}{2} \right\} \left(\frac{s^2}{2}\right)^{(n-1)/2-1} ds \frac{s^2}{2} ,
\end{align*}
\]
where we note that \( \Sigma d_i = 0, \Sigma d_i^2 = 1 \), and for integration we have used just the norming constants for the normal and gamma density functions. From the preceding expression we see that \( \bar{z} \) is normal \((0, 1/\sqrt{n})\), \( s^2 \) is chi-square \((n - 1)\), and \( \bar{z} \) and \( s \) are statistically independent. Note that this conditional distribution does not depend on \( d \) and is thus equal to the marginal distribution. This is a substantially shorter and more direct derivation than the usual derivation; compare with Section 8.2.

The distribution of the \( t \)-function, \( t = \sqrt{n} \bar{z}/s_x \), follows from the normal and chi distributions just discussed; it is the Student \((n - 1)\) distribution. The distribution of the scaling function \( s \) is, of course, the \( \text{chi}(n - 1) \) distribution.

Thus it follows that the tests and confidence intervals for \( \mu \) and \( \sigma \) will agree with those developed in Chapters 8, 9, and 10 by the methods rather special to the normal distribution. Recall the discussion in Section 8.5.

\section*{J PROBLEMS}

1. Let \((2.5, 2.8, 2.3)\) be an observed sequence for the model presented in Example 2. Determine a central 95 percent confidence interval (a) for \( \mu \); (b) for \( \sigma \).

2. Consider \( y = \mu 1 + \alpha z \), where \( z \) is a sample from the uniform \((0, 1)\); see Example 1. Show that the distribution of \( T = \bar{z}/s \) is given by

\[
(n - 1) \left( 1 + \frac{T + d_{11}}{d_{10} - d_{11}} \right)^{-n} \frac{dT}{d_{10} - d_{11}}
\]

on the interval \((-d_{11}, \infty)\).

3. (continuation) Show that the marginal distribution of \( s \) is given by

\[
n(n - 1)[(d_{10} - d_{11})^{n-2}s^{n-2} - (d_{10} - d_{11})^{n-1}s^{n-1}][d_{10} - d_{11}] ds
\]

on the interval \((0, 1/(d_{10} - d_{11}))\).

4. Let \((2.5, 2.8, 2.3)\) be an observed sequence for the model presented in Example 1. From Problem 2 determine the marginal distribution for \( T \) and derive a central 95 percent confidence interval for \( \mu \).

5. (continuation) From Problem 3 determine the marginal distribution for \( s \) and derive a central 95 percent confidence interval for \( \alpha \).

6. Consider the location-scale model with the following distribution for variation: \( f(\bar{z}) = \exp \{-\bar{z}\} \) on \((0, \infty)\) and zero elsewhere. Let \((y_1, \ldots, y_n)\) be an observed sequence.

(a) Determine the distribution describing \((\bar{z}, s)\).

(b) Determine the marginal distribution for \( T = \bar{z}/s \).

7. Show that the response distribution consistent with an observed \( d \) has probability differential

\[
k^{-1}(d) \prod f_k \left( \frac{\bar{y} + s(y)d_i - \mu}{\sigma} \right) \frac{s^{n-2}(y)}{s^n} ds(y) \sqrt{n} d\bar{y},
\]

where \((y, s(y))\) refers to possible responses with given \( d \). \textit{Method:} Use (8) with (11).
K COMPUTER PROBLEMS

A computer program that handles the calculations discussed in this section is available; see the Notes and References at the end of this section.

The Student(λ) distribution with $0 < \lambda \leq \infty$ provides a convenient family of distributions for variation: the family ranges from long-tailed distributions like the Cauchy with $\lambda = 1$ through to the short-tailed normal with $\lambda \to \infty$. Each distribution has median zero. The scaling, however, is more complicated. The normal has 68.26 percent probability in $(-1, +1)$. The Student(λ) distribution has 68.26 percent probability in $(-\lambda, +\lambda)$; some values of $\lambda$ are

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>1</th>
<th>3</th>
<th>6</th>
<th>10</th>
<th>15</th>
<th>25</th>
<th>$\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$</td>
<td>1.8367</td>
<td>1.1966</td>
<td>1.0903</td>
<td>1.0524</td>
<td>1.0343</td>
<td>1.0202</td>
<td>1.000</td>
</tr>
</tbody>
</table>

We shall speak of the standardized Student(λ) distribution if the Student(λ) distribution is rescaled so that $(-1, +1)$ contains 68.26 percent probability. Use the standardized Student family as the model describing variation.

For each of the following problems calculate the likelihood function for $\lambda$ and the density functions for $t$ and $s_x$. Form appropriate confidence intervals for the parameter $\mu$.

8 The British statistician William S. Gosset, who published under the pseudonym "Student," examined in 1908 the following data, which compared the additional hours of sleep gained by using two soporific drugs. The following data represent the difference in additional hours gained from 10 patients given the two drugs at different times:

1.2, 2.4, 1.3, 1.3, 0.0, 1.0, 1.8, 0.8, 4.6, 1.4

Report 90 percent confidence intervals for $\mu$.

9 In 1876, Charles Darwin examined the effects of cross- and self-fertilization in plants. The data from his experiment are given here as the differences (in eights of an inch) between cross- and self-fertilized plants of the same pair:

49 − 67 8 16 6 23 28 41
14 29 56 24 75 60 − 48

Report 90 percent confidence intervals for $\mu$.

10 The following 35 determinations on serum potassium were obtained:

5.20 5.20 5.20 5.20 5.20
5.20 5.20 5.10 5.10 5.10
5.00 5.00 5.00 5.00 5.00
4.90 6.60 6.10 6.10 5.30
5.30 5.30 5.30 5.20 4.90
4.90 4.80 4.80 4.80 4.80
4.80 4.70 4.70 4.60 4.50

Report 95 percent confidence intervals for $\mu$.

11 The weight of coating, in hundredths of an ounce per square foot, is measured for 30 sheets:

160 153 158 155 160
159 148 153 138 150
160 138 173 160 148
137 170 154 147 164
160 159 162 148 160
149 162 151 155 177

Report 95 percent confidence intervals for $\mu$. What do the results indicate about the choice of model for these data?
11.2

CALCULATING PROJECTIONS AND RESIDUALS

We now examine some convenient procedures for calculating projections and corresponding squared lengths. This is supplemental material for the least-squares method in Section 9.2. And it is basic material for the routine calculations used with the linear model in the present chapter.

In Section 8.2A we calculated the projection of a vector \( y \) onto each axis of a new orthogonal set of axes. We saw that the projections add to reconstruct the vector \( y \) and the squared lengths add to reconstruct the squared length of \( y \).

In Section 9.2B we calculated the projection of a vector \( y \) onto a subspace generated by the \( r \) independent column vectors in a matrix \( X \) and we noted that the residual vector was the projection onto the orthogonal complement \( \mathcal{L}^\perp(X) \). We saw that the projections add to reconstruct the vector \( y \) and the squared lengths add to reconstruct the squared length of \( y \).

In this section we examine these calculations for a succession of orthogonal subspaces.

A PROJECTIONS AND SQUARED LENGTHS
BY DIFFERENCING

Formulas are given in Section 9.2B for calculating the projection into a subspace of \( \mathbb{R}^n \) and the squared length of the projection. First, consider an \( s \)-dimensional subspace \( \mathcal{L}(X_1) \) generated by the \( s \) linearly independent column vectors of the matrix

\[
X_1 = (x_1 \cdots x_s).
\]

The projection of \( y \) and the squared length of the projection are

\[
(1) \quad p_{(1)} = X_1 b_1, \quad SS_{(1)} = y' X_1 b_1,
\]

where \( b_1 = (X_1' X_1)^{-1} X_1' y \) gives the regression coefficients on the column vectors of \( X_1 \).

Now consider an \( r \)-dimensional subspace \( \mathcal{L}(X) \) where \( r > s \) and where \( \mathcal{L}(X) \) contains the preceding subspace \( \mathcal{L}(X_1) \). The column vectors of \( X \) could all be different from those of \( X_1 \); or they could include the vectors of \( X_1 \) and have \( r - s \) additional vectors.

\[
(2) \quad X = (x_1 \cdots x_s x_{s+1} \cdots x_r) = (X_1 X_2);
\]

or in particular the vectors of \( X_2 \) could each be orthogonal to the vectors of \( X_1 \). The projection of \( y \) and the squared length of the projection are

\[
(3) \quad p_{(2)} = X b, \quad SS_{(2)} = y' X b,
\]

where \( b = (X' X)^{-1} X' y \) gives the regression coefficients on the column vectors of \( X \).

Let \( \mathcal{L}(X \perp X_1) \) be the orthogonal complement of \( \mathcal{L}(X_1) \) in \( \mathcal{L}(X) \); that is, it consists of all vectors \( y \) in \( \mathcal{L}(X) \) that are orthogonal to \( X_1 \):

\[
\mathcal{L}(X \perp X_1) = \{ y : y \in \mathcal{L}(X) ; y \perp x_1, \ldots, x_s \}.
\]

In particular, if the vectors of \( X_2 \) in (2) are orthogonal to those of \( X_1 \), then \( \mathcal{L}(X \perp X_1) = \mathcal{L}(X_2) \).
We now consider the \((r - s)\)-dimensional subspace \(L(X \perp X_1)\) and show that the projection of \(y\) and the squared length of the projection are

\[
P_2 = P_{(2)} - P_{(1)}, \quad SS_2 = SS_{(2)} - SS_{(1)}
\]

and can thus be obtained by differencing. We summarize this as shown in Table 11.2.

<table>
<thead>
<tr>
<th>Subspace</th>
<th>Dimension</th>
<th>Projection</th>
<th>Squared Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>(L(X_1))</td>
<td>(s)</td>
<td>(X_1b_1)</td>
<td>(y'X_1b_1)</td>
</tr>
<tr>
<td>(L(X \perp X_1))</td>
<td>(r - s)</td>
<td>(Xb - X_1b_1)</td>
<td>(y'Xb - y'X_1b_1)</td>
</tr>
<tr>
<td>(L(X))</td>
<td>(n - r)</td>
<td>(y - Xb)</td>
<td>(y'y - y'Xb)</td>
</tr>
<tr>
<td></td>
<td>(n)</td>
<td>(y)</td>
<td>(\sum y_i^2)</td>
</tr>
</tbody>
</table>

The proof is straightforward. The projection and squared length of the projection are determined by a subspace and of course do not depend on the particular basis vectors for the subspace. For the proof, then, it is convenient to have the vectors of \(X_2\) in (2) orthogonal to the vectors of \(X_1\); that is, \(X_2'X_1 = 0\), the zero matrix. Thus for the subspace \(L(X)\) we have

\[
(X'X) = \begin{pmatrix} X'_1X'_1 & 0 \\ 0 & X'_2X'_2 \end{pmatrix}, \quad (X'X)^{-1} = \begin{pmatrix} (X'_1X'_1)^{-1} & 0 \\ 0 & (X'_2X'_2)^{-1} \end{pmatrix}.
\]

\[
b = (X'X)^{-1}X'y = \begin{pmatrix} (X'_1X'_1)^{-1} \\ 0 \end{pmatrix}X'_2y = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix},
\]

where \(b_2 = (X'_2X'_2)^{-1}X'_2y\) gives the regression coefficients on the column vectors of \(X_2\). Thus the projection and squared length for \(L(X)\) can be written

\[
P_{(2)} = X_1b_1 + X_2b_2, \quad SS_{(2)} = y'X_1b_1 + y'X_2b_2,
\]

and it follows that the projection \(X_2b_2\) and squared length \(y'X_2b_2\) for the space \(L(X \perp X_1) = L(X_2)\) are obtained by differencing. Of course, for differencing we do not need the special form for the vectors of \(X_2\).

**B INNER PRODUCTS OF RESIDUALS**

Consider \(r\) linearly independent vectors \(X = (x_1 \cdots x_r)\) and an additional set of \(s\) vectors \(Y = (y_1 \cdots y_s)\). The inner product matrix for the \(r + s\) vectors is

\[
S = X'Y = \begin{pmatrix} X'X & X'Y \\ Y'X & Y'Y \end{pmatrix}.
\]

Suppose that we operate on the first \(r\) rows of the matrix \(S\): multiplying by real numbers and adding and subtracting rows until the first \(r\) rows and \(r\) columns become the identity matrix. We are then effectively performing the following premultiplication:

\[
\begin{pmatrix} (X'X)^{-1} & 0 \\ 0 & I \end{pmatrix}S = \begin{pmatrix} I & b(y_1) \cdots b(y_s) \\ Y'X & Y'Y \end{pmatrix} = \begin{pmatrix} I & B(Y) \\ Y'X & Y'Y \end{pmatrix},
\]

where \(b(y_i) = (X'X)^{-1}X'y_i\) is the vector of regression coefficients of \(y_i\) on \(x_1, \ldots, x_r\).
thus the upper right matrix
\[ B(Y) = (X'X)^{-1}X'Y = (b(y_1) \ldots b(y_s)) \]
contains the various \( b \) vectors obtained from \( y_1, \ldots, y_s \) successively.

Now suppose that we take multiples of the first \( r \) rows and subtract them from the last \( s \) rows to produce zeros in the first \( r \) columns of those rows. We are then effectively performing the following matrix premultiplication:

\[
\begin{pmatrix}
  I \\
-\gamma'X \\
  0
\end{pmatrix}
\begin{pmatrix}
  (X'X)^{-1} \\
  0 \\
  I
\end{pmatrix}
= \begin{pmatrix}
  I \\
  \gamma'Y - \gamma'XB(Y)
\end{pmatrix}.
\]

We now show the interesting result that the bottom right \( s \times s \) matrix is the inner product matrix of the residual vectors \( y_i - Xb(y_i) \). We first assemble the residual vectors in an \( n \times s \) matrix
\[
(y_1 - Xb(y_1), \ldots, y_s - Xb(y_s)) = Y - XB(Y);
\]
we then calculate the inner product matrix

\[
(Y - XB(Y))'(Y - XB(Y))
= Y'Y - Y'XB(Y) - B'(Y)X'Y + B'(Y)X'XB(Y)
= Y'Y - Y'XB(Y).
\]

Note that the last three terms are simplified by substituting the expression for \( B(Y) \).

Thus the two left multiplications on the matrix \( S \) give the coefficients for the projection vectors and give the inner product matrix for the residual vectors.

### C  EXAMPLE

Consider a biological system with an input variable \( x \) and a response variable \( y \). The following five observations were obtained:

<table>
<thead>
<tr>
<th>( x )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>5</td>
<td>7</td>
<td>11</td>
<td>10</td>
<td>13</td>
</tr>
</tbody>
</table>

We first examine the trivial least-squares fit of the constant

\[(8) \quad y = b \cdot 1.\]

We then examine the least-squares fit of an affine function of \( x \):

\[(9) \quad y = b_1 1 + b_2 x.\]

We then examine the least-squares fit of a quadratic function of \( x \):

\[(10) \quad y = b_1 1 + b_2 x + b_3 x^2.\]

For the calculations we use the methods and results discussed earlier in Sections A and B.

Let \( x_1, x_2, \) and \( x_3 \) be the vectors recording the values of \( 1, x, \) and \( x^2 \), respectively. Some preliminary calculations are given in Table 11.3. The inner product matrix of \( x_1, \)
TABLE 11.3
Calculations for some variables

<table>
<thead>
<tr>
<th></th>
<th>x</th>
<th>x²</th>
<th>x³</th>
<th>x⁴</th>
<th>y</th>
<th>xy</th>
<th>x²y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>7</td>
<td>14</td>
<td>28</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>9</td>
<td>27</td>
<td>81</td>
<td>11</td>
<td>33</td>
<td>99</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>16</td>
<td>64</td>
<td>256</td>
<td>10</td>
<td>40</td>
<td>160</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>25</td>
<td>125</td>
<td>625</td>
<td>13</td>
<td>65</td>
<td>325</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>55</td>
<td>225</td>
<td>979</td>
<td>46</td>
<td>157</td>
<td>617</td>
</tr>
</tbody>
</table>

x₂, and x₃ with x₁, x₂, x₃, and y is recorded in the following array. An identity matrix is adjoined on the right; it will absorb and thus keep a record of any left matrix multiplication. The sum of squares of the y's is y'y = Σ y² = 464.

<table>
<thead>
<tr>
<th>x₁</th>
<th>x₂</th>
<th>x₃</th>
<th>y</th>
<th>x₁</th>
<th>x₂</th>
<th>x₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>15</td>
<td>55</td>
<td>46</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>55</td>
<td>225</td>
<td>157</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>55</td>
<td>225</td>
<td>979</td>
<td>617</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

D FITTING A CONSTANT

First consider the least-squares fit of the constant as represented by equation (8). This corresponds to finding the projection in the subspace generated by the 1-vector x₁ (= w₁ for later reference) or, equivalently, generated by the corresponding unit vector

v₁ = x₁/√5 = (1, 1, 1, 1)/√5;

the squared length 5 is the one–one element in the first array (11). In the pattern of Section B we perform row operations (left multiplication by a matrix) on (11) to reduce the first row–first column to 1 and then reduce the remainder of the first column to 0's:

<table>
<thead>
<tr>
<th>x₁</th>
<th>x₂</th>
<th>x₃</th>
<th>y</th>
<th>x₁</th>
<th>x₂</th>
<th>x₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>11</td>
<td>9.2</td>
<td>0.2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>60</td>
<td>19</td>
<td>-3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>60</td>
<td>374</td>
<td>111</td>
<td>-11</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

The second, third, and fourth columns contain regression coefficients on x₁ above the dashed line and contain inner products of residual vectors below the dashed line. The right-hand array contains the matrix that has been left-multiplied into (11).

Consider the projection of y on the vector x₁ and the squared length of the projection. From formulas (1) and (5) we have

p₁ = 9.2x₁, \[SS₁ = 46 \times 9.2 = 423.2.\]

We can summarize this in Table 11.4.
TABLE 11.4
Projections and corresponding squared lengths

<table>
<thead>
<tr>
<th>Space</th>
<th>Basis</th>
<th>Dimension</th>
<th>Projection</th>
<th>Squared Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L(x_1)$</td>
<td>$v_1$</td>
<td>1</td>
<td>$9.2x_1$</td>
<td>423.2</td>
</tr>
<tr>
<td>$L^2(x_1)$</td>
<td></td>
<td>4</td>
<td>$y - 9.2x_1$</td>
<td>40.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>$y$</td>
<td>464.0</td>
</tr>
</tbody>
</table>

E FITTING AN AFFINE FUNCTION

Now consider the least-squares fit of the affine function represented by equation (9). This corresponds to finding the projection in the subspace generated by the vectors $x_1$ and $x_2$. We examine this in part as an extension from the preceding simpler case; as an extension we are concerned with $L(x_2 \perp x_1)$.

The array (12) gives the coefficients for the projection of $x_2$ on $x_1$; we thus obtain the residual

$$w_2 = x_2 - 3x_1 = (-2, -1, 0, 1, 2)' .$$

The array (12) also gives the squared length of this residual: the squared length is 10. The corresponding unit vector is

$$v_2 = (x_2 - 3x_1)/\sqrt{10} = (-2, -1, 0, 1, 2)'/\sqrt{10} .$$

Note that $w_2$ or $v_2$ generates the orthogonal complement $L(x_2 \perp x_1)$. Also note that the vector $w_2$ records the values of the function $x - 3$, which is orthogonal to the function 1 at the points \{1, 2, 3, 4, 5\}. Now in the pattern of Section B we perform row operations (left multiplication by a matrix) on (11) to reduce the first two rows—first two columns to the identity matrix and the remainder of the first two columns to 0's. Note that this is equivalent to reducing the second row—second column element of (12) to 1 and the remainder of the second column to 0's:

$$\begin{array}{cccc|ccc}
 x_1 & x_2 & x_3 & y & x_1 & x_2 & x_3 \\
\hline
 1 & 0 & -7 & 3.5 & 1.1 & -0.3 & 0 \\
 0 & 1 & 6 & 1.9 & 0.3 & 0.1 & 0 \\
 0 & 0 & 14 & -3 & 7 & -6 & 1 \\
\end{array} \tag{13}$$

The third and fourth columns contain regression coefficients on $x_1$ and $x_2$ above the dashed line and contain inner products of the residual vectors below the dashed line. The right-hand array contains the matrix that has been left-multiplied into (11).

Consider the projection of $y$ on the vectors $x_1$ and $x_2$ and the squared length of the projection. From formulas (3) and (5) we have

$$p_{(2)} = 3.5x_1 + 1.9x_2, \quad SS_{(2)} = 46 \times 3.5 + 157 \times 1.9 = 459.3 .$$

The projection of $y$ on the orthogonal complement $L(x_2 \perp x_1)$ is the projection of $y$ on $w_2$ or on $v_2$; it is available using the step from (12) to (13),

$$p_2 = 1.9w_2, \quad SS_2 = 19 \times 1.9 = 36.1 .$$

We can then summarize in Table 11.5.
TABLE 11.5
Projections and corresponding squared lengths

<table>
<thead>
<tr>
<th>Space</th>
<th>Basis</th>
<th>Dimension</th>
<th>Projection</th>
<th>Squared Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathcal{L}(x_1) )</td>
<td>( v_1 )</td>
<td>1</td>
<td>9.2w_1</td>
<td>423.2</td>
</tr>
<tr>
<td>( \mathcal{L}(x_2 \perp x_1) )</td>
<td>( v_2 )</td>
<td>1</td>
<td>1.9w_2</td>
<td>36.1</td>
</tr>
<tr>
<td>( \mathcal{L}^\perp(x_1, x_2) )</td>
<td>y - 3.5x_1 - 1.9x_2</td>
<td>3</td>
<td>4.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>y</td>
<td>464.0</td>
</tr>
</tbody>
</table>

\( \text{Fitting a quadratic function of } x \)

Now consider the least-squares fit of the quadratic function represented by equation (10). This corresponds to finding the projection in the subspace generated by \( x_1, x_2, \) and \( x_3 \). We examine this in part as an extension from the preceding case; as an extension we are concerned with \( \mathcal{L}(x_3 \perp x_1, x_2) \).

The array (13) gives the coefficients for the projection of \( x_3 \) on \( x_1 \) and \( x_2 \); we thus obtain the residual

\[
\mathbf{w}_3 = x_3 - (-7)x_1 - 6x_2 = (2, -1, -2, -1, 2)^t.
\]

The array (13) also gives the squared length of this residual: the squared length is 14. The corresponding vector is

\[
\mathbf{v}_3 = (x_3 + 7x_1 - 6x_2)/\sqrt{14} = (2, -1, -2, -1, 2)^t/\sqrt{14}.
\]

Note that \( \mathbf{w}_3 \) or \( \mathbf{v}_3 \) generates the orthogonal complement \( \mathcal{L}(x_3 \perp x_1, x_2) \). Also note that the vector \( \mathbf{w}_3 \) records the values of the function \( x^2 + 7 - 6x \), which is orthogonal to the functions 1 and \( x \) at the points \( \{1, 2, 3, 4, 5\} \).

Now in the pattern of Section B we perform row operations on (11) to reduce the first three rows—first three columns to the identity matrix and the remainder (here none) of the first three columns to 0's. Note that this is equivalent to reducing the third row—third column element of (13) to 1 and the remainder of the third column to 0's:

\[
\begin{align*}
\begin{array}{cccc}
  x_1 & x_2 & x_3 & y \\
  1 & 0 & 0 & 2 \\
  0 & 1 & 0 & 31/10 \\
  0 & 0 & 1 & -3/14 \\
\end{array}
& \quad \begin{array}{cccc}
  x_1 & x_2 & x_3 \\
  4.6 & -3.3 & 1/2 \\
  -3.3 & 23/10 & -3/7 \\
  1/2 & -3/7 & 1/14 \\
\end{array}
\end{align*}
\]

The fourth column contains regression coefficients of \( y \) on \( x_1, x_2, \) and \( x_3 \). The right-hand array contains the matrix that has been left-multiplied into (11).

Consider the projection of \( y \) on the vectors \( x_1, x_2, \) and \( x_3 \) and the squared length of the projection. From formulas (3) and (5) we have

\[
\mathbf{p}_{(3)} = 2x_1 + 3\frac{3}{10}x_2 - \frac{3}{14}x_3.
\]

\[
\text{SS}_{(3)} = 46 \times 2 + 157 \times 3\frac{3}{10} + 617 \times (-\frac{3}{14})
= 459.96.
\]
The projection of $\mathbf{y}$ on the orthogonal complement $\mathcal{L}(x_3 \perp x_1, x_2)$ is the projection of $\mathbf{y}$ on $\mathbf{w}_3$ or on $\mathbf{v}_3$: it is available using the step from (13) to (14):

$$p_3 = -(3/14)w_3, \quad SS_3 = -3(-3/14) = 9/14 = 0.64.$$ 

We can then summarize in Table 11.6.

<table>
<thead>
<tr>
<th>Space</th>
<th>Basis</th>
<th>Dimension</th>
<th>Projection</th>
<th>Squared Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{L}(x_1)$</td>
<td>$v_1$</td>
<td>1</td>
<td>$9.2w_1$</td>
<td>423.2</td>
</tr>
<tr>
<td>$\mathcal{L}(x_2 \perp x_1)$</td>
<td>$v_2$</td>
<td>1</td>
<td>$1.9w_2$</td>
<td>36.1</td>
</tr>
<tr>
<td>$\mathcal{L}(x_3 \perp x_1, x_2)$</td>
<td>$v_3$</td>
<td>1</td>
<td>$-\frac{9}{14}w_3$</td>
<td>0.64</td>
</tr>
<tr>
<td>$\mathcal{L}^1(x_1, x_2, x_3)$</td>
<td>$\frac{2}{\sqrt{5}}x_1 - \frac{3}{\sqrt{14}}x_2 + \frac{1}{\sqrt{14}}x_3$</td>
<td>2</td>
<td>$y - 2x_1 - 3\frac{3}{14}x_2 + \frac{1}{\sqrt{14}}x_3$</td>
<td>4.06</td>
</tr>
<tr>
<td></td>
<td>$\frac{5}{\sqrt{14}}x_2$</td>
<td></td>
<td>$y$</td>
<td>464.00</td>
</tr>
</tbody>
</table>

**G SUPPLEMENT: AN ORTHONORMAL BASIS**

We have been considering vectors $x_1$, $x_2$, and $x_3$ which generated successively the subspaces $\mathcal{L}(x_1)$, $\mathcal{L}(x_1, x_2)$, and $\mathcal{L}(x_1, x_2, x_3)$. As part of this we derived orthogonal vectors $w_1$, $w_2$, and $w_3$ which generated successively the same three subspaces. From the second column of (12) and the third column of (13), we have obtained the coefficients for expressing the $w$'s in terms of the $x$'s:

$$W = (w_1, w_2, w_3) = (x_1, x_2, x_3) \begin{pmatrix} 1 & -3 & 7 \\ 0 & 1 & -6 \\ 0 & 0 & 1 \end{pmatrix} = X \begin{pmatrix} 1 & -3 & 7 \\ 0 & 1 & -6 \\ 0 & 0 & 1 \end{pmatrix}.$$ 

The corresponding orthonormal vectors $v_1$, $v_2$, and $v_3$ were obtained from the $w_1$, $w_2$, and $w_3$ using squared lengths from (11), (12), and (13):

$$V = (w_1, w_2, w_3) \begin{pmatrix} 1/\sqrt{5} & 0 & 0 \\ 0 & 1/\sqrt{10} & 0 \\ 0 & 0 & 1/\sqrt{14} \end{pmatrix} = X \begin{pmatrix} 1/\sqrt{5} & -3/\sqrt{10} & 7/\sqrt{14} \\ 0 & 1/\sqrt{10} & -6/\sqrt{14} \\ 0 & 0 & 1/\sqrt{14} \end{pmatrix}.$$ 

We can also determine from the arrays (12) and (13) the coefficients for the projection of the $x$'s on the orthogonal $w$'s. From the first row of (12) and the second row of (13) we obtain the coefficients for expressing the $x$'s in terms of the $w$'s:

$$X = (x_1, x_2, x_3) = (w_1, w_2, w_3) \begin{pmatrix} 1 & 3 & 11 \\ 0 & 1 & 6 \\ 0 & 0 & 1 \end{pmatrix} = W \begin{pmatrix} 1 & 3 & 11 \\ 0 & 1 & 6 \\ 0 & 0 & 1 \end{pmatrix}.$$ 

The vectors $x_1$, $x_2$, $x_3$ can also be expressed directly in terms of the orthonormal vectors $v_1$, $v_2$, and $v_3$: 

$$X = (x_1, x_2, x_3) = (w_1, w_2, w_3) \begin{pmatrix} 1 & 3 & 11 \\ 0 & 1 & 6 \\ 0 & 0 & 1 \end{pmatrix}.$$
(18) \[ X = W \begin{pmatrix} 1 & 3 & 11 \\ 0 & 1 & 6 \\ 0 & 0 & 1 \end{pmatrix} = V \begin{pmatrix} \sqrt{5} & 0 & 0 \\ 0 & \sqrt{10} & 0 \\ 0 & 0 & \sqrt{14} \end{pmatrix} \begin{pmatrix} 1 & 3 & 11 \\ 0 & 1 & 6 \\ \sqrt{5} & 3\sqrt{5} & 11\sqrt{5} \\ 0 & \sqrt{10} & 6\sqrt{10} \\ 0 & 0 & \sqrt{14} \end{pmatrix} = VT, \]

where \( T \) is positive upper triangular (upper triangular with positive values on the diagonal).

The right-hand triangular matrices in (16) and (18) are inverses of each other.

**H  EXERCISES**

A large number of significant figures can be needed to ensure accuracy for the regression coefficients and inner products. In general, fractions as illustrated in this section should be avoided; hand, desk, and large-scale computers make the necessary arithmetic accessible.

1. For an input variable \( x \) and response \( y \) the following data were obtained:

| \( x \) | 1.6 | 1.8 | 1.9 | 1.9 | 2.1 | 2.2 |
| \( y \) | 0.5 | 0.4 | 0.7 | 0.9 | 1.2 | 1.1 |

First fit the constant \( y = b \cdot 1 \) and then fit the affine function \( y = b_1 \cdot 1 + b_2 x \). Record the projections and summary tables as in Sections D and E.

2. For an input variable \( x \) and response \( y \) the following data were obtained:

| \( x \) | 2.6 | 2.7 | 2.8 | 2.9 | 3.1 |
| \( y \) | 12.1 | 12.5 | 12.7 | 13.0 | 13.5 |

First fit the constant, then the affine function, then the quadratic function. Record the projections and summary tables as in Sections D, E, and F. Carry sufficient figures in calculations.

3. (continuation) Let \( X = (x_1, x_2, x_3) \), where \( x_1, x_2, \) and \( x_3 \) record the values of 1, \( x \), and \( x^2 \), respectively; and let \( W = (w_1, w_2, w_3) \) be the corresponding orthogonal vectors, and \( V = (v_1, v_2, v_3) \) be the corresponding orthonormal vectors. Determine the matrices \( T \) and \( V \) as in the decomposition (18).

4. The response SCI \( y \) was observed corresponding to input variables saturation \( x_1 \) and transisomers \( x_2 \):

| \( x_1 \) | 1.0 | 1.1 | 1.2 | 1.3 | 1.0 | 1.1 | 1.2 | 1.3 | 1.0 | 1.1 | 1.2 | 1.3 |
| \( x_2 \) | 0.04 | 0.04 | 0.04 | 0.04 | 0.20 | 0.20 | 0.20 | 0.20 | 0.38 | 0.38 | 0.38 | 0.38 |
| \( y \) | 67.1 | 64.0 | 44.3 | 45.1 | 69.8 | 58.5 | 46.3 | 44.1 | 74.5 | 60.7 | 49.1 | 47.6 |

Fit the constant \( y = b1 \); fit an affine function of the first input \( y = b_1 \cdot 1 + b_2 x_1 \); fit an affine function of the two inputs \( y = b_1 \cdot 1 + b_2 x_1 + b_3 x_2 \). Record the projections and summary tables as in Sections D, E, and F.
11.3

THE REGRESSION MODEL

We now examine the linear or regression model, in which the general level of the response vector is linear in terms of location parameters. For a variety of examples see Section 9.2. These examples indicate a broad range of applications in which input variables to a system cause a change in the general level of the responses. The regression model generalizes the location-scale model of Section 11.1.

A THE REGRESSION MODEL

Consider a real-valued response \( y \). Suppose that earlier experience with the system under investigation has identified the distribution of the variation affecting the response. Let \( z \) designate a variable presenting this variation, and \( f \) be the known density function for \( z \). And suppose that under present conditions the response scaling of the variation is unknown and the general level of the response varies in certain unknown ways relative to input variables for the system. Let \( \sigma \) designate the response scaling of the variation.

Consider the standardizations for \( f \) as presented in Section 11.1. If \( f \) can be standardized so that \( E(z) = 0, \) \( \text{var}(z) = 1 \), then the general level \( \mu \) of the response is presented as the mean response and the scaling \( \sigma \) is the response standard deviation. Or if \( f \) is standardized so that the median of \( z \) is zero and the interval \((-1, +1)\) contains 68.26 percent of the probability, then the general level \( \mu \) of the response is presented as the median response and the interval \((\mu - \sigma, \mu + \sigma)\) contains 68.26 percent of the response values; in this case \( \sigma \) is the response standard error.

Now consider a sequence of repetitions of the system under various settings for input variables. Let \( z = (z_1, \ldots, z_n)' \) designate the sequence for the variation and \( y = (y_1, \ldots, y_n)' \) designate the sequence for the response. And suppose that the response has a linear location model

\[
(1) \quad y = X \beta + \sigma z
\]

as discussed in Section 9.2A, where \( X = (x_1, \ldots, x_r) \) records \( r \) linearly independent vectors based on the input variables and \( \beta \) is the location parameter taking values in \( \mathbb{R}^r \).

Then for the variation \( z \) we have the probability model as given by the density function

\[
(2) \quad f(z_1) \cdots f(z_n)
\]

on \( \mathbb{R}^n \). And for the response we have the statistical model

\[
(3) \quad \{y = X \beta + \sigma z : (\beta, \sigma) \in \mathbb{R}^r \times \mathbb{R}^+\},
\]

giving the set of possible functions for the response \( y \) in terms of the variation \( z \).

B DATA FOR \( n \) INDEPENDENT REPETITIONS

Now consider an observed response sequence \( y = (y_1, \ldots, y_n)' \) from \( n \) independent repetitions of the system under input conditions represented by \( X \). The
observed response sequence is a function

\[ y = X\beta + \sigma z \]

of a realized sequence \( z \) from the distribution (2).

Consider the information concerning the \( z \) realized from the distribution (2). With no information concerning the values of \( \beta \) and \( \sigma \) it follows that

\[ z = -\sigma^{-1}\beta_1 x_1 + \cdots + \sigma^{-1}\beta_r x_r + \sigma^{-1}\gamma \]

can be any point in the half-\((r+1)\) space

\[ \mathcal{L}^+(X; y) = \{ a_1 x_1 + \cdots + a_r x_r + cy : a_j \in R, c \in R^+ \} \]

based on the subspace \( \mathcal{L}(X) \) and passing through the observed response \( y \); see Figure 11.6. Thus all that we effectively observe concerning the realized \( z \) is the half-space passing through it; we have no differential information concerning the position of \( z \) in that half-space.

In Section C we obtain convenient coordinates for the half-space and convenient coordinates within the half-space. Then in Section D we derive the conditional distribution describing the realized \( z \) in the observed half-space.

**C. SIMPLE COORDINATES**

We now use some least-squares results to obtain simple coordinates for a half-space \( \mathcal{L}^+(X; z) \) and simple coordinates for a point \( z \) in such a half-space; see Figure 11.7. The half-space \( \mathcal{L}^+(X; z) \) is hinged on the \( r \)-dimensional subspace \( \mathcal{L}(X) \). Accordingly, it is reasonable to use the \( r \) vectors in \( X \) as basis vectors. The vector \( z \)

![Figure 11.6](image)

The \( r \)-dimensional base space \( \mathcal{L}(X) \); an additional dimension by introducing the vector \( y \) gives the \((r+1)\)-dimensional space \( \mathcal{L}(X, y) \). \( \mathcal{L}^+(X; y) \) is a positive half of \( \mathcal{L}(X, y) \) obtained with a positive coordinate for \( y \).
has projection

\[(6) \quad X b(z) = b_1(z)x_1 + \cdots + b_r(z)x_r \]

on the space \(L(X)\); the regression coefficients forming \(b(z)\) are given by formula (9.2.6). And the vector \(z\) has projection \(z - Xb(z)\) on the orthogonal complement \(L(X)\). Thus

\[(7) \quad z = Xb(z) + (z - Xb(z)) = Xb(z) + s(z)d(z), \]

where \(s(z)\) is the length of the residual vector and \(d(z)\) is the unit residual vector. This suggests using the vectors \(X, d(z)\), that is, \(x_1, \ldots, x_r, d(z)\), as the \(r + 1\) basis vectors for \(L(X, z)\); thus

\[(8) \quad z = Xb(z) + s(z)d(z) = b_1(z)x_1 + \cdots + b_r(z)x_r + s(z)d(z) \]

with coordinates \(b_1(z), \ldots, b_r(z), s(z)\) relative to the vectors \(x_1, \ldots, x_r, d(z)\). Note that this is not an orthogonal basis but that the vector \(d(z)\) is orthogonal to the vectors of \(X\).

The points \(d(z)\) are at unit distance from the origin and lie in the \((n - r)\)-dimensional subspace \(L(X)\). Thus the points \(d(z)\) form the unit sphere \(Q_{n-r}\) in \(L(X)\).

Now consider the possible points \(z\) given the information provided by the observed response \(y\). From Section B we see that \(z\) and \(y\) must lie on the same half-space \(L^+(X; z) = L^+(X; y)\) hinged on \(L(X)\). Thus \(d(z) = d(y)\), which says they have the same \((r + 1)\)th basis vector. For convenience we abbreviate \(d(y)\) as \(d\). We then see
that the information provided by the observed response \( y \) is that \( d(z) = d \); we have the "observed" value \( d \) for the function \( d(z) \).

Now consider the relation (4) between the observed response \( y \) and the realized \( z \):

\[
y = Xb(y) + s(y)d = X\beta + \sigma z = X\beta + \sigma(Xb(z) + s(z)d) \\
= X(\beta + \sigma b(z)) + \sigma s(z)d.
\]

By equating coefficients relative to the basis vectors \( x_1, \ldots, x_r, d \), we obtain

\[
\begin{align*}
b(y) &= \beta + \sigma b, \\
s(y) &= \sigma s,
\end{align*}
\]

where we abbreviate \((b(z), s(z)) = (b, s)\). These can be compared with equations (11.1.8); we now have \( r \) location coordinates instead of 1.

**D THE CONDITIONAL DISTRIBUTION GIVEN THE DATA**

We now determine the marginal probability for what has been observed, the half-space through \( z \) as given by \( d(z) = d \), and we determine the conditional probability for what cannot be observed, the position \((b, s)\) of \( z \) in the observed half-space. As in Section 11.1, we can be quite general and allow a parameter \( \lambda \) for the distribution \( f = f_\lambda \) for the variation.

The probability differential for the distribution on \( \mathbb{R}^n \) is given by

\[
\prod_{i=1}^{n} f_{\lambda}(z_i) \prod_{i=1}^{n} dz_i.
\]

To obtain the marginal and conditional distributions we make the change of variable (8) to the new coordinates \( b(z), d \). For the density function the substitution is trivial. For the differential we must argue in several steps. At the point \( z \) we first measure area or volume parallel to the subspace \( \mathcal{L}(X) \). In the subspace we have coordinate \( b_1(z), \ldots, b_r(z) \) relative to the vectors \( X \); suppose that we have alternative coordinates \( a_1(z), \ldots, a_r(z) \) relative to an orthonormal set of vectors \( V \) that also generate the subspace \( \mathcal{L}(X) \). Let \( X = VT \) express the \( X \) vectors as linear combinations of the \( V \) vectors; see (11.2.18) and note that \( T \) is \( r \times r \). Then

\[
X = VT, \quad Xb = V Tb = Va,
\]

giving the relation \( Tb = a \) between the original coordinates \( b \) and the new coordinates \( a \). Euclidean area or volume is given simply in terms of the \( a \) coordinates:

\[
da_1 \cdots da_r = da = |T| d'b = |V'V|^{1/2}|T| d'b \\
= |T'V'VT|^{1/2} d'b = |X'X|^{1/2} d'b,
\]

where of course \(|V'V|\) is the determinant of an identity matrix. Thus \(|X'X|^{1/2} d'b \) measures area or volume parallel to the subspace \( \mathcal{L}(X) \). At \( z \) we can then measure distance in the orthogonal direction \( d \) by \( ds \). Then in directions orthogonal to the half-space we can measure area or volume on the sphere formed by the points \( sd(z) \).
with \( s \) fixed. Let \( d\alpha \) be area or volume on the unit sphere \( O_{n-r} \) formed by the points \( d(z) \); then \( s^{n-r-1} d\alpha \) is area or volume on the sphere of radius \( s \). We then put together these three length or volume measures in the three orthogonal spaces and obtain

\[
\prod dz_i = |X'X|^{1/2} \, d\mathbf{b} \, ds \, s^{n-r-1} \, d\alpha.
\]

Thus the probability differential for the distribution on \( R^n \) can be reexpressed as

\[
\prod f_{\lambda} \left( \sum b_u x_{ui} + sd_i \right) |X'X|^{1/2} \, d\mathbf{b} \, ds \, s^{n-r-1} \, d\alpha.
\]

By integrating out the variables \( \mathbf{b} \) and \( s \) over \( R^r \times R^+ \), we obtain the marginal density for \( \mathbf{d} \):

\[
(14) \quad k_{\lambda}(\mathbf{d}) = \int_{-\infty}^{\infty} \int_{0}^{\infty} \prod f_{\lambda} \left( \sum b_u x_{ui} + sd_i \right) s^{n-r-1} \, ds \, |X'X|^{1/2} \, d\mathbf{b}.
\]

This then gives the following factorization of the probability differential (11):

\[
(15) \quad k_{\lambda}(\mathbf{d}) \, d\mathbf{d} \cdot k_{\lambda}(\mathbf{d}) \prod f_{\lambda} \left( \sum b_u x_{ui} + sd_i \right) s^{n-r-1} \, ds \, |X'X|^{1/2} \, d\mathbf{b};
\]

the probability differential is factored into the marginal probability for the observed \( \mathbf{d} \) and the conditional probability for the unobserved \( (\mathbf{b}, s) \). Note that the normalizing constant \( k_{\lambda}(\mathbf{d}) \) is obtained by integrating over the half-space \( R^r \times R^+ \); see formula (14).

**E INFERENCE CONCERNING THE PARAMETERS**

For statistical inference we follow the pattern in Section 11.1E. We have an observed \( \mathbf{d} \) from a distribution involving \( \lambda \). We can examine the likelihood function \( L(\mathbf{d} \mid \cdot) \) for \( \lambda \), where

\[
(16) \quad L(\mathbf{d} \mid \lambda) = cK_{\lambda}(\mathbf{d})
\]

and perhaps obtain sharp discriminations concerning possible \( \lambda \) values. Perhaps the model \( k_{\lambda}(\cdot) \, d\alpha \) may be accessible and of use.

Then for a given value for \( \lambda \) we can use the distribution

\[
(17) \quad k_{\lambda}^{-1}(\mathbf{d}) \prod f_{\lambda} \left( \sum b_u x_{ui} + sd_i \right) s^{n-r-1} \, ds \, |X'X| \, d\mathbf{b}
\]

together with the equations (10) to make tests concerning \( \beta_1, \ldots, \beta_r, \sigma \) or to calculate confidence intervals. In the remainder of this chapter we examine inference for the case of normal variation.

**F DISTRIBUTIONS IN THE NORMAL CASE**

Consider the case of a standard normal distribution for the variation \( z \). Let \( y \) be a realized response and let \( \mathbf{d} = d(y) \) be the identified characteristic \( d(z) \) of the variation \( z \). Then from (17) we have the following distribution describing \( b_1, \ldots, b_r, s \):

\[
k^{-1}(\mathbf{d}) (2\pi)^{-n/2} \exp \left\{ -\frac{1}{2} (Xb + sd)'(Xb + sd) \right\} s^{n-r-1} \, ds \, |X'X|^{1/2} \, d\mathbf{b}
\]

\[
= \frac{|X'X|^{1/2}}{(2\pi)^{r/2}} \exp \left\{ -\frac{1}{2} b'(X'X)b \right\} \, d\mathbf{b} \cdot \frac{1}{\Gamma \left( \frac{n-r}{2} \right)} \exp \left\{ -\frac{s^2}{2} \left( \frac{s^2}{2} \right)^{(n-r)/2-1} \right\} \, d \frac{s^2}{2},
\]
where we have used the orthogonality $X'd = 0$ of the vectors $x_1, \ldots, x_r$ with $d$, and for integration we have used just the norming constants for the normal and gamma densities. Thus we obtain the following:

**LEMMA 1**

*Under the preceding assumption, $b$ is multivariate normal($0; (X'X)^{-1}$); $s^2$ is chi-square($n - r$); $b$ is statistically independent of $s$.***

Note that the distribution does not depend on $d$ and is thus equal to the marginal distribution.

The response distribution consistent with the observed value $d$ can be obtained immediately from equation (10):

**LEMMA 2**

*Under the preceding assumption: $b(y)$ is multivariate normal($\beta; \sigma^2(X'X)^{-1}$); $s^2(y)$ is $\sigma^2$ times a chi-square($n - r$) variable; $b(y)$ is statistically independent of $s(y)$.***

It is of interest to record the analogous results for the use of coordinates $a_1, \ldots, a_r$ with respect to an orthonormal basis $v_1, \ldots, v_r$. From Section 11.2G we have $X = VT$; thus

\[(18) \quad Xb = VTb = Va,\]

where $a = Tb$ and the particular matrix $T$ is upper triangular. The inner product matrix $V'V$ is the $r \times r$ identity matrix.

**LEMMA 3**

*Under the preceding assumptions: $a_1, \ldots, a_r, s$ are statistically independent; $a_i$ is standard normal; $s^2$ is chi-square($n - r$).***

For the response distribution relative to the orthonormal basis we have, from equation (10),

\[a(y) = \alpha + \sigma a,\]

\[s(y) = \sigma s,\]

where the new parameters $\alpha_1, \ldots, \alpha_r$ are given by $\alpha = T\beta$, as is seen from

\[(19) \quad X = VT, \quad X\beta = VT\beta = Va.\]

**LEMMA 4**

*Under the preceding assumptions: $a_1(y), \ldots, a_r(y), s(y)$ are statistically independent; $a_i(y)$ is normal ($\alpha_i, \sigma$); $s^2(y)$ is $\sigma^2$ times a chi-square($n - r$) variable.*
TABLE 11.7
Projections and corresponding squared lengths

<table>
<thead>
<tr>
<th>Source</th>
<th>Dimension</th>
<th>Projection</th>
<th>Squared Length</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_1$</td>
<td>1</td>
<td>$a_1(y)v_1$</td>
<td>$a_1^2(y)$</td>
<td>$\sigma^2 \chi_1^2$ if $\alpha_1 = 0$</td>
</tr>
<tr>
<td>$v_2$</td>
<td>1</td>
<td>$a_2(y)v_2$</td>
<td>$a_2^2(y)$</td>
<td>$\sigma^2 \chi_1^2$ if $\alpha_2 = 0$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$v_r$</td>
<td>1</td>
<td>$a_r(y)v_r$</td>
<td>$a_r^2(y)$</td>
<td>$\sigma^2 \chi_1^2$ if $\alpha_r = 0$</td>
</tr>
<tr>
<td>$\Omega^+(X)$</td>
<td>$\frac{n-r}{n}$</td>
<td>$\frac{s(y)d}{y}$</td>
<td>$\frac{s^2(y)}{\sum y_i^2}$</td>
<td>$\sigma^2 \chi_{n-r}^2$</td>
</tr>
</tbody>
</table>

These results are recorded partially in Table 11.7 following the pattern discussed in Section 11.2. The different projections are statistically independent, as are the different squared lengths. The residual squared length $s^2(y)$ has the distribution of $\sigma^2$ times a chi-square $(n - r)$ variable; if divided by its degrees of freedom, it provides an estimate of $\sigma^2$. The squared length for a component based on $v_u$ has the distribution of $\sigma^2$ times a chi-square (1) distribution provided that the corresponding location parameter $\alpha_u = 0$; otherwise, it has a noncentral chi-square-type distribution (see Problem 3.6.3) and tends to be larger in value.

In Section A we noted that the vectors $x_1$, $\ldots$, $x_r$ are often obtained as values of $r$ input variables. In such cases it can be of interest to estimate the response level at settings, say $l_1$, $\ldots$, $l_r$, for the input variables. This response level is given by

$$(20) \quad y = \beta_1 l_1 + \cdots + \beta_r l_r = \beta' l,$$

and the corresponding estimate is

$$(21) \quad \hat{y} = b_1(y)l_1 + \cdots + b_r(y)l_r = b'(y)l.$$

The distribution for this estimate is then obtained from, say, Example 5.7.5:

**LEMMA 5**

*Under the model (4) with standard normal variation, the distribution of the estimate $\hat{y}$ is normal with mean $\beta' l$ and variance $l'(X'X)^{-1}l \sigma^2$.***

**11.4**

**INFERENCES WITH THE REGRESSION MODEL**

We now examine statistical inference for the regression model of the preceding section. We do this in terms of the numerical example introduced in Section 11.2C.

The numerical example involved a random system with input variable $x$ and response $y$. The data are as follows:

<table>
<thead>
<tr>
<th>$x$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>5</td>
<td>7</td>
<td>11</td>
<td>10</td>
<td>13</td>
</tr>
</tbody>
</table>

These are plotted in Figure 11.8.
A THE QUADRATIC MODEL

We analyze the example on the assumption that the quadratic model

\[(1) \quad y = \beta_1 + \beta_2 x + \beta_3 x^2 + \sigma z\]

adequately describes the response and that the underlying variation \(z\) is standard normal.

The analysis-of-variance table for fitting constant, linear, and then quadratic dependence on \(x\) was obtained in Section 11.2F and is presented here as Table 11.8. Dimension has been relabeled DF, for degree of freedom, and squared length has been relabeled SS, for sums of squares.

<table>
<thead>
<tr>
<th>Source</th>
<th>Basis</th>
<th>DF</th>
<th>Projection</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>(v_1)</td>
<td>1</td>
<td>(9.2x_1 = 9.2w_1)</td>
<td>423.20</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(= (9.2, 9.2, 9.2, 9.2, 9.2))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear</td>
<td>(v_2)</td>
<td>1</td>
<td>(-5.7x_1 + 1.9x_2 = 1.9w_2)</td>
<td>36.10</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(= (-3.8, -1.9, 0, 1.9, 3.8))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quadratic</td>
<td>(v_3)</td>
<td>1</td>
<td>(-1.5x_1 + 1.5x_2 - \frac{3}{4}x_3 = -\frac{3}{4}w_3)</td>
<td>0.64</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(= (-\frac{5}{4}, \frac{3}{4}, \frac{3}{4}, \frac{3}{4}, -\frac{5}{4}))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>2</td>
<td>(y - 2x_1 - 3\frac{3}{2}x_2 + \frac{3}{2}x_3)</td>
<td>4.06</td>
<td>2.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(= (\frac{5}{2}, -\frac{3}{2}, 1\frac{3}{4}, -1\frac{3}{2}, \frac{5}{2}))</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>(= (5, 7, 11, 10, 13))</td>
<td>464.00</td>
<td></td>
</tr>
</tbody>
</table>
**B DISTRIBUTION THEORY AND OBSERVED VALUES, QUADRATIC MODEL**

Consider the quadratic model derived from formula (1) together with (11.3.19):

\[
(2) \quad y = \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \sigma z \\
= \alpha_1v_1 + \alpha_2v_2 + \alpha_3v_3 + \sigma z,
\]

where \( z \) is a sample of five from the standard normal.

Under the preceding model we have from Section 11.3F that

\[
(b_1(y), b_2(y), b_3(y))
\]

has a multivariate normal distribution with mean

\[
(\beta_1, \beta_2, \beta_3)
\]

and variance matrix

\[
(3) \quad \sigma^2(X'X)^{-1} = \sigma^2 \begin{pmatrix}
4.6 & -3.3 & \frac{1}{2} \\
-3.3 & 2\frac{7}{10} & -\frac{3}{2} \\
\frac{1}{2} & -\frac{3}{2} & 1\frac{1}{4}
\end{pmatrix},
\]

with values from the array (11.2.14). The observed value of \((b_1(y), b_2(y), b_3(y))\) is

\[
(4) \quad (2, 3\frac{7}{10}, -\frac{3}{4}),
\]

and the fitted quadratic function is represented by the vector

\[
(5) \quad 2x_1 + 3\frac{7}{10}x_2 - \frac{3}{4}x_3 = 9.2w_1 + 1.9w_2 - \frac{3}{4}w_3,
\]

as calculated in Section 11.2F.

And also we have that \(s^2(y)\) has the distribution of \(\sigma^2\chi^2\), where \(\chi^2\) has the chi-square\((n - r)\) distribution; here \(n - r = 2\). The observed value of \(s^2(y) = 4.06\).

**C INERENCE CONCERNING THE VARIANCE**

The bottom element \(s^2(y) = 4.06\) in the SS column of the analysis-of-variance table is an observed value of \(\sigma^2\chi^2\), where \(\chi^2\) is chi-square\((n - r = 2)\); the degrees of freedom \(n - r\) is given in the corresponding DF column, \(n - r = 2\). We take the ratio of sum of squares (SS) to degrees of freedom (DF), obtaining the observed mean square (MS):

\[
(6) \quad MS = \frac{SS}{DF} = \frac{4.06}{2} = 2.03;
\]

this mean square is an unbiased estimate \(s_E^2\) of the variance \(\sigma^2\). A confidence interval for the variance \(\sigma^2\) can be obtained from the chi-square\((n - r)\) distribution for

\[
\chi^2 = \frac{(n - r)s_E^2}{\sigma^2},
\]

following the pattern in Section 8.3B.
D INFERENCE CONCERNING
QUADRATIC EFFECT

For the present example it is natural to investigate whether the quadratic term in (2) is needed for an adequate presentation of the response level. More generally it is natural to investigate whether the last introduced source for response variation is needed for an adequate response presentation.

We can test for the presence of a nonzero value for $\beta_3$ by calculating

$$t_3 = \frac{b_3(y)}{s_E \sqrt{1/14}}$$

and comparing it with the Student($n - r = 2$) distribution. Note that $1/14$ is the reciprocal of the squared length of $\mathbf{w}_3$, and it is available from the right-hand portion of (3). Or more directly we can test by calculating the ratio of mean squares given in the analysis-of-variance table,

$$F = \frac{\text{MS(quadratic)}}{\text{MS(residual)}}$$

and comparing it with the $F(1, 2)$ distribution. Note that if there is no quadratic effect, that is, $\alpha_3 = 0$ in the notation of Sections 11.2 and 11.3, then $F$ is a ratio of mean squares based on ordinary chi-square variables and accordingly has the ordinary $F$ distribution. On the other hand, if there is a quadratic effect, that is, $\alpha_3 \neq 0$, then $F$ has the noncentral $F$ distribution and tends to be larger in value. The $t$ and $F$ calculations are easily seen to be equivalent; they are just different ways of comparing the projection in $\mathcal{L}(\mathbf{v}_3)$ with the projection in $\mathcal{L}(X)$.

The observed value of $F = 0.64/2.03 = 0.32$. This is not an extremely large value; in fact, it is to the left of the center of the $F$ distribution (which is approximately unity). Thus we have no evidence of a quadratic dependence. It is natural, then, to consider withdrawing to the linear model

$$y = \beta_1 + \beta_2 x + \sigma z$$

and analyzing accordingly. Note that our calculation sequence allows for this: examining quadratic after linear and constant.

On the assumption of the quadratic model we can, of course, calculate confidence intervals for the regression coefficients $\beta_1$, $\beta_2$, and $\beta_3$. For $\beta_3$ we have a Student($n - r = 2$) distribution for

$$t_3 = \frac{b_3(y) - \beta_3}{s_E \sqrt{1/14}}$$

This gives the following 95 percent confidence interval for $\beta_3$:

$$(3/14 \pm 4.30 \sqrt{2.03} \sqrt{1/14}) = (-1.852, 1.423).$$

Recall the note following (7). Confidence intervals for $\beta_2$ and $\beta_1$ can be calculated analogously; for example, for $\beta_2$ we have

$$(3/2^3 \pm 4.30 \sqrt{2.03} \sqrt{24/14}) = (-6.83, 13.20).$$
E THE AFFINE MODEL

We now consider the analysis on the assumption that the affine model

\[ y = \beta_1 + \beta_2 x + cz \]

adequately describes the response and the underlying variation \( z \) is standard normal. The analysis-of-variance table for fitting constant, and then linear dependence on \( x \), was obtained in Section 11.2E; now see Table 11.9.

<table>
<thead>
<tr>
<th>Source</th>
<th>Basis</th>
<th>DF</th>
<th>Projection</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
</table>
| Constant  | \( \nu_1 \) | 1  | 9.2\( x_1 \) = 9.2\( w_1 \)  
| Linear    | \( \nu_2 \) | 1  | -5.7\( x_1 \) + 1.9\( x_2 \) = 1.9\( w_2 \)  
\[ = (-3.8, -1.9, 0, 1.9, 3.8) \] | 36.10  | 36.10 |
| Residual  | \( y - 3.5x_1 - 1.9x_2 \)  
\[ = (-0.4, -0.3, 1.8, -1.1, 0) \] | 4.70   | 1.57 |
|           | \( 5 \) \( (5, 7, 11, 10, 13) \) | 464.00 |    |

F DISTRIBUTION THEORY
AND OBSERVED VALUES:
AFFINE MODEL

Consider the quadratic model derived from formula (11) together with (11.3.19):

\[ y = \beta_1 x_1 + \beta_2 x_2 + cz \]

\[ = \alpha_1 \nu_1 + \alpha_2 \nu_2 + cz, \]

where \( z \) is a sample of five from the standard normal.

Under the preceding model we have from Section 11.3 that

\[(b_1(y), b_2(y))\]

has a bivariate normal distribution with mean

\[(\beta_1, \beta_2)\]

and variance matrix

\[\sigma^2(X'X)^{-1} = \sigma^2 \begin{pmatrix} 1.1 & -0.3 \\ -0.3 & 0.1 \end{pmatrix}\]

with values from the array (11.2.13). Note that \((b_1(y), b_2(y))\) refers to the calculations for the affine model as examined in Section 11.2E; thus these \( b \)'s are different from those considered earlier in Section B. The \( a \)'s for the model here would, however, be the same as those for the model in Section B; an advantage from orthogonality. The
observed value of \((b_1(y), b_2(y))\) is
\[(14) \quad (3.5, 1.9)\]
and the fitted affine function is represented by the vector
\[(15) \quad 3.5x_1 + 1.9x_2 = 9.2w_1 + 1.9w_2,\]
as calculated in Section 11.2E.
And also we have that \(s^2(y)\) has the distribution of \(\sigma^2\chi^2\), where \(\chi^2\) has the chi-square\((n - r)\) distribution; here \(n - r = 3\). The observed value of \(s^2(y) = 4.70\).

\section{Inference Concerning Linear Effect}

We now assume there is negligible quadratic effect and consider the analysis based essentially on the simpler model
\[(16) \quad y = \beta_1x_1 + \beta_2x_2 + \sigma z.\]
Tests and confidence intervals are available following the pattern in Section D.

This procedure based on negligible quadratic effect requires some caution. In Section D we tested for quadratic effect and we have now withdrawn to the simpler model. As part of withdrawing we pooled the quadratic SS and the residual SS of the table in Section A to obtain the residual SS of the table in Section E. This pooling procedure has some rather serious potentials for bias. For the present we can allow that the quadratic SS may contain minor quadratic effects and use the original table in Section A to obtain our error variance estimate and follow the pattern in Section C for inference concerning the variance.

Thus on the assumption of the affine model together with the precaution just mentioned for error variance, we can test for the presence of a nonzero value for \(\beta_2\) by calculating
\[(17) \quad t_2 = \frac{b_2(y)}{s_\theta \sqrt{0.1}}\]
with \(s_\theta\) from Section A and comparing it with the Student\((n - r)\) distribution; here \(n - r = 2\), the DF for the error variance. Note that 0.1 is the reciprocal of the squared length of \(w_2\), and it is available from the right-hand portion of (13). Or more directly we can test by calculating
\[(18) \quad F = \frac{\text{MS(linear)}}{\text{MS(residual)}}\]
and comparing it with the \(F(1, 2)\) distribution; note that these DF come from the table in Section A.

The observed value of \(F = 36.10/2.03 = 17.8\). The 5 percent point is 18.51 and the 1 percent point is 98.49. Thus there is some suggestion of a nonzero value for \(\beta_2\). Note that if we had used the pooled estimate for the variance we would have had an additional DF for the denominator, and the \(F\) test would, in fact, give significance.

On the assumption of the linear model together with the precaution we can
calculate confidence intervals for the regression coefficients $\beta_1$ and $\beta_2$. For $\beta_2$ we have a Student($m = 2$) distribution for

$$t_2 = \frac{b_2(y_i - \beta_2)}{s_x \sqrt{0.1}}$$

using the unpooled estimate. This gives the following 95 percent confidence interval for $\beta_2$:

$$1.9 \pm 4.30 \sqrt{2.03 \sqrt{0.1}} = (-0.04, 3.84).$$

**H PROBLEMS**

1. Calculate formula (5) in terms of the orthonormal vectors $v_1, v_2$, and $v_3$. Calculate formula (15) in terms of the orthonormal vectors $v_1$ and $v_2$.

2. For the model in Sections A and B, record numerical values for the entries in the matrix equation $Xb = Va$ from (11.3.18); here $r = 3$. For the model in Sections E and B, record numerical values for the entries in the equation $Xb = Va$; here $r = 2$. Recall Problem 11.2.3; compare with the preceding problem.

3. (continuation of Problem 11.2.1) Assume the affine model $y = \beta_1 + \beta_2 x + az$, where $z$ is standard normal. Test the hypothesis: $\beta_2 = 0$. Form a 95 percent confidence interval for $\beta_2$.

4. (continuation of Problem 11.2.2) Assume the quadratic model $y = \beta_1 + \beta_2 x + \beta_3 x^2 + az$, where $z$ is standard normal. Test the hypothesis $\beta_3 = 0$. If no significance, assume an affine model and test the hypothesis $\beta_2 = 0$; form a 95 percent confidence interval for $\beta_2$; recall the precaution in Section G.

5. (continuation of Problem 11.2.4) Assume the affine model $y = \beta_1 + \beta_2 x_1 + \beta_3 x_2$. Test the hypothesis $\beta_3 = 0$. If no significance, assume an affine model and test the hypothesis $\beta_2 = 0$; form a 95 percent confidence interval for $\beta_2$; recall the precaution in Section G.

**NOTES AND REFERENCES**

The linear model is the basic model for large areas of statistics, including the fundamental areas of experimental design. For material on linear models primarily based on normal variation, see the references cited below. For some discussion of linear models with other distribution forms for error, see Fraser (1976).

A computer program for analyzing the location-scale model is available to instructors on request; write to: Statistics Section, Attn.: Gordon Fick, Mathematics Department, University of Toronto, Canada.


In preceding chapters we have examined the basic methods of statistical inference and applied the methods to the very widely used linear models.

In Chapter 12 we shall survey experimental design, the design of experimental investigations to determine cause–effect relations; in particular, we discuss the basic principles of experimentation, control, randomization, and replication. In Section 12.1 the factorial design is introduced and the single-factor design is examined in detail. In Section 12.2 the two-factor design is examined in detail. Then in Supplement Sections 12.3, 12.4, and 12.5 we examine three-factor designs, randomized blocks, and analysis of covariance.
THE DESIGN OF EXPERIMENTS

12.1 ONE-FACTOR DESIGN

The design of experiments, the most important part of statistics, has been mentioned briefly in Sections 1.1 and 7.1. In this chapter we discuss some central material of the design of experiments.

The investigation of a random system becomes an experiment if the investigator makes designed changes in the input variables or factors with the express purpose of determining cause–effect relations. In Sections 12.1 and 12.2 and Supplement Section 12.3 we examine the basic factorial design for one, two, and three factors, respectively. Then in Supplement Sections 12.4 and 12.5 we acknowledge the inevitable variation that occurs in the initial conditions and the input materials to a system and we examine two methods for minimizing the influence of such variation. If the conditions or materials can be grouped into blocks with less variation within a block, the randomized block design in Supplement Section 12.4 can be used with the factorial design. And if a concomitant variable that describes the variation is available, the analysis of covariance in Supplement Section 12.5 can be used with the factorial design. The development in this chapter emphasizes basic principles: the basic principles necessary to the design of experiments, and the basic principles underlying the algebraic results.

On an individual basis the algebraic results in this chapter can each be verified fairly directly. But the why is important—not just for the developments in this chapter but for the flexibility to handle new methods and new designs in a direct way. Thus at several points we may have somewhat more vector notation than is usual, but only so we can use the basic principles in a direct way to obtain the needed algebraic results.

A FACTORIAL DESIGN

We first discuss briefly some aspects of experimental design—and we do this in the context of a random system with a variety of input variables that may influence the response of the system. A factor is an input or input variable that may affect the response of the system: for example, the pressure at which a chemical process is operated, the kind of soil preparation for an agricultural experiment, the type of diet in a growth investigation, the variety of a grain in an agricultural experiment. The level of a factor is the particular value of the input or input variable: for example, a particular pressure, a particular diet, a particular variety of grain.
The design factors are the factors under study, the factors that the investigator changes in his designed investigation of the system. A particular combination of levels for the design factors is called a treatment.

The investigator when possible holds constant or controls other factors. This is done to ensure that such factors cannot change and thus cause changes in the response that might be attributable to the design factors.

This leaves the factors or influences that are not directly changed or controlled. These cover differences between experimental units, the basic units such as animals or plots of land that receive the treatments being investigated. They also cover things such as variation in input material, and show trends with respect to location and time. By extending the definition of experimental unit appropriately to include the basic unit at a given location and given time, these various factors or influences can all be embraced under "experimental units."

In practice there rarely is reliable information as to the actual variation from experimental unit to unit. Thus with treatments applied to units, an apparent difference in the effects of the treatments can be due, in fact, to differences between the units. The only reliable way of dealing with the actual variation from unit to unit is to randomly assign the treatments to the units. This randomization changes the variation from unit to unit into a random variation affecting the response; and this response variation or error can be directly assessed by the probability methods we have been developing.

We indicate design factors by capital letters: A, B, C, . . . . And we add subscripts to denote the chosen levels for the factors:

\[ A : A_1, A_2, \ldots, A_a, \]
\[ B : B_1, B_2, \ldots, B_b, \]
\[ C : C_1, C_2, \ldots, C_c. \]

An experimental design is called a factorial design if each combination of levels is examined the same number of times, say \( n \). For example, with three factors as indicated above, each treatment \( A_i B_j C_k \) is examined \( n \) times, giving a total of \( N = abc n \) observations on the system. The design is called fully randomized if the \( N \) observations, \( n \) for each treatment, are randomly assigned to \( N \) experimental units or to the \( N \) times or locations at which the system is performed. In this section we examine the fully randomized factorial design with one design factor, and then in Section 12.2 and Supplement: Section 12.3 we examine the factorial design with two and three design factors, respectively.

In Supplement Section 12.4 we consider the case where experimental units can be grouped to form blocks with less variation within blocks than between blocks. The treatments can then be randomly assigned within blocks, producing the randomized block design.

In Supplement Section 12.5 we consider how some related or concomitant variable can give information on the variation attributable to the experimental units and how the analysis of covariance with such a concomitant variable can be used to assess more precisely the dependence of the response on the treatments.

**B ONE-FACTOR DESIGN**

Consider a single design factor \( A \) with chosen levels \( A_1, \ldots, A_a \). And suppose that \( n_1 \) observations are planned for level \( A_1, \ldots, \) and \( n_a \) observations for level \( A_a \), making
a total of $N = \sum n_j$ observations. For the one-factor design we can allow differing numbers at the different factor levels without serious complication. However, for several factors and with randomized blocks this departure from the formal definition of a factorial design can lead to severe difficulties. We assume that the observations are fully randomized against the experimental units, time, or any uncontrolled factor.

The response vector for the $N$ observations can be presented in an array such as the following:

$$
\begin{array}{c|c|c}
  x_1 & A_1 & y_{11} \ldots y_{1n_1} \\
  \vdots & \vdots & \vdots \\
  x_a & A_a & y_{a1} \ldots y_{an_a} \\
\end{array}
$$

In certain cases the factor $A$ may be represented by a controllable variable $x$; the values for such a variable are indicated in the left column.

An investigation is concerned with whether the factor $A$ affects the response and, if so, in what manner. We assume that the response is presented in such a form that the variation has the same scaling at the various possible response levels. And we consider the analysis for the case of approximately normal variation. The analysis for other distribution forms with a parameter $\lambda$ can be investigated using the theory in Section 11.3; this does, however, require integration over a space with dimension equal to the number of primary parameters.

If the factor $A$ does not affect the response, then the model for a single observation and the model for the vector of $N$ observations are

$$
(1) \quad y_{i2} = \mu + \sigma z_{i2}, \quad y = \mu \mathbf{1} + \sigma z,
$$

where $z$ is a sample of $N$ from the standard normal. The fitted model for the general level is then the projection $p_t(y)$ into $\mathcal{L}(1)$; the projection and squared length of the projection are

$$
(2) \quad \begin{array}{c|c|c}
  A_1 & \bar{y} & \ldots \bar{y} \\
  \vdots & \vdots & \vdots \\
  A_a & \bar{y} & \ldots \bar{y} \\
\end{array}
$$

$$
SS_t = SS_{(1)} = N\bar{y}^2 = T^2/N.
$$

Note that $T_\cdot = \sum i y_{i2}$ is the overall total and $\bar{y} = T_\cdot / N$ is the overall average.

If the factor $A$ does affect the response and if we allow for this in full generality with a level $\mu_i$ at the factor level $A_i$, then the model is

$$
(3) \quad y_{i2} = \mu_i + \sigma z_{i2}, \quad y = \sum_{i=1}^a \mu_i \mathbf{1}_i + \sigma z,
$$

where $z$ is a sample of $N$ from the standard normal and $\mathbf{1}_i$ is the vector with 1's in positions corresponding to the level $A_i$ and 0's elsewhere, as given by the following array:

$$
\begin{array}{c|cccc}
  A_1 & 0 & \cdots & 0 \\
  \vdots & \vdots & \ddots & \vdots \\
  A_i & 1 & \cdots & 1 \\
  \vdots & \vdots & \ddots & \vdots \\
  A_a & 0 & \cdots & 0 \\
\end{array}
$$

The fitted model is then the projection $p_{(2)}(y)$ into $\mathcal{L}(1, \ldots, 1_a)$; it gives a separate average for each factor level. The projection and squared length are
\[ A_1 \begin{array}{ccc} \bar{y}_1 & \cdots & \bar{y}_1 \\ \vdots & & \vdots \\ A_a & \bar{y}_a & \cdots & \bar{y}_a \end{array} \]

\[ SS_{(2)} = \sum_i n_i \bar{y}_i^2 = \sum_i T_i^2 / n_i. \]

Note that \( T_i = \sum y_{is} \) is the total for the level \( A_i \) and \( \bar{y}_i = T_i / n_i \) is the average for that level; for the justification note that the vectors \( 1_1, \ldots, 1_a \) are orthogonal.

In going from the first model to the second model we are allowing the general level to change with the factor \( A \). The fitted vector for such changes is the projection \( p_2 = p_{(2)} - p_1 \) into \( L(1_1, \ldots, 1_a \perp 1) \); recall the discussion in Section 11.2A. The projection and squared length are

\[ A_1 \begin{array}{ccc} \bar{y}_1 - \bar{y} & \cdots & \bar{y}_1 - \bar{y} \\ \vdots & & \vdots \\ A_a & \bar{y}_a - \bar{y} & \cdots & \bar{y}_a - \bar{y} \end{array} \]

\[ SS_2 = \sum_i n_i (\bar{y}_i - \bar{y})^2 = SS_{(2)} - SS_{(1)} = \sum_i T_i^2 / n_i - T^2 / N. \]

Note that the projection and the squared length are obtained easily by differencing. And note also the interesting identity for the squared length, as calculated directly from the projection and by differencing.

The response vector itself and the square length of the response are

\[ A_1 \begin{array}{cc} y_{11} & \cdots & y_{1n_1} \\ \vdots & & \vdots \\ A_a & y_{a1} & \cdots & y_{an_a} \end{array} \]

\[ SS_{(3)} = \sum_i y_{is}^2. \]

The step from the second fitted model to the response vector is concerned with the deviations from the average at each factor level. The fitted vector is the projection \( p_3 = y - p_{(2)} \) into \( L^1(1_1, \ldots, 1_a) \); the projection and squared length are

\[ A_1 \begin{array}{ccc} y_{11} - \bar{y}_1 & \cdots & y_{1n_1} - \bar{y}_1 \\ \vdots & & \vdots \\ A_a & y_{a1} - \bar{y}_a & \cdots & y_{an_a} - \bar{y}_a \end{array} \]

\[ SS_3 = \sum_i (y_{is} - \bar{y}_i)^2 = SS_{(3)} - SS_{(2)} = \sum_i y_{is}^2 - \sum_i T_i^2 / n_i. \]

**TABLE 12.1**

<table>
<thead>
<tr>
<th>Source</th>
<th>Subspace</th>
<th>DF</th>
<th>Projection</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>Mean</td>
<td>( L(1) )</td>
<td>1</td>
<td>( \bar{y} )</td>
</tr>
<tr>
<td>(2)</td>
<td>( A )</td>
<td>( L(1_1, \ldots, 1_a) )</td>
<td>( a )</td>
<td>( \sum \bar{y}_i )</td>
</tr>
<tr>
<td>(3)</td>
<td>Response</td>
<td>( R^N )</td>
<td>( N )</td>
<td>( y )</td>
</tr>
</tbody>
</table>
The cumulative table recording projections and squared lengths is given in Table 12.1. The difference table recording projections and squared lengths for orthogonal subspaces is then given in Table 12.2. This is an analysis-of-variance table, analyzing the variation among response values in accord with patterns based on the various models.

**TABLE 12.2**

Analysis-of-variance table

<table>
<thead>
<tr>
<th>Source</th>
<th>Space</th>
<th>DF</th>
<th>Projection</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mean</td>
<td>1</td>
<td>$\bar{y}$</td>
<td>$N\bar{y}^2 = SS_{(1)}$</td>
</tr>
<tr>
<td>2</td>
<td>$A$</td>
<td>$a - 1$</td>
<td>$\bar{y}_i$</td>
<td>$\sum n_i(\bar{y}<em>i - \bar{y})^2 = SS</em>{(a)} - SS_{(1)}$</td>
</tr>
<tr>
<td>3</td>
<td>Residual $L^*(1, \ldots, 1_a)$</td>
<td>$N - a$</td>
<td>$\bar{y}$</td>
<td>$\sum (y_i - \bar{y})^2 = SS_{(a)} - SS_{(2)}$</td>
</tr>
</tbody>
</table>

**C THE ANALYSIS**

Now consider the problem of testing the hypothesis that changes in the factor $A$ do not affect the response level.

For this we note some distribution results available from Section 11.3. Under the general model (3) we have that

$$(7) \quad SS_3 = \sum (y_{is} - \bar{y})^2$$

has the distribution of $\sigma^2 \chi^2_3$, where $\chi^2_3$ has the chi-square $(N - a)$ distribution. Under the special model (1) we have that

$$(8) \quad SS_2 = \sum n_i(\bar{y}_i - \bar{y})^2$$

has the distribution of $\sigma^2 \chi^2_i$, where $\chi^2_i$ has the chi-square $(a - 1)$ distribution; however, under the general model with the $\mu_i$ not all equal it tends to be larger in value and has the noncentral chi-square distribution in Problem 3.6.3.

If the hypothesis is true, then both (7) and (8) lead (division by appropriate DF) to mean squares (MS) that are estimates of $\sigma^2$. If the hypothesis is not true, then (8) tends to be larger. The hypothesis can be tested by calculating

$$F = \frac{SS_2/(a - 1)}{SS_3/(N - a)} = \frac{MS_2}{MS_3}$$

and comparing the value with the $F$ distribution $(a - 1, N - a)$; unusually large values are indicative of effects due to the factor $A$.

**EXAMPLE 1**

Consider a catalyst $A$ at two levels $A_1$ and $A_2$ with three observations at level $A_1$ and two observations at level $A_2$. The five observations were randomized against time, giving the response data shown in Table 12.3 (rounded to the nearest integer for
easy arithmetic. Some calculations based on the formulas in Section B are recorded on the right. The cumulative table is shown in Table 12.4. The analysis-of-variance table is obtained by differencing (Table 12.5).

### Table 12.3

**Response data**

<table>
<thead>
<tr>
<th>Data</th>
<th>Totals</th>
<th>Calculations</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>4 3 5</td>
<td>12</td>
</tr>
<tr>
<td>$A_2$</td>
<td>2 1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
</tr>
</tbody>
</table>

### Table 12.4

**Cumulative table**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Projection</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Mean</td>
<td>1</td>
<td>$\begin{pmatrix} 3 &amp; 3 &amp; 3 \ 3 &amp; 3 \end{pmatrix}$</td>
<td>45</td>
</tr>
<tr>
<td>(2) $A$</td>
<td>2</td>
<td>$\begin{pmatrix} 4 &amp; 4 &amp; 4 \ 1.5 &amp; 1.5 \end{pmatrix}$</td>
<td>52.5</td>
</tr>
<tr>
<td>(3) Response</td>
<td>5</td>
<td>$\begin{pmatrix} 4 &amp; 3 &amp; 5 \ 2 &amp; 1 \end{pmatrix}$</td>
<td>55</td>
</tr>
</tbody>
</table>

### Table 12.5

**Analysis-of-variance table**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Projection</th>
<th>SS</th>
<th>MS</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Mean</td>
<td>1</td>
<td>$\begin{pmatrix} 3 &amp; 3 &amp; 3 \ 3 &amp; 3 \end{pmatrix}$</td>
<td>45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 $A$</td>
<td>1</td>
<td>$\begin{pmatrix} 1 &amp; 1 &amp; 1 \ -1.5 &amp; -1.5 \end{pmatrix}$</td>
<td>$7.5$</td>
<td>$7.5$</td>
<td>$7.5/\frac{5}{6} = 9$</td>
</tr>
<tr>
<td>3 Residual</td>
<td>3</td>
<td>$\begin{pmatrix} 0 &amp; -1 &amp; 1 \ -\frac{1}{6} &amp; -\frac{1}{6} \end{pmatrix}$</td>
<td>$2\frac{1}{6}$</td>
<td>$\frac{8}{6}$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>5</td>
<td>$\begin{pmatrix} 4 &amp; 3 &amp; 5 \ 2 &amp; 1 \end{pmatrix}$</td>
<td>55</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Now suppose that the general model (3) holds and we want to test the hypothesis that the factor $A$ does not affect the general level of the response. We can test for the factor $A$ by calculating

$$F = \frac{MS_2}{MS_3} = \frac{7\frac{1}{2}}{\frac{8}{6}} = 9$$

and comparing it with the $F(1, 3)$ distribution. The value 9 is slightly less than the 5 percent point 10.1, giving some moderate suggestion that the factor $A$ affects the response at the levels tested.
D WITH A QUANTITATIVE FACTOR

Now suppose that the factor A represents an input variable x. If there is dependence on the factor A, the experimenter will want to know the form of the dependence, whether it depends, say, linearly, quadratically, or perhaps more generally on x. He can then examine some models that fall between the no-dependence case

(1) \( y_{is} = \mu + \sigma z_{is}, \quad y = \mu \mathbf{1} + \sigma \mathbf{z} \)

and the general dependence case

(3) \( \mu_{ij} = \mu_i + \sigma z_{is}, \quad y = \sum \mu_i \mathbf{1}_i + \sigma \mathbf{z} \).

Consider the case of linear and quadratic dependence on x. The model presenting dependence that is at most linear is

(9) \( y_{is} = \beta_1 x_i + \beta_2 x_i^2 + \sigma z_{is}, \quad y = \beta_1 \mathbf{1} + \beta_2 \mathbf{x}_i + \sigma \mathbf{z} \).

And the model presenting dependence that is at most quadratic is

(10) \( y_{is} = \beta_1 \mathbf{1} + \beta_2 \mathbf{x}_i + \beta_3 \mathbf{x}_i^2 + \sigma z_{is}, \quad y = \beta_1 \mathbf{1} + \beta_2 \mathbf{x}_i + \beta_3 \mathbf{x}_i^2 + \sigma \mathbf{z} \).

The projections and squared lengths for the succession of models (1), (9), (10), and (3) can be obtained following the methods of Section 11.2. Alternatively some of them can be obtained directly by finding basis vectors for the successive orthogonal subspaces. For this, consider the sequence of vectors corresponding to the constant, linear, quadratic dependence on x:

\[
\mathbf{1} = (1, \ldots, 1; \ldots; 1, \ldots, 1)',
\]

\[
\sum x_i \mathbf{1}_i = (x_1, \ldots, x_1; \ldots; x_a, \ldots, x_a)',
\]

\[
\sum x_i^2 \mathbf{1}_i = (x_1^2, \ldots, x_1^2; \ldots; x_a^2, \ldots, x_a^2)'.
\]

The second vector can be orthogonalized to the first vector, giving

\[
\sum (x_i - \bar{x}) \mathbf{1}_i = (x_1 - \bar{x}, \ldots, x_1 - \bar{x}; \ldots; x_a - \bar{x}, \ldots, x_a - \bar{x})',
\]

where \( \bar{x} = \Sigma n_i x_i / N \) is the average for the N values of x. The third vector as orthogonalized to the first two vectors is

\[
\sum (x_i^2 - c - dx_i) \mathbf{1}_i = (x_1^2 - c - dx_1, \ldots, x_1^2 - c - dx_1; \ldots; x_a^2 - c - dx_a, \ldots, x_a^2 - c - dx_a)',
\]

where the coefficients c and d are obtained from the usual least-squares equations,

\[
Nc + \sum n_i x_i d = \sum n_i x_i^2.
\]
\[ \sum n_i x_i c + \sum n_i x_i^2 d = \sum n_i x_i^3. \]

Note, of course, that the solution for \( c \) and \( d \) can be obtained as part of the matrix procedures in Section 11.2 applied to

\[
\begin{align*}
N & \quad \sum n_i x_i \quad \sum n_i x_i^2 \quad \sum y_{is} \\
\sum n_i x_i \quad \sum n_i x_i^2 & \quad \sum n_i x_i^3 \quad \sum x_i y_{is} \\
\sum n_i x_i^2 & \quad \sum n_i x_i^3 \quad \sum n_i x_i^4 \quad \sum x_i^2 y_{is}.
\end{align*}
\]

The analysis-of-variance table can be presented in terms of the orthogonalized vectors (Table 12.6). The projections are simply related to the expressions given in the table; for example, the quadratic projection and squared length are

\[
\frac{\sum (x_i^2 - c - dx_i)y_{is}}{\sum n_i(x_i^2 - c - dx_i)^2} = \frac{\sum (x_i^2 - c - dx_i)y_{is}}{\sum n_i(x_i^2 - c - dx_i)^2}.
\]

These are special cases of the formulas for the projection of a vector \( y \) on a vector \( w \) and the corresponding squared length.

\[
\frac{w'y}{w'w}, \quad \frac{w'y}{w'w} \cdot w'y.
\]

To test for dependence beyond quadratic we calculate

\[
F = \frac{MS_4}{MS_5}
\]

**TABLE 12.6**

*Analysis-of-variance table*

<table>
<thead>
<tr>
<th>Source</th>
<th>Subspace</th>
<th>DF</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mean</td>
<td>1</td>
<td>( \frac{\left( \sum y_{is} \right)^2}{N} )</td>
</tr>
<tr>
<td>2</td>
<td>A linear</td>
<td>1</td>
<td>( \frac{\sum (x_i - \bar{x})y_{is}}{\sum n_i(x_i - \bar{x})^2} )</td>
</tr>
<tr>
<td>3</td>
<td>quadratic</td>
<td>1</td>
<td>( \frac{\sum (x_i^2 - c - dx_i)y_{is}}{\sum n_i(x_i^2 - c - dx_i)^2} )</td>
</tr>
<tr>
<td>4</td>
<td>other</td>
<td>( a - 3 )</td>
<td>Difference</td>
</tr>
<tr>
<td></td>
<td>Subtotal</td>
<td>( a - 1 )</td>
<td>( \sum n_i(\bar{y}_i - \bar{y})^2 )</td>
</tr>
<tr>
<td>5</td>
<td>Residual</td>
<td>( \sum_{i=1}^{n} \mathbf{1}_i )</td>
<td>( N - a )</td>
</tr>
<tr>
<td>Total</td>
<td>( R^n )</td>
<td>( \frac{N}{N} )</td>
<td>( \sum y_{is}^2 )</td>
</tr>
</tbody>
</table>
and compare it with the \( F(a - 3, N - a) \) distribution.

Now suppose that we are prepared to assume that dependence is at most quadratic; then we can test for quadratic dependence by calculating

\[
F = \frac{MS_3}{MS_5}
\]

and comparing it with the \( F(1, N - a) \) distribution.

And suppose we are prepared to assume that dependence is at most linear; then we can test for dependence by calculating

\[
F = \frac{MS_2}{MS_5}
\]

and comparing it with the \( F \) distribution \((1, N - a)\).

E EXAMPLE WITH A QUANTITATIVE FACTOR

Consider a biological system with a factor \( A \) at five levels corresponding to \( x = 1, 2, 3, 4, 5 \). Two observations were obtained at each factor level, giving the data of Table 12.7 (rounded to the nearest integer for easy arithmetic). The analysis-of-variance table without separation of linear and quadratic components is shown as Table 12.8.

**TABLE 12.7**
Response data

<table>
<thead>
<tr>
<th>( x )</th>
<th>Data</th>
<th>Totals</th>
<th>Calculations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>6</td>
<td>13</td>
</tr>
</tbody>
</table>

\[ y_{1a}^2 = 236 \]

\[ (5^2 + \cdots + 13^2)/2 = 232 \]

\[ 46^2/10 = 211.6 \]

**TABLE 12.8**
Analysis-of-variance table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1</td>
<td>211.6</td>
</tr>
<tr>
<td>( A )</td>
<td>4</td>
<td>20.4 = 232 - 211.6</td>
</tr>
<tr>
<td>Residual</td>
<td>5</td>
<td>4  = 236 - 232</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>236</td>
</tr>
</tbody>
</table>

The vector for linear dependence orthogonal to the mean is based on the input variable \( x - 3 \); recall Section 11.2E. The corresponding squared projection is

\[
\frac{[-2(5) - 1(7) + 0(11) + 1(10) + 2(13)]^2}{8 + 2 + 0 + 2 + 8} = \frac{19^2}{20} = 18.05;
\]

note that the value \(-2\) for \( x - 3 \) occurs twice, \(-1\) occurs twice, and so on.
The vector for quadratic dependence, orthogonal to the mean and linear vectors, is based on the input variable $x^2 - 6x + 7$; recall Section 11.2F. The corresponding squared projection is

$$\frac{[2(5) - 1(7) - 2(11) - 1(10) + 2(13)]^2}{8 + 2 + 8 + 2 + 8} = \frac{(-3)^2}{28} = 0.32.$$

The coefficients for these orthogonalized vectors can be obtained from the matrix procedures in Section 11.2:

$$\begin{array}{ccc}
10 & 30 & 110 \\
30 & 110 & 450 \\
1 & 3 & 11 \\
0 & 20 & 120 \\
1 & 0 & -7 \\
0 & 1 & 6 \\
\end{array}$$

The analysis-of-variance table can now be rewritten as shown in Table 12.9.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1</td>
<td>211.6</td>
<td></td>
</tr>
<tr>
<td>A linear</td>
<td>1</td>
<td>18.05</td>
<td>18.05</td>
</tr>
<tr>
<td>A quadratic</td>
<td>1</td>
<td>0.32</td>
<td>0.32</td>
</tr>
<tr>
<td>Other</td>
<td>2</td>
<td>2.03</td>
<td>1.02</td>
</tr>
<tr>
<td>Residual</td>
<td>5</td>
<td>4</td>
<td>0.8</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>236</td>
<td></td>
</tr>
</tbody>
</table>

The more-than-quadratic and quadratic components show no significance. Then on the assumption that dependence is at most linear, the test for dependence is obtained by calculating

$$F = \frac{18.05}{0.8} = 22.5$$

and comparing it with $F(1, 5)$ distribution; the 5 percent value is 6.6 and the 1 percent value is 16.3. The observed value is extreme, thus providing strong evidence against the hypothesis of no linear dependence.

The example in Sections 11.2 and 11.4 is based on the row totals of the present example. Correspondingly, the sum of squares in the first four rows of the preceding table are $1/2$ the corresponding values in the first table in Section 11.4A. The present analysis has two observations at each factor level and thus produces an estimate of the variance independent of the form of dependence on the input variable $x$. This permits a test for dependence beyond quadratic.

Comparison of the two analyses also illustrates two different calculations (by differencing and direct) for some elements in the analysis-of-variance table.
F EXERCISES

Recall the cautionary remarks at the beginning of Section 11.2H.

1 Fourteen pigs were randomly assigned to three diets for a two-month period. The following data record weight increase:

\[
\begin{array}{c|ccccc}
T_1 & 23 & 13 & 28 & 23 & 35 \\
T_2 & 33 & 31 & 22 & 40 & 31 \\
T_3 & 40 & 33 & 38 & 44 & \\
\end{array}
\]

Test for dependence on the diets.

2 Three methods of catalyzing a chemical process produced the following data:

\[
\begin{array}{c|cccc}
C_1 & 47.2 & 49.8 & 48.4 & 48.7 \\
C_2 & 50.1 & 49.3 & 51.5 & 50.9 \\
C_3 & 49.1 & 53.2 & 51.2 & 52.8 & 52.3 \\
\end{array}
\]

The variation is known to be approximately normal. Construct the analysis-of-variance table and test for the equivalence of the catalytic methods.

3 A fertilizer \( A \) was used at five levels \( x \) on 10 plots of land, giving two yields for each level; the variation is known to be approximately normal.

<table>
<thead>
<tr>
<th>( x )</th>
<th>2.6</th>
<th>2.7</th>
<th>2.8</th>
<th>2.9</th>
<th>3.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>12.2</td>
<td>12.5</td>
<td>12.5</td>
<td>13.1</td>
<td>13.5</td>
</tr>
<tr>
<td></td>
<td>12.0</td>
<td>12.5</td>
<td>12.9</td>
<td>12.9</td>
<td>13.5</td>
</tr>
</tbody>
</table>

Construct the analysis-of-variance table with entries for Mean, Factor \( A \), and Residual. Test for dependence on the factor \( A \).

4 (continuation) Separate out components for linear, quadratic, and other dependence on the factor \( A \); use the analysis from Problems 11.2.2 and 11.4.4. Determine a 95 percent confidence interval for the coefficient, giving the linear dependence on \( x \).

5 Three procedures, \( T_1, T_2, \) and \( T_3 \), were used to determine the melting point of a hydrocarbon; the variation is approximately normal.

\[
\begin{array}{lccc}
T_1 & T_2 & T_3 \\
214.2 & 213.2 & 211.5 \\
213.1 & 212.1 & 211.3 \\
213.8 & 213.2 & \\
\end{array}
\]

Test the equivalence of the procedures.

6 The compressive load at breakdown of three alloys, \( A_1, A_2, \) and \( A_3 \), were determined for samples of three from each; the variation is approximately normal.
<table>
<thead>
<tr>
<th>(A_1)</th>
<th>(A_2)</th>
<th>(A_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>38.5</td>
<td>40.1</td>
<td></td>
</tr>
<tr>
<td>39.8</td>
<td>41.5</td>
<td>43.2</td>
</tr>
<tr>
<td>37.2</td>
<td>39.3</td>
<td>42.2</td>
</tr>
</tbody>
</table>

Test the equivalence of the alloys for compressive strength.

12.2

TWO-FACTOR DESIGN

We now examine the fully randomized two-factor design. The analogous design with randomized blocks is examined briefly in Supplement Section 12.4.

A TWO-FACTOR DESIGN

Consider a random system with two design factors, \(A\) and \(B\); let \(A_1, \ldots, A_a\) be the chosen levels for the first factor and \(B_1, \ldots, B_b\) be the chosen levels for the second factor. And suppose that \(n\) response observations are planned for each treatment \(A_iB_j\), making a total of \(N = abn\) observations. We assume that the observations are fully randomized against experimental units, time, or any uncontrolled factor.

The response vector for the \(N\) observations can be presented in a matrix array as follows:

\[
\begin{array}{cccc}
\left|\begin{array}{cccc}
\vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots \\
\end{array}\right|
\end{array}
\]

\[
\begin{array}{cccc}
x_{11} & x_{12} & x_{13} & x_{14} \\
A_1 & y_{111} & \cdots & y_{11n} \\
\vdots & \vdots & \vdots & \vdots \\
A_a & y_{a11} & \cdots & y_{a1n} \\
\end{array}
\begin{array}{cccc}
x_{21} & x_{22} & x_{23} & x_{24} \\
B_1 & y_{1b1} & \cdots & y_{1bn} \\
\vdots & \vdots & \vdots & \vdots \\
B_b & y_{ab1} & \cdots & y_{abn} \\
\end{array}
\]

In certain cases the factors may correspond to input variables \(x_1\) and \(x_2\); the values for such variables are indicated beside the factor levels.

An investigator is concerned with whether the factors \(A\) and \(B\) affect the response—and, if so, whether it is a simple dependence on one of the factors alone; or whether it is an additive dependence on both factors, so that a change in one factor shifts the response level by the same amount whatever the level of the other factor; or whether it is an interactive dependence on both factors with a response level special to each treatment \(A_iB_j\). As before, we assume that the response is presented in such a form that the variation has the same scaling at the various possible response levels. And we consider the analysis for the case of approximately normal variation. The analysis for other distribution forms or with a parameter \(\lambda\) can be investigated using the
theory in Section 11.3; and such analysis will typically need actual projections rather than just the squared lengths as recorded in the ordinary analysis-of-variance table.

If the factors do not affect the response, then the model for the response is

\[ y = \mu 1 + \sigma z, \]

where \( z \) is a sample of \( N \) from the standard normal. The fitted general-level vector and the corresponding squared length are

\[ p_1 = \bar{y} 1, \]
\[ SS_1 = SS_{(1)} = N \bar{y}^2 = T_{..}/N, \]

where \( T_{..} = \sum_{i} \sum_{js} y_{ij} \) is the overall total and \( \bar{y} \) is the overall average.

If only the factor \( A \) affects the response, the model is

\[ y = \sum \mu_i 1_i + \sigma z, \]

where \( 1_i \) is a vector with 1's in positions corresponding to the level \( A_i \) and 0's elsewhere; note that \( \Sigma 1_i = 1 \). The fitted general-level vector and the corresponding squared length are

\[ p_{(2)} = \sum \bar{y}_i 1_i, \]
\[ SS_{(2)} = bn \sum \bar{y}_i^2 = T_{..}/bn, \]

where \( T_{i..} = \sum_{j} \sum_{js} y_{ij} \) is the total for the level \( A_i \) and \( \bar{y}_i = T_{i..}/bn \).

The projection into the space describing differences between the levels of the factor \( A \), and the corresponding squared length are obtained by differencing

\[ p_2 = \sum (\bar{y}_i - \bar{y}) 1_i, \]
\[ SS_2 = bn \sum (\bar{y}_i - \bar{y})^2 = SS_{(2)} - SS_{(1)} = \sum_i T_{i..}/bn - T_{..}/N. \]

Similarly, if only the factor \( B \) affects the response, the model is

\[ y = \sum \nu_j 1_{.j} + \sigma z, \]

where \( 1_{.j} \) is a vector with 1's in positions corresponding to the level \( B_j \) and 0's elsewhere; note that \( \Sigma 1_{.j} = 1 \). The fitted general-level vector and the corresponding squared length are

\[ p_{(2)} = \sum \bar{y}_j 1_{.j}, \]
\[ SS_{(2)} = an \sum \bar{y}_j^2 = \sum_j T_{.j}/an, \]

where \( T_{.j} = \sum_{i} \sum_{js} y_{ij} \) is the total for the level \( B_j \) and \( \bar{y}_j = T_{.j}/an \) is the average for that level.
The projection into the space describing differences between the levels of the factor $B$ and the corresponding squared length are obtained by differencing.

$$p_x = \sum (\bar{y}_j - \bar{y})1_{.,j}$$

$$SS_x = an \sum (\bar{y}_j - \bar{y})^2 = SS_{(x)} = SS_{(i)} = \sum_i T_i^2 / an - T^2 / N.$$  

The component $\Sigma (\bar{y}_j - \bar{y})1_{i,j}$ describes $A$-differences is the projection of $\bf{y}$ into the subspace $\mathcal{L}(1_{i}, \ldots, 1_{a}, \bot 1)$. The component describing $B$-differences is the projection of $\bf{y}$ into the subspace $\mathcal{L}(1_{.,1}, \ldots, 1_{.,b}, \bot 1)$. These subspaces are orthogonal and thus the $A$-difference component is orthogonal to the $B$-difference component. This is easily seen from an example:

\[
\begin{pmatrix}
1 & 1 & 1 \\
-1 & -1 & -1 \\
\end{pmatrix}
\begin{pmatrix}
2 & -1 & -1 \\
2 & -1 & -1 \\
\end{pmatrix}.
\]

And it is easily verified generally: note that the inner product of $\bf{1}_i$ and $\bf{1}_j$ is $n$; thus the inner product of $\Sigma l_i 1_{i}$ and $\Sigma m_j 1_{.,j}$, where $\Sigma l_i = 0 = \Sigma m_j$ is $\Sigma l_i m_j n = 0 \cdot 0 \cdot n = 0$.

If the factors $A$ and $B$ can interact in their effect on the response, then the model is

$$\bf{y} = \sum \mu_{ij} 1_{ij} + \sigma \bf{z},$$

where $1_{ij}$ is a vector with 1's in positions corresponding to $A_i B_j$ and 0's elsewhere. The vectors $1_{ij}$ are orthogonal; hence the vectors $\Sigma \mu_{ij} 1_{ij}$ form an $ab$-dimensional subspace. The fitted general-level vector and the corresponding squared length

$$p_{(3)} = \sum \bar{y}_{ij} 1_{ij},$$

$$SS_{(3)} = n \sum \bar{y}_{ij}^2 = \sum_{ij} T_{ij}^2 / n,$$

where $T_{ij} = \sum \gamma_{ij}$ is the total for the treatment $A_i B_j$ and $\bar{y}_{ij} = T_{ij} / n$ is the average value for that treatment.

Note the following vector relations:

$$\sum_i 1_{ij} = 1_{.,j}, \quad \sum_j 1_{ij} = 1_{i,.}, \quad \sum_{ij} 1_{ij} = 1.$$  

Hence the subspaces for the mean, for the $A$ effects, and for the $B$ effects are component spaces of the $ab$-dimensional space $\mathcal{L}(1_{ij}: all \ i, j)$. Now consider the $AB$ interaction, the $AB$ effects beyond that for the mean, for $A$ levels, and for $B$ levels. The projection into the space describing the $AB$ interaction and the corresponding squared length can be obtained by differencing.

$$p_3 = p_{(3)} - p_2 - p_x - p_1 = \sum (\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}) 1_{ij},$$

$$SS_3 = n \sum (\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y})^2$$
\[
= \sum_{ij} T_{ij}^2/n - \sum_i T_i^2/n - \sum_j T_j^2/an + T_{..}^2/N.
\]

Note the algebraic identity obtained from the two ways of deriving the sum of squares.

The analysis-of-variance table (Table 12.10) summarizes some of the preceding results. Note that the SS entries are best computed by differencing, as in the second expressions in (5), (8), and (11). MS entries are calculated as SS/DF.

### Table 12.10

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Projection</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Mean</td>
<td>1</td>
<td>(\bar{y}^1)</td>
<td>(N\bar{y}^2)</td>
</tr>
<tr>
<td>2 A</td>
<td>(a - 1)</td>
<td>(\sum (\bar{y}<em>{i} - \bar{y})^1</em>{i})</td>
<td>(bn \sum (\bar{y}_{i} - \bar{y})^2)</td>
</tr>
<tr>
<td>2' B</td>
<td>(b - 1)</td>
<td>(\sum (\bar{y}<em>{.j} - \bar{y})^1</em>{.j})</td>
<td>(an \sum (\bar{y}_{.j} - \bar{y})^2)</td>
</tr>
<tr>
<td>3 (A \times B)</td>
<td>((a - 1)(b - 1))</td>
<td>(\sum (\bar{y}<em>{ij} - \bar{y}</em>{i} - \bar{y}<em>{.j} + \bar{y})^1</em>{ij})</td>
<td>(n \sum (\bar{y}<em>{ij} - \bar{y}</em>{i} - \bar{y}_{.j} + \bar{y})^2)</td>
</tr>
<tr>
<td>4 Residual</td>
<td>(ab(n - 1))</td>
<td>(y - \sum \bar{y}<em>{ij}^1</em>{ij})</td>
<td>(\sum (y_{ij} - \bar{y}_{ij})^2)</td>
</tr>
</tbody>
</table>

\(abn = N\)

### B THE ANALYSIS

We consider the analysis for the case of normal variation. The analysis for other distribution forms is based on the theory in Section 11.3. We assume that the response \(y\) has variation \(\sigma^2\), where \(z\) is standard normal.

The distribution theory is available from Section 11.3F and follows the pattern discussed in Section 12.1C. The distribution of an SS entry is that of \(\sigma^2 \chi^2\), where \(\chi^2\) has the "chi-square distribution" with degrees of freedom given by the corresponding DF entry; the chi-square distribution is ordinary chi-square if the true mean vector has zero projection in the particular subspace, and it is noncentral chi-square with values that tend to be larger if the mean vector has a nonzero projection.

Now consider the problem of testing for dependence on the factors \(A\) and \(B\). The first test for dependence is made by calculating \(F = MS_A/MS_B\) and comparing the value with the appropriate \(F\) distribution. Significance indicates dependence on the two factors; further analysis would be concerned with the form of the dependence on the two factors.

If the preceding test gives no significance, then an assumption that there is no interaction could be entertained; more detailed investigations would, of course, be needed to support such an assumption. Under such an assumption we can test separately for \(A\) effects and for \(B\) effects by calculating the \(F\) ratios.
\[ F = \frac{MS_2}{MS_4}, \quad F = \frac{MS_2'}{MS_4} \]

and comparing the values with the appropriate \( F \) distribution.

Some of the arithmetic can be illustrated by a very simple numerical example. The example is so simple, however, that we have no degrees of freedom left for estimating the error variance and making appropriate tests; but earlier examples illustrate the procedure for making the tests.

**EXAMPLE 1**

Consider a response observed at three levels for a factor \( A \) and two levels for a factor \( B \). The six observations are

<table>
<thead>
<tr>
<th></th>
<th>( B_1 )</th>
<th>( B_2 )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A_1 )</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>( A_2 )</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>( A_3 )</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>9</strong></td>
<td><strong>3</strong></td>
<td><strong>12</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>30</strong></td>
</tr>
</tbody>
</table>

Together with some calculated numbers: \( 12^2/6 = 24; (6^2 + 2^2 + 4^2)/2 = 28; \) and \( (9^2 + 3^2)/3 = 30 \). The sum of squares of the observations is 38.

The projections for the various models can be calculated from the array, row, column, and cell averages; the sums of squares are available from the numbers just recorded. It is convenient to arrange these in a partial ordering corresponding to relative generality:

\[
\begin{pmatrix}
2 & 2 \\
2 & 2 \\
2 & 2
\end{pmatrix}
\begin{pmatrix}
3 & 3 \\
1 & 1 \\
2 & 2
\end{pmatrix}
\begin{pmatrix}
5 & 1 \\
1 & 1 \\
3 & 1
\end{pmatrix}
\]

\[
\begin{pmatrix}
24 \\
28 \\
38
\end{pmatrix}
\begin{pmatrix}
30 \\
38
\end{pmatrix}
\]

By differencing we then obtain the projection and squared lengths corresponding to orthogonal subspaces:
The analysis-of-variance table is available from the preliminary calculations (Table 12.11).

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Mean</td>
<td>1</td>
<td>24</td>
</tr>
<tr>
<td>2 A</td>
<td>2</td>
<td>4 = 28 − 24</td>
</tr>
<tr>
<td>2' B</td>
<td>1</td>
<td>6 = 30 − 24</td>
</tr>
<tr>
<td>3 A × B</td>
<td>2</td>
<td>4 = 38 − preceding</td>
</tr>
<tr>
<td>4 Residual</td>
<td>—</td>
<td>38</td>
</tr>
</tbody>
</table>

**C FORM OF THE DEPENDENCE ON FACTORS A AND B**

Dependence on the factors A and B can be investigated in part by separating out individual dimensions in the analysis-of-variance table.

Differences between various averages can be used as a basis for individual dimensions. For an example, suppose that the factor A has three levels. There are then two dimensions for A-differences. One dimension could be used say to measure how the average for A2 and A3 exceeds the value at A1 and a second dimension (orthogonal) could examine the difference between A2 and A3. Such a separation is based on the orthogonal vectors (−1, 1/2, 1/2) and (0, −1, 1); of course, for the full design each coordinate as given would appear bn times. The resulting components of the sum of squares for the factor A would be

\[
\left( -\sum_{ja} y_{1ja} + (1/2) \sum_{ja} y_{2ja} + (1/2) \sum_{ja} y_{3ja} \right)^2 \over bn + (1/4)bn + (1/4)bn
\]
\[
\frac{(- \sum_{js} y_{2js} + \sum_{js} y_{3js})^2}{bn + bn}.
\]

And suppose that the factor \( B \) has two levels. There is then one dimension for \( B \) differences, and the vector \((-1, +1)\) for the two levels generates the subspace; of course, for the full design each coordinate as given would appear an \( n \) times. The resulting component of the sum of squares for the factor \( B \) would be

\[
\frac{(- \sum_{is} y'_{i1s} + \sum_{is} y'_{i2s})^2}{an + an}.
\]

The vectors for the factor \( A \) and the vectors for the factor \( B \) can be combined to give multipliers for the interaction:

\[
\begin{align*}
(12) & \quad \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} (-1, 1) = \begin{pmatrix} 1 \\ \frac{1}{2} \end{pmatrix} = (a_{ij}), \\
(13) & \quad \begin{pmatrix} 0 \\ \frac{1}{2} \end{pmatrix} (-1, 1) = \begin{pmatrix} 0 \\ \frac{1}{2} \end{pmatrix} = (b_{ij});
\end{align*}
\]

of course, for the full design each coordinate in the arrays at right would appear \( n \) times. The resulting components of the sum of squares for the \( AB \) interaction would be

\[
\frac{\left( \sum_{js} a_{ij} y_{ijs} \right)^2}{n \sum a_{ij}^2},
\]

\[
\frac{\left( \sum_{js} b_{ij} y_{ijs} \right)^2}{n \sum b_{ij}^2}.
\]

If an input variable is available for the factor \( A \), then it and various functions of it can be used to separate out individual dimensions following the pattern in Section 12.1D. Similarly, if an input variable is available for the factor \( B \), then in a similar way various individual dimensions can be separated out again in the pattern of Section 12.1D.

Now consider the case of possible interaction and suppose that input variables \( x_1 \) and \( x_2 \) are available for the factors \( A \) and \( B \). Various functions of the \( x \)'s could provide components for structuring the response level. We illustrate the methods here by continuing with the linear and quadratic dependencies examined in Section 12.1. A quadratic expression in \( x_1 \) and \( x_2 \) has the form

\[ a + a_1 x_1 + a_{11} x_1^2 + a_2 x_2 + a_{22} x_2^2 + a_{12} x_1 x_2. \]

For \( x_1 \) we can follow the pattern in Section 12.1 and use the orthogonal functions

\[ x_1 - \bar{x}, \quad x_1^2 - c_1 - d_1 x_1, \]

where
\[ \sum (x_{1i}^2 - c_1 - d_1 x_{1i}) = 0, \]
\[ \sum (x_{2i}^2 - c_2 - d_2 x_{2i}) = 0, \]

and \( \bar{x}_1 = \sum x_{1i} / \alpha \). These generate two dimensions corresponding to \( A \)-differences. And similarly for \( x_2 \), we can use the orthogonal functions
\[ x_2 - \bar{x}_2, \quad x_2^2 - c_2 - d_2 x_2, \]

where
\[ \sum (x_{2ij}^2 - c_2 - d_2 x_{2ij}) = 0, \]
\[ \sum (x_{2ij}^2 - c_2 - d_2 x_{2ij}) x_{2ij} = 0. \]

### Table 12.12
Analysis-of-variance table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1</td>
<td>( \frac{\left( \sum y_{1i} \right)^2}{N} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( \frac{\left( \sum (x_{1i} - \bar{x}<em>1) y</em>{1i} \right)^2}{bn \sum (x_{1i} - \bar{x}_1)^2} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( \frac{\left( \sum (x_{2i}^2 - c_2 - d_2 x_{2i}) y_{1i} \right)^2}{bn \sum (x_{2i}^2 - c_2 - d_2 x_{2i})^2} )</td>
</tr>
<tr>
<td>A Linear</td>
<td>1</td>
<td>By differencing</td>
</tr>
<tr>
<td>Quadratic</td>
<td>1</td>
<td>( \frac{\left( \sum (x_{2i} - \bar{x}<em>2) y</em>{1i} \right)^2}{an \sum (x_{2i} - \bar{x}_2)^2} )</td>
</tr>
<tr>
<td>Other</td>
<td>( a - 3 )</td>
<td>By differencing</td>
</tr>
<tr>
<td>B Linear</td>
<td>1</td>
<td>( \frac{\left( \sum (x_{2ij}^2 - c_2 - d_2 x_{2ij}) y_{1i} \right)^2}{an \sum (x_{2ij}^2 - c_2 - d_2 x_{2ij})^2} )</td>
</tr>
<tr>
<td>Quadratic</td>
<td>1</td>
<td>By differencing</td>
</tr>
<tr>
<td>Other</td>
<td>( b - 3 )</td>
<td>By differencing</td>
</tr>
<tr>
<td>A \times B</td>
<td>Quadratic</td>
<td>( \frac{\sum (x_{1i} - \bar{x}<em>1)(x</em>{2j} - \bar{x}<em>2) y</em>{1i}}{n \sum (x_{1i} - \bar{x}<em>1)^2 \sum (x</em>{2j} - \bar{x}_2)^2} )</td>
</tr>
<tr>
<td>Other</td>
<td>( (a - 1)(b - 1) - 1 )</td>
<td>By differencing</td>
</tr>
<tr>
<td>Residual</td>
<td>( ab(n - 1) )</td>
<td>( \frac{\sum (y_{1i} - \bar{y}<em>i)^2}{\sum y</em>{1i}^2} )</td>
</tr>
</tbody>
</table>
and $\bar{x}_2 = \Sigma x_{2j}/b$. These generate two dimensions corresponding to the $B$-differences. An orthogonal component for $A$-differences can be combined with an orthogonal component for $B$-differences to give an orthogonal component for the $AB$ interaction; thus in the at-most-quadratic case we obtain the function

$$(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)$$

as with the formulas (12) and (13).

The analysis-of-variance table then takes the form shown in Table 12.12. Some of the entries are available by differencing with respect to the earlier table. The various projections can be given explicitly by analogy from the sum-of-squares entries. For example, the projection onto $\xi(\Sigma (x_{1i} - \bar{x}_1)1_{i1})$ is

$$\frac{\sum (x_{1i} - \bar{x}_1)y_{ija}}{bn \sum (x_{1i} - \bar{x}_1)^2} \cdot \sum (x_{1i} - \bar{x}_1)1_{i1}.$$ 

Tests for dependence beyond the quadratic can be made by comparing the entries marked "other" against the residual.

On the assumption that dependence is at most quadratic, tests can be made by comparing the various quadratic and linear entries against the residual following the general pattern described in Section 12.1.

**D Exercises**

1. Twelve response observations were made randomly in time, two at each combination of three levels of a first factor temperature and two levels of a second factor pressure; the variation is known to be approximately normal. The data are as follows:

<table>
<thead>
<tr>
<th>Temperature, $x_1$</th>
<th>Pressure, $x_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$-1$</td>
</tr>
<tr>
<td>73</td>
<td>14.2</td>
</tr>
<tr>
<td></td>
<td>14.3</td>
</tr>
<tr>
<td>74</td>
<td>17.4</td>
</tr>
<tr>
<td></td>
<td>17.6</td>
</tr>
<tr>
<td>75</td>
<td>20.7</td>
</tr>
<tr>
<td></td>
<td>20.4</td>
</tr>
</tbody>
</table>

Construct the analysis-of-variance table separating Mean, $T$, $P$, $T \times P$, and Residual. Test for dependence on the factors.

2. (continuation) Separate the dependence on factor $T$ into Linear and Other and make appropriate tests.

3. The effect of plate temperature $T$ and filament lighting $L$ on conductance was investigated for a certain kind of tube. Three observations were made at each combination of two levels for temperature and four levels for filament lighting:
Make appropriate tests for dependence on the factors $T$ and $L$.

The effects of fertilizer $F$ and chemical seed treatment $C$ on growth of wheat were investigated on 48 plots of land; three determinations were made at each combination of four levels for $F$ and four levels for $C$:

<table>
<thead>
<tr>
<th></th>
<th>$C_1$</th>
<th>$C_2$</th>
<th>$C_3$</th>
<th>$C_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$</td>
<td>21.4</td>
<td>20.9</td>
<td>19.6</td>
<td>17.6</td>
</tr>
<tr>
<td></td>
<td>21.2</td>
<td>20.3</td>
<td>18.8</td>
<td>16.6</td>
</tr>
<tr>
<td></td>
<td>20.1</td>
<td>19.8</td>
<td>16.6</td>
<td>17.5</td>
</tr>
<tr>
<td>$F_2$</td>
<td>12.0</td>
<td>13.6</td>
<td>13.0</td>
<td>13.3</td>
</tr>
<tr>
<td></td>
<td>14.2</td>
<td>13.3</td>
<td>13.7</td>
<td>14.0</td>
</tr>
<tr>
<td></td>
<td>12.1</td>
<td>11.6</td>
<td>12.0</td>
<td>13.9</td>
</tr>
<tr>
<td>$F_3$</td>
<td>13.5</td>
<td>14.0</td>
<td>12.7</td>
<td>12.4</td>
</tr>
<tr>
<td></td>
<td>11.9</td>
<td>15.6</td>
<td>12.9</td>
<td>13.7</td>
</tr>
<tr>
<td></td>
<td>13.4</td>
<td>13.8</td>
<td>13.1</td>
<td>13.0</td>
</tr>
<tr>
<td>$F_4$</td>
<td>12.8</td>
<td>14.1</td>
<td>14.2</td>
<td>12.0</td>
</tr>
<tr>
<td></td>
<td>13.8</td>
<td>13.2</td>
<td>13.6</td>
<td>14.6</td>
</tr>
<tr>
<td></td>
<td>13.7</td>
<td>15.3</td>
<td>13.3</td>
<td>14.0</td>
</tr>
</tbody>
</table>

Make appropriate tests for dependence on the factors $F$ and $C$.

SUPPLEMENTARY MATERIAL

Supplement Sections 12.3, 12.4, and 12.5 are on pages 584–598.

The one- and two-factor designs in preceding sections are the basic designs for experimentation. The supplementary material examines the extension to the three-factor design and by implication the general multifactor design. It also examines how information concerning variability from experimental unit to experimental unit can be used to increase the sensitivity of the experimental design: specifically, randomized blocks and the analysis of covariance.

NOTES AND REFERENCES

The emphasis in this chapter has been on principles of experimentation and on certain principles in the related mathematical analyses. The author believes that an introductory course on statistics should include some contact with these principles.

A full appreciation of the importance and substance of experimental design, however, requires close contact with the design process in various applications. For some additional readings, see the following:


1.5

PROBABILITY

We have developed the basic mathematics for handling probabilities. A small amount of additional development gives some very useful results.

A SIMPLE FUNCTIONS AND PROBABILITY FORMULAS

Indicator functions can be very useful. In Section 1.1 we noted that the long-run average of an indicator function was the corresponding probability. In Section 1.2 we proved the relation

\[ I_{A \cup B} = I_A + I_B - I_{AB}, \]

which corresponds closely to the relation

\[ P(A \cup B) = P(A) + P(B) - P(AB) \]

in Section 1.4. In this section we generalize the notion of an indicator function and establish more generally the correspondence just noted.

Consider a sample space \( S \) and a class \( \mathcal{G} \) of subsets that form a \( \sigma \)-algebra.

DEFINITION 1

A real-valued function \( Y \) on \( S \) is a simple function if it takes only a finite number of values.

For an example, see Figure S.1.

Let \( a_1, \ldots, a_m \) be the different values for the function \( Y \), and let \( A_1, \ldots, A_m \) be the corresponding sets on which these values are taken:

\[ A_j = \{ s : Y(s) = a_j \} = Y^{-1}(a_j). \]

We assume, of course, that the sets \( A_i \) belong to \( \mathcal{G} \). The different sets \( A_1, \ldots, A_m \) are clearly disjoint and they cover \( S \); they thus form a partition of \( S \). The simple function \( Y \) can then be written in the canonical form,

\[ Y = a_1 I_{A_1} + \cdots + a_m I_{A_m}; \]

note that at any point \( s \) at most one term on the right is different from zero and the \( a \)-value of that term is the value of the function.
FIGURE S.1
Simple function $Y$ takes the values $a_1$, $a_2$, and $a_3$, respectively, on $A_1$, $A_2$, and $A_3$.

**EXAMPLE 1**

Consider the rolling of a symmetric die and suppose that various characteristics are observed including the number of points. Let $S$ be the sample space and $\Omega$ the class of events; and let $Y$ be the function on $S$ that gives the number of points.

The function $Y$ is a simple function that takes only the values $1, 2, \ldots, 6$; it can be put in the canonical form

$$Y = 1/\Omega_1 + 2/\Omega_2 + 3/\Omega_3 + 4/\Omega_4 + 5/\Omega_5 + 6/\Omega_6,$$

where $\Omega_j$ is the event "the number of points is $j$." Of course, $Y$ can be expressed in other ways, for example,

$$Y = \Omega_{1(1)} + \cdots + \Omega_{6(6)},$$

where $\Omega_{j(1)}$ is the event "the number of points is $j$ or more." Note that if the sample space $S$ is $\{1, 2, \ldots, 6\}$ recording just the number of points, then $Y$ is the identity function on $S$.

Now suppose that we have found someone who will give us $1 \text{ for} \text{ every dot that appears}$

when the die is rolled. If this game is played a large number of times, say $N$, then $Y = 1$ will

occur approximately $N/6 \text{ times}$, and $Y = 6$ will occur approximately $N/6 \text{ times}$. The

total amount received will be approximately

$$1 \cdot (1/6)N + 2 \cdot (1/6)N + \cdots + 6 \cdot (1/6)N$$

dollars and the average received per play will be

$$1 \cdot (1/6) + 2 \cdot (1/6) + \cdots + 6 \cdot (1/6) = 3.50$$

dollars. A fair price to play this game would then be $3.50. Note that the preceding average

is obtained by multiplying values by corresponding probabilities and summing. The same

rule also works with the second expression for $Y$:

$$1 \cdot (6/6) + 1 \cdot (5/6) + \cdots + 1 \cdot (1/6) = 3.50.$$

Now consider in general this average value, or mean value, or "expected" value. Let $P$ be a

probability measure over the sample space $S$; then

**DEFINITION 2**

The **mean value** of a simple function $Y$ having canonical form

$$Y = a_1/\Omega_1 + \cdots + a_m/\Omega_m$$

is

$$\mu(Y) = a_1P(\Omega_1) + \cdots + a_mP(\Omega_m).$$
is given by

\[ E(Y) = a_1 P(A_1) + \cdots + a_m P(A_m). \]

For a random system with variation described by \( P \), the mean \( E(Y) \) represents the long-run average of the values of the function \( Y \); compare with the example. This capital \( E \) is often called expectation.

Formula (2) for \( E(Y) \) works whether or not \( Y \) is expressed in canonical form:

**PROPOSITION 3**

If a simple function \( Y \) is expressed as

\[ Y = \sum_{i=1}^{k} b_i I_{A_i}, \]

then

\[ E(Y) = \sum_{i=1}^{k} b_i P(B_i). \]

The proof is deferred until Section 1.

Consider the constant function taking the value 1; then trivially from (2) we have that \( E(1) = 1 \). And, similarly, for a constant \( k \) we have \( E(k) = k \).

The usefulness of \( E \) is largely based on the following linearity property:

**THEOREM 4**

If \( Y_1 \) and \( Y_2 \) are simple functions, then the mean of a linear combination is the linear combination of the means:

\[ E(k_1 Y_1 + k_2 Y_2) = k_1 E(Y_1) + k_2 E(Y_2). \]

**Proof** The proof is trivial using Proposition 3. Let

\[ Y_1 = \sum_{i} a_i I_{A_i}, \quad Y_2 = \sum_{i} b_i I_{A_i}, \]

be representations for \( Y_1 \) and \( Y_2 \) in terms of simple functions. Then

\[
E(k_1 Y_1 + k_2 Y_2) = E \left( \sum_{i} k_1 a_i I_{A_i} + \sum_{i} k_2 b_i I_{A_i} \right) \\
= k_1 \sum_{i} a_i P(A_i) + k_2 \sum_{i} b_i P(B_i) \\
= k_1 E(Y_1) + k_2 E(Y_2).
\]

As a simple illustration, consider the relation

\[ I_{A \cup B} = I_A + I_B - I_{AB} \]

from Section 1.2E. Applying \( E \) to both sides gives immediately the relation \(1.4.3\).

\[ P(A \cup B) = P(A) + P(B) - P(AB) \]

in Section 1.4.

More generally, we have

\[
I_{A \cup B \cup C} = 1 - (1 - I_A)(1 - I_B)(1 - I_C) \\
= I_A + I_B + I_C - I_{AB} - I_{BC} - I_{AC} + I_{ABC}.
\]

Applying \( E \) to both sides gives the relation \(1.4.4\).
\[ P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1A_2) - P(A_2A_3) - P(A_1A_3) + P(A_1A_2A_3) \]

which was recorded in Section 1.4.

B PROBLEMS: PROBABILITY FORMULAS

The problems here use properties of the combinatorial function

\[
\binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{n(n-1) \cdots (n-r+1)}{r!},
\]

called \( n \text{ choose } r \). These will not be used in the main body of the text until Section 3.4.

1. Let \( A_1, \ldots, A_m \) be subsets of the sample space \( S \) and \( P \) be a probability measure over \( S \). Use Problem 1.2.22 to prove that

\[ P(\bigcup_i A_i) = S_1 - S_2 + S_3 - \cdots + (-1)^{m-1}S_m, \]

where \( S_r = \sum_{i_1 < \cdots < i_r} P(A_{i_1} \cdots A_{i_r}) \).

2. (continuation) Verify that the indicator function for \( E_r \) "exactly \( r \) of the events \( A_1, \ldots, A_m \)" can be expressed as

\[ I_{E_r} = \sum_{i_1 < \cdots < i_r} I_{i_1} \cdots I_{i_r}(1 - I_{i_{r+1}}) \cdots (1 - I_{i_m}), \]

where the summation is over those permutations \( (i_1, \ldots, i_m) \) of \( (1, \ldots, m) \) that satisfy the constraints indicated. Hence prove that

\[ I_{E_r} = T_r - \binom{r+1}{r} T_{r+1} + \binom{r+2}{r} T_{r+2} - \cdots + (-1)^{m-r}\binom{m}{r} T_m. \]

where

\[ T_r = \sum_{i_1 < \cdots < i_r} I_{i_1} \cdots I_{i_r}. \]

3. (continuation) Prove the following generalization of formula (5):

\[ P(E_r) = S_r - \binom{r+1}{r} S_{r+1} + \binom{r+2}{r} S_{r+2} - \cdots + (-1)^{m-r}\binom{m}{r} S_m \]

\[ = S_r - (r + 1)S_{r+1} + \frac{(r+2)(r+1)}{2} S_{r+2} - \cdots; \]

see Problem 1.4.14 and note that \( S_0 = 1 \).

4. A deck contains \( N \) cards numbered 1, 2, \ldots, \( N \). The cards are thoroughly shuffled and dealt successively as the dealer calls 1, 2, \ldots, \( N \). Let \( E_r \) be the event "exactly \( r \) called correctly." Use Problem 3 to show that

\[ P(E_r) = \frac{1}{r!} \left( 1 - 1 + \frac{1}{2!} - \frac{1}{3!} \cdots \pm \frac{1}{(N-r)!} \right) \]

\[ \approx \frac{1}{r!} e^{-1}; \]

see Problem 1.4.16. Calculate an approximate value for the probability of no cards called correctly. Note that for medium to large \( N \) the number of cards called correctly is approximately Poisson(\( \lambda \)) with \( \lambda = 1 \) (see Example 1.3.6).
5. (continuation of Problem 3) Let $E_{[r]}$ be the event "at least $r$ of $A_1, \ldots, A_m$." Recall Problem 1.4.15 and prove that

\[
P(E_{[r]}) = S_r - \binom{r}{r-1} S_{r+1} + \binom{r+1}{r-1} S_{r+2} - \cdots (-1)^{m-1} \binom{m-1}{r-1} S_m
\]

\[
= S_r - rS_{r+1} + \frac{(r+1)r}{2} S_{r+2} - + \cdots;
\]

use the combinatorial result in Problem 3.4.9.

C PROBLEMS: SIMPLE CLASSES OF SETS

As we have noted before, it is a long step from the intervals on $R$ to the Borel sets on $R$. Something slightly more general than intervals is useful for studying probability measures. Let $\mathcal{R}$ be a nonempty class of subsets of a sample space $\mathbb{S}$. Then we define the following properties (note that $A_1 - A_2 = A_1 A_2$):

Closure under difference: If $A_1, A_2 \in \mathcal{R}$, then $A_1 - A_2 \in \mathcal{R}$.

Closure under union: If $A_1, A_2 \in \mathcal{R}$, then $A_1 \cup A_2 \in \mathcal{R}$.

Closure under $\sigma$-union: If $A_i \in \mathcal{R}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{R}$.

6. $\mathcal{R}$ is a ring (set-theoretic) if it is closed under differences and unions. Show that an equivalent definition is closure under symmetric difference and union; the symmetric difference $A_1 \Delta A_2$ is defined to be $(A_1 - A_2) \cup (A_2 - A_1)$.

7. $\mathcal{R}$ is an algebra if it is closed under union and complementation. Show that an equivalent definition is a ring containing $\mathbb{S}$.

8. $\mathcal{R}$ is a $\sigma$-ring if it is closed under difference and $\sigma$-union. Show that a $\sigma$-algebra can be defined as a $\sigma$-ring containing $\mathbb{S}$.

9. Let $\mathcal{R}$ be the class of finite unions of disjoint left open–right closed intervals $(a, b]$ on $R$; that is,

\[
\mathcal{R} = \{ \bigcup_{i=1}^{n} (a_i, b_i) : a_1 < b_1 \leq a_2 < \cdots \leq a_n < b_n \}.
\]

Show that $\mathcal{R}$ is a ring. Some tedious but routine extensions of the proof establish the following: If $\mathcal{R}$ is the class of finite unions of disjoint rectangles $(a_1, b_1] \times \cdots \times (a_k, b_k]$ on $R^k$, then $\mathcal{R}$ is a ring.

10. Let $\mathcal{A}_0 = \{A\}$ be a class of subsets $A$ on a sample space $\mathbb{S}$. Show that there exists a $\sigma$-ring on $\mathbb{S}$ that contains $\mathcal{A}_0$.

11. (continuation) Let $K = \{\mathcal{A}_a\}$ be the set of all $\sigma$-rings $\mathcal{A}_a$ such that $\mathcal{A}_a \subseteq \mathcal{A}_a$ and $a$ ranges over an index set. Show that $\mathcal{A} = \cap \mathcal{A}_a$ is a $\sigma$-ring on $\mathbb{S}$.

12. (continuation) If $\mathcal{B}_a \subset \mathcal{B}_b$, we say that $\mathcal{B}_a$ is smaller than $\mathcal{B}_b$. Show that it is meaningful to speak of the smallest $\sigma$-ring containing a class $\mathcal{A}_0$ of subsets of $\mathbb{S}$; let $\mathcal{S}(\mathcal{A}_0)$ designate this $\sigma$-ring.

D THE EXTENSION THEOREM

It is easy to define the length or some other measure of an interval. In this section we quote an important extension theorem that then says a measure is uniquely determined for the logical consequences of intervals—the Borel sets.

Consider a sample space $\mathbb{S}$ and a ring $\mathcal{R}$ of subsets of $\mathbb{S}$; a ring is defined in Problem 6. And suppose that we have a measure $\mu$ defined on the ring $\mathcal{R}$; a measure on a ring is defined as before: $\sigma$-additive (for countable unions that belong to $\mathcal{R}$) and nontrivial. A measure $\mu$ on $\mathcal{R}$ is called $\sigma$-finite if each $A$ in $\mathcal{R}$ can be expressed as $\bigcup_{i=1}^{\infty} A_i$, where each $\mu(A_i) < \infty$. 


THEOREM 5

If \( \mu \) is a measure on a ring \( \mathcal{R} \) and if \( \mu \) is \( \sigma \)-finite, then there is a unique extension of \( \mu \) to the smallest \( \sigma \)-ring \( S(\mathcal{R}) \) containing \( \mathcal{R} \).

The proof is not simple and may be found in advanced books on measure theory.

E MEASURING LENGTH, AREA, AND VOLUME

In Section 1.3A we considered the real line \( R \) as a uniform rod having unit weight for unit length. The weight measure of a set \( A \) as a piece of the rod then gives the length measure of the set \( A \). Consider this more formally.

On the real line \( R \) we define the length of an interval \([a, b]\) by

\[ \mu([a, b]) = b - a. \]

We could consider open intervals or closed intervals, but the left-open–right-closed intervals have some advantages; see Problem 9. In the same way we define the length of \( \bigcup_{i=1}^{k} [a_i, b_i] \), where the components are disjoint, by

\[ \mu(\bigcup_{i=1}^{k} [a_i, b_i]) = \sum_{i=1}^{k} (b_i - a_i). \]

The class of finite unions of left-closed–right-open intervals forms a ring \( \mathcal{R} \); Problem 9. The Extension Theorem in Section D then says there is a unique extension of \( \mu \) to the smallest \( \sigma \)-ring \( S(\mathcal{R}) \) containing \( \mathcal{R} \). This \( \sigma \)-ring is, however, the \( \sigma \)-algebra of Borel sets. Thus length as defined for intervals extends uniquely to a length measure for all Borel sets on \( R \). This is called Lebesgue measure on the line \( R \).

Now consider the plane \( R^2 \), or, more generally, the Cartesian product \( R^k \). We define the volume of a rectangle \([a_1, b_1] \times \cdots \times [a_k, b_k] \) by

\[ \mu\left( [a_1, b_1] \times \cdots \times [a_k, b_k] \right) = \prod_{i=1}^{k} (b_i - a_i). \]

And for a finite union of disjoint rectangles we define it to be the sum of the component volumes. The class of finite unions of left-open–right-closed rectangles forms a ring \( \mathcal{R} \); Problem 9. The Extension Theorem then gives a unique extension of \( \mu \) to the smallest \( \sigma \)-ring containing \( \mathcal{R} \); this \( \sigma \)-ring is the Borel class \( \mathcal{B}^k \) on \( R^k \). Thus volume as defined for rectangles extends uniquely to a volume measure for all Borel sets on \( R^k \). This extension is called Lebesgue measure on \( R^k \).

F PROBLEMS: LENGTH MEASURE

13 Let \( \mu \) be Lebesgue measure on \( R \).
(a) Show that \( \mu(\{x\}) = 0 \), the measure of a single point is zero.
(b) Show that \( \mu(A) = 0 \) if \( A \) is finite or countable.

14 Consider the real line \( R \) and let \( A_0, A_1, \ldots \) be a sequence of sets defined as follows:

\[ A_0 = [0, 1], \quad A_1 = [0, 1/3] \cup [2/3, 1], \quad A_2 = [0, 1/9] \cup [2/9, 1/3] \cup [2/3, 7/9] \cup [8/9, 1], \ldots, \]

where \( A_{n+1} \) is obtained from \( A_n \) by deleting the open middle third of each component interval. The intersection set \( A = \bigcap_{i=1}^{\infty} A_i \) is called the Cantor ternary set. Let \( \mu \) be Lebesgue measure.
(a) Calculate \( \mu(A_0), \mu(A_1), \text{ and } \mu(A_2) \).
(b) Calculate \( \mu(A_n) \) and deduce \( \mu(A) \).
Sec. 1.5: Probability 517

15 (continuation) Let \( x \) belong to \( A \), and hence to \( A_0, A_1, A_2, \ldots \); let \( i_1 = O(1) \) if \( x \) is in the lower (upper) interval formed in going from \( A_0 \) to \( A_1 \); let \( i_2 = O(1) \) if \( x \) is in the lower (upper) interval in going from \( A_1 \) to \( A_2, \ldots \). Thus \( x \) produces \( (i_1, i_2, \ldots) \), or \( i_1 i_2 i_3 \cdots \) as a binary-decimal sequence. Note that all binary-decimal sequences can be obtained. Show that \( A \) has the same cardinality as the interval \([0, 1]\) itself, that is, the same as the continuum.

G PROBLEMS: LIMIT SETS

16 Let \( A_1, A_2, \ldots \) be a sequence of subsets of a sample space \( S \).
(a) Let \( U \) be the set of points that belong to an infinity of the sets \( A_1, A_2, \ldots \); we write \( U = \limsup_{n \to \infty} A_n \). Prove that
\[
U = \cap_{n=1}^{\infty} \cup_{i=n}^{\infty} A_i = \cap_{n=1}^{\infty} B_n;
\]
note that \( B_1, B_2, \ldots \) is a decreasing sequence where \( B_n = \cup_{i=n}^{\infty} A_i \).

(b) Let \( I_1 \) be the indicator function for \( A_1 \). Prove that
\[
I_U = \limsup_{i \to \infty} I_i.
\]
This is a generalization of part of Problem 1.4.19. The definition of \( \limsup \) for functions is recorded in Section 3.6G.

17 Let \( A_1, A_2, \ldots \) be a sequence of subsets of a sample space \( S \).
(a) Let \( L \) be the set of points that belong to all but a finite number of the sets \( A_1, A_2, \ldots \); we write \( L = \liminf_{i \to \infty} A_i \). Prove that
\[
L = \cup_{n=1}^{\infty} \cap_{i=n}^{\infty} A_i = \cup_{n=1}^{\infty} C_n;
\]
note that \( C_1, C_2, \ldots \) is an increasing sequence, where \( C_n = \cap_{i=n}^{\infty} A_i \).

(b) Let \( I_i \) be the indicator function for \( A_i \). Prove that
\[
I_1 = \liminf_{i \to \infty} I_i.
\]
This is a generalization of part of Problem 1.4.19.

18 Borel-Cantelli Lemma: Let \( A_1, A_2, \ldots \) be a sequence of elements of a \( \sigma \)-algebra \( \mathcal{A} \). If \( \Sigma_{i=1}^{\infty} P(A_i) < \infty \), then the probability that an infinity of the \( A_i \) occur is zero. See Problem 16 and note that the event just mentioned is \( \limsup A_i \). Also note that \( \cup_{i=1}^{\infty} A_i \) is monotone decreasing, and that \( P(\cup_{i=1}^{\infty} A_i) \leq \Sigma_{i=1}^{\infty} P(A_i) \). Prove the lemma.

H PROBLEMS: LIMIT FUNCTIONS

The following problem examines in depth what it means to have a sequence of functions converge to a limit function.

19 Let \( Y, Y_1, Y_2, \ldots \) be functions mapping \( S \) into \( R \); that is, \( Y_1 : S \to R \). Let \( C_i(\epsilon) \) be the set of points for which \( Y_i \) is at least \( \epsilon \) close to \( Y \):
\[
C_i(\epsilon) = \{ s : |Y_i(s) - Y(s)| \leq \epsilon \}.
\]
(a) Let \( C_{i1}(\epsilon) \) be the set of points for which each of \( Y_i, Y_{i+1}, \ldots \) is at least \( \epsilon \) close to \( Y \). Show that
\[
C_{i1}(\epsilon) = \bigcap_{j=i}^{\infty} C_i(\epsilon) = \left\{ s : \sup_{i=1}^{\infty} |Y_i(s) - Y(s)| \leq \epsilon \right\}.
\]
Note that \( C_{i1}(\epsilon) \) is monotone increasing in \( i \).

(b) Let \( C(\epsilon) \) be the set of points for which the sequence \( Y_1, Y_2, \ldots \) eventually is \( \epsilon \) close to \( Y \). Show that
\( C(\epsilon) = \bigcup_{n=1}^{\infty} C_{\epsilon,n}(\epsilon) \).

Note that \( C(\epsilon) \) is monotone decreasing as \( \epsilon \to 0 \).

(c) Let \( C \) be the set of points for which the sequence \( Y_1, Y_2, \ldots \) converges to \( Y \). Show that

\[
C = \left\{ s : \lim_{i \to \infty} Y_i(s) = Y(s) \right\} = \bigcap_{n=1}^{\infty} C \left( \frac{1}{n} \right).
\]

(d) If all the sets \( \{ s : |Y_i(s) - Y(s)| \leq a \} \) belong to a \( \sigma \)-algebra \( \mathcal{A} \), prove that \( C \in \mathcal{A} \). For a probability application of this problem, see Section 6.3E.

I MEAN VALUE OF SIMPLE FUNCTIONS

We now prove Proposition 3, which states that the mean of a simple function can be calculated directly from any expression for the simple function. The proposition asserts that the mean of a simple function

\[
Y = \sum_{i=1}^{k} b_i 1_{B_i}
\]

can be calculated as

\[
F = \sum_{i=1}^{k} b_i P(B_i)
\]

without reduction to canonical form. The method of proof is to make a succession of changes on \( Y \) and the corresponding changes on \( F \) until \( Y \) has been put in canonical form; we then check if the expression for \( F \) is as given in Definition 2.

The sets \( B_1, \ldots, B_k \) may not be a partition of \( \mathbb{S} \). Form all the intersection sets

\[
B_{1i} \cap \ldots \cap B_{ki},
\]

where \( i_k = 0, 1, \ldots, k \). Let \( C_1, \ldots, C_N \) be the nonempty intersection sets. Then \( \{C_1, \ldots, C_N\} \) is a partition of \( \mathbb{S} \) and each \( B_i \) is a union of \( C \)'s. If \( B_i \) is replaced by \( C_1 \cup \ldots \cup C_i \), say, then

\[
\lambda_{B_i} = \sum_{i=1}^{t} I_{C_i}, \quad P(B_i) = \sum_{i=1}^{t} P(C_i).
\]

It follows then that \( Y \) and \( F \) can be put in the form

\[
Y = b_1 I_{C_1} + \ldots, \quad F = b_1 P(C_1) + \ldots.
\]

If several terms, say \( b_1 I_{C_1}, \ldots, b_w I_{C_1} \) involve the same set \( C_1 \), then they can be collected. Let

\[
b_1 I_{C_1} + \ldots + b_w I_{C_1} = c_1 I_{C_1}, \quad b_1 P(C_1) + \ldots + b_w P(C_1) = c_1 P(C_1).
\]

It follows then that \( Y \) and \( F \) can be put in the form

\[
Y = \sum c_i I_{C_i}, \quad F = \sum c_i P(C_i).
\]

If several terms have the same coefficient, say \( c_1 = \ldots = c_v \), then they can be combined. Let \( a_1 = c_1 = \ldots = c_v \) and \( A_1 = C_1 \cup \ldots \cup C_v \); then

\[
c_1 I_{C_1} + \ldots + c_v I_{C_v} = a_1 I_{A_1}, \quad c_1 P(C_1) + \ldots + c_v P(C_v) = a_1 P(A_1).
\]
It follows then that $Y$ and $F$ can be put in the form

$$Y = \sum a_i I_{A_i}, \quad F = \sum a_i P(A_i);$$

but $Y$ is now in canonical form and $F$ is the defining expression for $E(Y)$.

\section{Mean of Discrete Functions}

We defined an indicator function $I_A$ as an alternative way of representing an event; and we associated a value to the indicator function, the probability $P(A)$. We then defined a simple function, in effect, a linear combination of indicator functions; and we associated a value to it, the linear combination of the values associated with the indicator functions. We now extend further and define a \textit{discrete function} $Y$ as a real-valued function that takes a countable number of values, say $a_1, a_2, \ldots$; thus

$$Y = a_1 I_{A_1} + a_2 I_{A_2} + \cdots,$$

where $I_{A_i}$ is the indicator function for the set $A_i$ of points at which the function takes the value $a_i$; it is assumed that the $A_i$ are in the class $\mathcal{A}$. And we associate a value to the discrete function, the mean value

$$E(Y) = a_1 P(A_1) + a_2 P(A_2) + \cdots$$

provided the series converges absolutely; otherwise, we say that $E(Y)$ does not exist.

We can then prove analogs of Proposition 3 and Theorem 4. The analog of Theorem 4 is: if $Y_1$ and $Y_2$ are discrete functions and if $E(Y_1)$ and $E(Y_2)$ exist, then $E(k_1 Y_1 + k_2 Y_2) = k_1 E(Y_1) + k_2 E(Y_2)$.

\section{Problems: Mean Value; Simple and Discrete Functions}

20. If the distribution of a discrete function $Y$ is given by the Poisson($\lambda$) distribution, then show that $E(Y) = \lambda$.

21. The mean value of a simple function $Y$ has been defined as the summation of values weighted by corresponding probabilities; thus $Y = \sum_{i=1}^k a_i I_{A_i}$ with distinct values $a_1, \ldots, a_k$ has mean

$$E(Y) = \sum_{i=1}^k a_i P(A_i) = \int_Y Y(s) \, dP(s),$$

and can be called the integral of $Y$ with respect to $P$. Now let $Y$ be a simple function with nonnegative values $a_1, \ldots, a_k$. Then the integral of $Y$ over the set $E$ with respect to the measure $\mu$ is defined by

$$\int_E Y(s) \, d\mu(s) = \sum_{i=1}^k a_i \mu(A_i \cap E),$$

where $0 < \infty$ if it occurs is taken to be zero.

(a) If $\nu(E) = \int_E Y(s) \, d\mu(s)$, then prove that $\nu$ is a measure. The function $Y$ modulates the measure $\mu$, giving an adjusted measure $\nu$.

(b) Let $Y_1$ and $Y_2$ be simple functions taking nonnegative values, and $k_1, k_2 \geq 0$, then prove

$$\int_E (k_1 Y_1(s) + k_2 Y_2(s)) \, d\mu(s) = k_1 \int_E Y_1(s) \, d\mu(s) + k_2 \int_E Y_2(s) \, d\mu(s).$$

The integral is linear in the integrand.
2.5

PROBABILITY ON
THE LINE AND PLANE

In applications we frequently obtain a sequence of determinations of a response or a
sequence of values from different responses, and then want to do similar numerical calculations on
the different elements in the sequence. For such calculations, vector and matrix methods are
essential. In this section we briefly survey vector and matrix methods; some have been touched on
in the preceding section and others will become important as we progress through later chapters.
We conclude the section with some material on components of a distribution: discrete, absolutely
continuous, and singular.

A VECTOR ALGEBRA

Consider points $z = (z_1, \ldots, z_n)'$ in $R^n$. We can add points $z_1 + z_2 = (z_{11} + z_{12}, \ldots,
\ldots, z_{n1} + z_{n2})'$ by adding coordinate by coordinate; and we can multiply by a scalar $cz = (cz_1, \ldots,
cz_n)'$, where $c$ is real. With these operations $R^n$ is a vector space. Call the points vectors.
Let $a_1, \ldots, a_r$ be vectors in $R^n$. Then the set
\begin{equation}
\mathcal{L}(a_1, \ldots, a_r) = \{x_1 a_1 + \cdots + x_r a_r : x_i \in R\}
\end{equation}
of all linear combinations of the $a_i$'s is closed under addition and scalar multiplication: it is a linear
subspace of $R^n$, called the linear subspace generated by $a_1, \ldots, a_r$. It is convenient sometimes to
present the vectors in the form of the matrix
\begin{equation}
A = (a_1 \cdots a_r) = \begin{pmatrix}
a_{11} & \cdots & a_{1r} \\
\vdots & \ddots & \vdots \\
a_{n1} & \cdots & a_{nr}
\end{pmatrix}
\end{equation}
and write $\mathcal{L}(a_1, \ldots, a_r)$ or $\mathcal{L}(A)$; see Figure S.2.

The vectors $a_1, \ldots, a_r$ are called linearly dependent if there is a linear combination
$x_1 a_1 + \cdots + x_r a_r = 0$ which gives the zero vector and has at least one coefficient different from
zero. Otherwise they are called linearly independent. For example, the vectors $(1, 1, 1, 1)$ and $(1,
2, 3, 4)$ are linearly independent and the vectors $(1, 1, 1, 1), (1, 2, 3, 4),$ and $(2, 3, 4, 5)$ are
linearly dependent.

Now consider $\mathcal{L}(a_1, \ldots, a_s)$. If the vectors $a_1, \ldots, a_s$ are linearly dependent, then the linear
combination can be solved for one vector in terms of the others and $\mathcal{L}(a_1, \ldots, a_s)$ can be generated
by one less vector. This can be continued until a linearly independent set of vectors remains. Let $s$ be the
number of vectors remaining. Then $s$ is uniquely determined; it is called the dimension of
the subspace $\mathcal{L}(a_1, \ldots, a_s)$ and the corresponding set of vectors is called a basis for the subspace;
s is also called the rank of $A$ and cannot be larger than $n$ or $r$.

Now consider $n$ vectors $a_1, \ldots, a_n$ with matrix $A$ and examine the equation
\begin{equation}
z = Ax \text{ or } z = x_1 a_1 + \cdots + x_n a_n,
\end{equation}
where $x$ is also a vector with $n$ coordinates. We can view this as a linear transformation from a
space $R^n$ for $x$ to a space $R^n$ for $z$. Or we can view $x_1, \ldots, x_n$ as a set of new coordinates for a
$z$ in the subspace $\mathcal{L}(a_1, \ldots, a_n)$. The $n$ vectors are linearly independent according as $\mathcal{L}(a_1, \ldots, a_n)$
is the whole of $R^n$, or as the transformation just described is nonsingular, or as the determinant of
the matrix $A$ is nonzero, or as the rank of $A$ is $n$.

B INNER PRODUCTS

Now consider the length of a vector and the inner product of two vectors. The squared length
of a vector $z$ in $R^n$ is given by
Plane \( \mathcal{L}(\mathbf{a}_1, \mathbf{a}_2) \) generated by \( \mathbf{a}_1 \) and \( \mathbf{a}_2 \) in \( \mathbb{R}^3 \); the unit cube at the origin suggests the spatial position of the axes.

\[
|z|^2 = z_1^2 + \cdots + z_n^2 = z'z,
\]

where \( z' \) as a row vector is multiplied into \( z \) as a column vector by matrix multiplication; the inner product of vectors \( z_1 \) and \( z_2 \) is given by

\[
(z_1, z_2) = z_{11}z_{12} + \cdots + z_{n1}z_{n2} = z_1^Tz_2.
\]

Now suppose that we have several vectors \( \mathbf{a}_1, \ldots, \mathbf{a}_r \), and want the squared lengths and inner products. Then we can calculate the inner product matrix,

\[
M = A'A = \begin{pmatrix}
\mathbf{a}_1^T \\
\vdots \\
\mathbf{a}_r^T
\end{pmatrix}
\begin{pmatrix}
\mathbf{a}_1 & \cdots & \mathbf{a}_r
\end{pmatrix}
= \begin{pmatrix}
\mathbf{a}_1'\mathbf{a}_1 & \cdots & \mathbf{a}_1'a_r \\
\vdots & \ddots & \vdots \\
\mathbf{a}_r'\mathbf{a}_1 & \cdots & \mathbf{a}_r'a_r
\end{pmatrix}
= \begin{pmatrix}
m_{11} & \cdots & m_{1r} \\
\vdots & \ddots & \vdots \\
m_{r1} & \cdots & m_{rr}
\end{pmatrix}
\]

by matrix multiplication. Note that \( m_{11} \) is the squared length of \( \mathbf{a}_1 \), and \( m_{12} = m_{21} \) is the inner product of \( \mathbf{a}_1 \) and \( \mathbf{a}_2 \). Also note that any square submatrix on the diagonal is also an inner product matrix—for a subset of the vectors \( \mathbf{a}_1, \ldots, \mathbf{a}_r \).

As an example, consider the vectors \( \mathbf{a}_1 = (1, 1, 1, 1, 1)' \) and \( \mathbf{a}_2 = (1, 2, 3, 4, 5)' \) in \( \mathbb{R}^5 \). The inner product matrix is

\[
M = \begin{pmatrix}
1 & 1 & 1 & 1 & 1 \\
1 & 2 & 3 & 4 & 5 \\
1 & 2 & 3 & 4 & 5 \\
1 & 2 & 3 & 4 & 5 \\
1 & 2 & 3 & 4 & 5
\end{pmatrix}
= \begin{pmatrix}
5 & & & & \\
& 15 & & & \\
& & 15 & & \\
& & & 55 & \\
& & & & 55
\end{pmatrix}
\]

Thus the length of \( \mathbf{a}_1 \) is \( \sqrt{5} \) and of \( \mathbf{a}_2 \) is \( \sqrt{55} \); the inner product of \( \mathbf{a}_1 \) and \( \mathbf{a}_2 \) is \( 15 \).

The angle \( \theta \) between two vectors \( z_1 \) and \( z_2 \) can be defined by

\[
(z_1, z_2) = |z_1||z_2| \cos \theta
\]
provided that $|z_1| \neq 0$ and $|z_2| \neq 0$; that is, provided neither $z_1$ nor $z_2$ is the zero vector 0. Two vectors $z_1$ and $z_2$ are called orthogonal if $\langle z_1, z_2 \rangle = z_1^t z_2 = 0$. In the preceding example the cosine of the angle between $a_1$ and $a_2$ is

$$\cos \theta = \frac{15}{\sqrt{5} \cdot \sqrt{55}} = 0.9045$$

and the angle $\theta = 0.4406$.

As a second example, consider the vectors $a_1 = (1, 1, 1, 1, 1)'$ and $a_{2,1} = (-2, -1, 0, 1, 2)'$. The inner product matrix is

$$N = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \\ \end{pmatrix} \begin{pmatrix} 1 & -2 \\ -1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ \end{pmatrix} = \begin{pmatrix} 5 & 0 \\ 0 & 10 \end{pmatrix}.$$ 

Thus the vectors $a_1$ and $a_{2,1}$ are orthogonal.

Consider two vectors $a_1$ and $a_2$. We can always find the component of $a_2$ that is orthogonal to $a_1$; that is, choose $b$ so that $a_2 - ba_1$ is orthogonal to $a_1$:

$$(a_2 - ba_1, a_1) = a_2^t a_1 - ba_1^t a_1 = 0.$$ 

For the initial example we have

$$15 - b \cdot 5 = 0 \text{ or } b = 3;$$

thus

$$a_2 - 3a_1 = (1, 2, 3, 4, 5)' - 3(1, 1, 1, 1, 1)' = (-2, -1, 0, 1, 2)'$$

is orthogonal to $a_1$. Note that this orthogonal component is the vector $a_{2,1}$ used in the second example. The part removed $3a_1$ is the projection of $a_2$ on $a_1$.

C POSITIVE DEFINITE QUADRATIC FORMS

Consider the equation (2) that expresses certain points $z$ in $R^n$ in terms of vectors $a_1, \ldots, a_n$:

$$z = Ax \text{ or } z = x_1 a_1 + \cdots + x_n a_n.$$ 

The squared length of $z$ can be expressed in terms of $x$:

$$z'z = x' A' Ax = x'Mx = \sum_{i,j=1}^n m_{ij} x_i x_j. 

(7)$$

An expression of the form (with $M$ symmetric)

$$x'Mx = \sum_{i,j=1}^n m_{ij} x_i x_j $$

(8)

is called a quadratic form and $M$ is called the matrix of the quadratic form. A matrix $M$ is called positive definite if $x'Mx > 0$ for all $x \neq 0$ and positive semidefinite if $x'Mx \geq 0$ for all $x$. Correspondingly, the quadratic form is called positive definite and positive semidefinite.

Note from equation (7) that an inner product matrix is positive semidefinite; it is also true conversely that a positive semidefinite matrix can be written as an inner product matrix. Also note that if $A$ has full rank $n$, then no nontrivial linear combination of $a_1, \ldots, a_n$ can be 0; thus an inner product matrix of linearly independent vectors is positive definite; it is also true conversely that a positive definite matrix can be written as the inner product matrix for linearly independent vectors.

Now consider the transformation $z = Ax$ and assume that $A$ is of full rank $n$. The transformation from $x$ to $z$ is then a one-to-one transformation of $R^n$ into $R^n$. The transformation dilates,
contracts, rotates, and shears. The sphere \( z'z = \Sigma z_i^2 = c^2 \) for \( z \) corresponds to the ellipsoid
\[ x'Mx = \Sigma m_{ij} x_i x_j = c^2 \] for \( x \).

A positive semidefinite matrix \( M_2 \) is called *greater than* or *equal to* a positive semidefinite matrix \( M_1 \) if \( x'M_2 x - x'M_1 x = x'(M_2 - M_1)x \) is positive semidefinite. If \( M_2 \) is greater than or equal to \( M_1 \), then clearly a diagonal element of \( M_2 \) is greater than or equal to the corresponding diagonal element of \( M_1 \). For the case of positive definite matrices, \( M_2 \) greater than \( M_1 \) can be interpreted as the ellipsoid \( x'M_2 x = c^2 \) inside the ellipsoid \( x'M_1 x = c^2 \).

For positive definite matrices it can be shown that \( M_2 \) greater than or equal to \( M_1 \) implies \( M_1^{-1} \) greater than or equal to \( M_2^{-1} \); see Problem 12.

## D THE SINGULAR COMPONENT ON \( R \)

Consider a distribution on \( R \) with measure \( P \) and distribution function \( F \). In Section 2.1 we defined and separated out the discrete component:

\[
S_d = \{ y : \rho(y) > 0 \}, \quad F_d(y) = P(( - \infty , y ) \cap S_d) = \sum_{t \leq y} \rho(t).
\]

And in Section 2.2 we defined and separated out the absolutely continuous component:

\[
S_a = \{ y : F(y) \text{ exists} \}, \quad F_a(y) = P(( - \infty , y ) \cap S_a) = \int_{-\infty}^{y} f(t) \, dt.
\]

Is there any probability left? The remaining sample space and "distribution function" are

\[
S_s = R - S_d - S_a, \quad F_s(y) = F(y) - F_d(y) - F_a(y) = P(( - \infty , y ) \cap S_s).
\]

Is \( S_s = \emptyset \) and \( F_s(y) = 0 \)?

The function \( F_d \) records cumulative discrete probability. The difference function \( F - F_a \) has the discrete probability removed; it is thus a continuous function. The function \( F_a \) records cumulative probability at points where the derivative exists. The difference function \( F - F_a - F_a \) thus has the derivative removed where it exists; it is thus a continuous function that has derivative 0 throughout \( S_a \). But in Section 2.2 we noted that the length measure of \( R - S_a \) is zero. Thus \( F_a \) is continuous and has derivative 0 except on a set of length measure zero. Is there a nontrivial function \( F_s \)?

Surprisingly, an example can be constructed fairly easily. Initially: consider a uniform distribution on \( A_0 = [0, 1] \); let \( G_0 \) be the distribution function. First: delete the middle third of the preceding interval and consider a uniform distribution on the remainder \( A_1 = [0, 1/3) \cup (2/3, 1] \); let \( G_1 \) be the distribution function. Second: delete the middle thirds of the preceding intervals and consider a uniform distribution on the remainder \( A_2 = [0, 1/9) \cup (2/9, 3/9) \cup (6/9, 7/9) \cup [8/9, 1] \); let \( G_2 \) be the distribution function. Continue in this manner and let \( G(y) = \lim G_n(y) \). See Figure S.3 and compare with Problems 1.5.14 and 1.5.15. In a sense \( G \) describes the uniform distribution on the Cantor ternary set.

From the construction it is easy to see intuitively that \( G \) is continuous with \( G(0) = 0 \), \( G(1) = 1 \); and also that \( G'(y) = 0 \) on all the removed intervals and thus is zero except on the Cantor ternary set \( A = \cap_1^\infty A_n \); Problem 1.5.14 is to show that the length measure of the Cantor set is zero.

An explicit expression can, in fact, be given for \( G(y) \). Let \( y \) in \([0, 1) \) be expressed in a ternary decimal expansion:

\[
y = j_1 j_2 j_3 \cdots \bigg( = \sum_{a=1}^{\infty} j_a 3^{-a^a} \bigg), \quad j_a = 0, 1, 2.
\]

This is unique except for instances involving repeating 2's: round up if preceded by a 1 (thus change 0.10122 to 0.10200); retain if preceded by 0 (thus use 0.1022 in preference to 0.1100).
A point $y$ gets removed by the construction procedure if a 1 occurs in the expansion; let $n(y)$ be the subscript of the first $j$ that is equal to 1 with $n(y) = \infty$ if there is no such $j$. The value of $G(y)$ can then be expressed in binary form as

$$G(y) = i_1 i_2 i_3 \cdots = \sum_{n=1}^{n(y)-1} i_n 2^{-n} + 2^{-n(y)},$$

where $i_n = i_n/2$. Thus $G(0.20220102) = 0.101101$ using ternary and binary decimals.

**E THE SINGULAR COMPONENT ON $R^k$**

A distribution on $R^k$ can be decomposed into a discrete component, an absolutely continuous component, and a singular component, just as on $R$. The possibilities for the singular component are, however, much wider with $k > 1$.

For an example, consider $(y_1, y_2)$ on $R^2$ and suppose that we put a normal distribution along certain line segments. Specifically, let $z$ designate a standard normal variable and define

$$(y_1, y_2) = \begin{cases} (z, z) & \text{if } |z| < c, \\ (-z, -z) & \text{if } |z| \geq c. \end{cases}$$

This is a singular distribution but a rather special one with its split components; this distribution
can be useful for other purposes later. Sketch this distribution as a one-dimensional density function sitting on the three line segments.

**F Exercises and Problems**

1. Calculate the angle between the two vectors for each of the following pairs:
   
   \[
   \begin{pmatrix}
   1 \\ 0 \\ 0
   \end{pmatrix}, \begin{pmatrix}
   1 \\ 1 \\ 0
   \end{pmatrix}, \begin{pmatrix}
   1 \\ 2 \\ 1
   \end{pmatrix}, \begin{pmatrix}
   1 \\ 4 \\ 1
   \end{pmatrix}, \begin{pmatrix}
   1 \\ 9 \\ 4
   \end{pmatrix}, \begin{pmatrix}
   1 \\ 16 \\ 4
   \end{pmatrix}
   \]

2. Calculate the inner product matrix for the column vectors in
   
   \[
   A = \begin{pmatrix}
   1 & 1 & 1 \\
   1 & -1 & -1 \\
   1 & -1 & 1
   \end{pmatrix}
   \]

3. Calculate the inner product matrix for the column vectors in
   
   \[
   A = \begin{pmatrix}
   1 & 1 & 1 \\
   1 & 2 & 4 \\
   1 & 3 & 9 \\
   1 & 4 & 16
   \end{pmatrix}
   \]

4. Let \(a_1 = (1, 1, 1, 1)'\) and \(a_2 = (1, 2, 3, 4)'\). Find \(b\) so that \(a_2, 1 = a_2 - ba_1\) is orthogonal to \(a_1\); record \(a_{2,1}\).

5. Let \(L^1(1)\) consist of all the vectors \(x_2\) which are orthogonal to the 1-vector \(1 = (1, \ldots, 1)'\). Show that \(L^1(1)\) is the plane \(\Sigma x_2 = 0\).

6. Show that the quadratic form \(\Sigma m_{ij}x_i x_j\) with
   
   \[
   M = \begin{pmatrix}
   1/n & \cdots & 1/n \\
   \vdots & \ddots & \vdots \\
   1/n & \cdots & 1/n
   \end{pmatrix}
   \]
   
   is equal to \(n \bar{x}^2\), where \(\bar{x} = \Sigma x_i/n\).

7. Show that the quadratic form \(\Sigma n_{ij}x_i x_j\) with
   
   \[
   N = \begin{pmatrix}
   1 & -1/n & -1/n & \cdots & -1/n \\
   -1/n & 1 & -1/n & \cdots & \vdots \\
   \vdots & \ddots & 1 & -1/n & -1/n \\
   -1/n & \cdots & -1/n & 1 & -1/n
   \end{pmatrix}
   \]
   
   is equal to \(\Sigma x_i^2 - n \bar{x}^2 = \Sigma (x_i - \bar{x})^2\).

8. A matrix \(M\) is idempotent if \(MM = M\). Show that \(M\) and \(N\) in Problems 6 and 7 are idempotent and symmetric.

9. (continuation of Problem 6) Show that \(Mx = \bar{x} 1 = (\bar{x}, \ldots, \bar{x})'\), a point in \(L^1(1)\); it is the projection of \(x\) on \(L^1(1)\).

10. (continuation of Problem 7) Show that \(Nx = x - \bar{x} 1 = (x_1 - \bar{x}, \ldots, x_n - \bar{x})'\), a point in \(L^1(1)\); it is the projection of \(x\) on \(L^1(1)\).

11. The distribution function \(G\) in Section D is called the Cantor function. Show that the Cantor function is continuous.

12. For positive definite matrices, show that if \(M_1\) is greater than \(M_2\), then \(M_1^{-1}\) is greater than \(M_2^{-1}\) (simultaneously diagonalize \(M_1\) and \(M_2\)).
3.6

MARGINAL PROBABILITY

In this section we develop further several of the topics in the preceding sections. This is optional material presented here to round out the development.

A FACTORIZATION OF THE DISTRIBUTION FUNCTION

Proposition 6 in Section 3.3D records the factorization

\[ F_{XY}(x, y) = F_X(x)F_Y(y) \]

as a necessary and sufficient condition for the statistical independence of \( X \) and \( Y \). In that section we proved the necessity—that independence implies the factorization. We now prove the sufficiency—that the factorization \( F_{XY}(x, y) = F_X(x)F_Y(y) \) implies independence. From Section 2.3B we know that the distribution function on \( \mathbb{R}^2 \) uniquely determines the probability measure \( P_{XY} \) for \( (X, Y) \) on the plane. What can we say about this measure on the basis of the factored expression \( F_XF_Y \) for the distribution function? From Section 3.3E we know that the measure \( P_x \) for the first coordinate and \( P_y \) for the second coordinate uniquely combine on the basis of independence to give a measure on \( \mathbb{R}^2 \). This product measure, of course, agrees with \( F_XF_Y \) on the sets \( (-\infty, x] \times (-\infty, y] \); thus it must be the unique probability measure \( P_{XY} \) determined from the distribution function. Hence \( X \) and \( Y \) are statistically independent.

B SOME NORMAL DISTRIBUTION THEORY

The normal distribution provides an approximation for the variation found with many measurement and response variables. And it happens to be essentially the only distribution for which certain calculations of statistical interest are simple and direct. These calculations concern certain marginal distributions of common functions. We shall now derive some of these distributions.

Consider a standard normal distribution for \( z \). From Example 3.2.2 we have that \( z^2 \) has a chi-square(1) distribution. If we rescale by a factor \( \sigma \) and have a normal(0, \( \sigma^2 \)) distribution for \( \nu = az \), then \( \nu^2/\sigma^2 = z^2 \) has a chi-square(1) distribution, or equivalently, \( \nu^2 \) has the distribution of \( \sigma^2\chi^2 \), where \( \chi^2 \) has the chi-square(1) distribution.

Now consider independent \( \chi_1^2 \) and \( \chi_2^2 \), where \( \chi_1^2 \) is chi-square(\( \eta_1 \)). The joint probability element is

\[
\frac{1}{\Gamma(n_1/2)} \frac{1}{\Gamma(n_2/2)} e^{-\chi_1^2/2} \left( \frac{\chi_1^2}{2} \right)^{(n_1/2) - 1} e^{-\chi_2^2/2} \left( \frac{\chi_2^2}{2} \right)^{(n_2/2) - 1} d\chi_1 d\chi_2
\]

We derive the marginal distribution of \( \chi^2 = \chi_1^2 + \chi_2^2 \). For this, we first change the variable from \( (\chi_1^2, \chi_2^2) \) to \( (\chi^2, \chi_2^2) \) as in the convolution example 3.2.8: the Jacobian is 1; hence, by substitution, we obtain the joint probability element

\[
\frac{1}{\Gamma(n_1/2)} \frac{1}{\Gamma(n_2/2)} e^{-\chi^2/2} \left( \frac{\chi^2}{2} \right)^{(n_1/2) - 1} \left( \frac{\chi_2^2}{2} \right)^{(n_2/2) - 1} d\chi d\chi^2
\]

\[
= \frac{1}{\Gamma(n_1/2)} \frac{1}{\Gamma(n_2/2)} e^{-\chi^2/2} \left( \frac{\chi^2}{\chi^2} \right)^{(n_1/2) - 1} \left( \frac{\chi_2^2}{\chi^2} \right)^{(n_2/2) - 1} d\chi d\chi^2
\]
on the range \(0 < \chi^2 < \infty, \quad 0 < \chi^3 < \chi^2\). We then integrate out the unwanted \(\chi^2\): the essential part of the integration is as follows:

\[
\int_0^{\chi^2} \left( \frac{x^2 - \chi^2}{\chi^2} \right)^{(n/2-1)} \left( \frac{\chi^2}{x^2} \right)^{(n/2-1)} \left(\frac{1}{2}\right) dx^2 = \int_0^{1} (1 - \frac{u^{n/2-1} - 1}{\Gamma(n/2-1)} u^{(n/2-1)} du \cdot \frac{\chi^2}{2},
\]

\[
= \frac{\Gamma(n/2) \Gamma(n_2/2)}{\Gamma(n_1 + n_2/2)} \frac{\chi^2}{2}
\]

where we have substituted \(u = \chi^2/\chi^2\) and have used Problem 2.2.22. The result of the integration on the full expression is

\[
(1) \quad \frac{1}{\Gamma(n_1 + n_2/2)} e^{-x^2/2} \left( \frac{\chi^2}{2} \right)^{(n_1 + n_2/2)-1} \frac{d\chi^2}{2}.
\]

Thus \(\chi^2 = \chi^3 + \chi^2\) is chi-square\(n_1 + n_2\).

We now combine the results in the preceding paragraphs. Consider a sample \((z_1, \ldots, z_n)\) from the standard normal. It follows first that \((z^2_1, \ldots, z^2_n)\) is a sample from the chi-square\((1)\); and then that \(\chi^2 = z^2_1 + \cdots + z^2_n\) is chi-square\(n\). If we rescale by a factor \(\sigma\) and have a sample \((v_1, \ldots, v_n)\) from the normal\((0, \sigma)\), then \(\sum v_i^2/\sigma^2\) has the distribution of \(\sigma^2 \chi^2\), where \(\chi^2\) is chi-square\(n\).

Now consider independent \(z, \chi^2\) where \(z\) is standard normal and \(\chi^2\) is chi-square\(n\). The joint probability element is

\[
\frac{1}{\sqrt{2\pi}} e^{-z^2/2} \frac{1}{\Gamma(n/2)} e^{-\chi^2/2} \left( \frac{\chi^2}{2} \right)^{(n/2)-1} \frac{d\chi^2}{2}.
\]

We derive the marginal distribution of \(T = z/\chi\). For this we first change the variable from \((z, \chi^2)\) to \((T, \chi^2)\) parallel to the convolution problem 3.2.30: the Jacobian change is given by \(dz = \chi dt\); hence by substitution we obtain

\[
\frac{1}{\sqrt{\pi}} \frac{1}{\Gamma(n/2)} e^{-z^2/2} e^{-\chi^2/2} \left( \frac{\chi^2}{2} \right)^{(n+1)/2-1} \frac{d\chi^2}{2} dT
\]

on the range \(0 < \chi^2 < \infty, -\infty < T < \infty\). We then integrate out the unwanted \(\chi^2\): the essential part of the integration is as follows:

\[
\int_0^{\infty} e^{-\chi^2/(1+T^2/2)} \left( \frac{\chi^2}{2} \right)^{(n+1)/2-1} \frac{d\chi^2}{2} = \int_0^{\infty} e^{-x^2/(1+T^2)} \left(1 + T^2\right)^{-(n+1)/2} dx
\]

\[
= \Gamma\left(\frac{n + 1}{2}\right) (1 + T^2)^{-(n+1)/2}
\]

where \(x = \chi^2(1 + T^2)/2\) and Problem 2.2.21 is used; the result of the integration on the full expression is

\[
(2) \quad \Gamma\left(\frac{n + 1}{2}\right) \frac{\Gamma(n/2)}{\Gamma(1/2)\Gamma(n/2)} (1 + T^2)^{-(n+1)/2} dT,
\]

using \(\sqrt{\pi} = \Gamma(1/2)\). Thus \(T = z/\chi\) is canonical Student\(n\) and \(t = z/(\chi/\sqrt{n})\) is ordinary Student\(n\); see Problem 2.2.19.

We now combine the results in the preceding paragraphs. If we have a sample \((v, v_1, \ldots, v_n)\) from the normal\((0, \sigma)\) distribution, then
\[ t = \frac{\nu}{\left( \sum v_i^2/n \right)^{1/2}} \]

has the Student(\(\nu\)) distribution.

Now consider independent \(\chi_i^2\) and \(\chi_j^2\), where \(\chi_i^2\) is chi-square(\(n_i\)). The joint probability element is

\[ \frac{1}{\Gamma(n_i/2)} \frac{1}{\Gamma(n_j/2)} e^{-\chi_i^2/2} \left( \frac{\chi_i^2}{2} \right)^{(n_i/2)-1} e^{-\chi_j^2/2} \left( \frac{\chi_j^2}{2} \right)^{(n_j/2)-1} \frac{d\chi_i^2}{2} \frac{d\chi_j^2}{2}. \]

We derive the marginal distribution of \(G = \chi_i^2/\chi_j^2\). For this we first change the variable from \((\chi_i^2, \chi_j^2)\) to \((G, \chi_j^2)\) as in the convolution problem 3.2.30: the Jacobian change is given by \(d\chi_i^2 = \chi_j^2 \, dG\); hence by substitution we obtain

\[ \frac{1}{\Gamma(n_i/2)} \frac{1}{\Gamma(n_j/2)} e^{-\chi_j^2/(1+G)/2} \left( \frac{\chi_j^2}{2} \right)^{(n_i+n_j)/2 - 1} G^{(n_i-1)/2} \, dG \frac{d\chi_j^2}{2} \]

on the range \(0 < G < \infty, 0 < \chi_j^2 < \infty\). We then integrate out the unwanted \(\chi_j^2\); the essential part of the integration is as follows:

\[ \int_0^\infty e^{-x^{(1+G)/2}\left( \frac{\chi_j^2}{2} \right)^{(n_i+n_j)/2 - 1} G^{(n_i-1)/2}} \, dx(1+G)^{(n_i+n_j)/2}, \]

\[ = \Gamma \left( \frac{n_i + n_j}{2} \right) (1+G)^{(n_i+n_j)/2}, \]

where we have substituted \(x = \chi_j^2(1+G)\) and have used Problem 2.2.21; the result of the integration on the full expression is

\[ \frac{\Gamma \left( \frac{n_i + n_j}{2} \right)}{\Gamma(n_i/2) \Gamma(n_j/2)} \frac{G^{(n_i-1)/2}}{(1+G)^{(n_i+n_j)/2}} \, dG. \]

Thus \(G = \chi_i^2/\chi_j^2\) is canonical \(F(n_i, n_j)\); and \(F = (\chi_i^2/n_i)/(\chi_j^2/n_j)\) is ordinary \(F(n_i, n_j)\); see Problem 2.2.18.

We now combine results in preceding paragraphs. If we have a sample \((v_1, \ldots, v_{n_1}, v_{n_1+1}, \ldots, v_{n_1+n_2})\) from the normal(\(\mu, \sigma\)) distribution, then

\[ F = \frac{\sum_{n_i} v_i^2/n_i}{\sum_{n_i+n_2} v_i^2/n_2} \]

has the \(F(n_1, n_2)\) distribution.

**C FURTHER NORMAL DISTRIBUTION RESULTS: THEORY AND PROBLEMS**

Consider the normal(\(\mu, \sigma\)) distribution for \(x\). We derive the marginal distribution for \(\chi^2 = x^2\).

The probability element for \(x\) is

\[ g(x) \, dx = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} e^{-y^2/2} \, dy. \]

The probability element for \(\chi^2 = x^2\) is then obtained from Example 3.2.7:
(4) \[ h(\chi^2) \, d\chi^2 = (1/2)(g(\chi) + g(-\chi)) \chi^{-1} \, d\chi^2 \]

\[ = \frac{1}{\sqrt{2\pi}} \, e^{-\chi^2/2} \left( 1 + \chi^2 \frac{\delta^2}{2!} + \chi^4 \frac{\delta^4}{4!} + \cdots \right) e^{-\chi^2/2} \chi^{-1} \, d\chi^2 \]

\[ = e^{-\chi^2/2} \sum_{r=0}^{\infty} \frac{(\delta^2/2)^r}{r!} \cdot \frac{1}{\Gamma \left( \frac{2r+1}{2} \right)} \left( \frac{\chi^2}{2} \right)^{\frac{2r+1}{2}-1} e^{-\chi^2/2} \frac{d\chi^2}{2}, \]

where the constants have been rearranged according to

\[ \sqrt{2\pi} \, 2r(2r-1) \cdots 2.1 = 2^r \Gamma(2r+1) \Gamma(1/2) \]

to give chi-square density expressions. The distribution (4) is called the noncentral chi-square distribution with noncentrality \( \delta^2 \): note the Poisson(\( \delta^2/2 \)) mixture of chi-square(1 + 2r)'s.

1. **Noncentral normal(\( \delta, 1 \)) distribution**: Consider the normal(\( \delta, 1 \)) distribution for \( x \). Show that the density function for \( x \) can be expressed as

\[ f(x) = e^{-x^2/2} \frac{1}{\sqrt{2\pi}} \sum_{r=0}^{\infty} \frac{\delta^r}{r!} \, x^r. \]

2. **Noncentral canonical Student(\( n \)) distribution with noncentrality \( \delta \)**: Consider the noncentral normal(\( \delta, 1 \)) distribution for \( x \) and an independent chi(\( n \)) distribution for \( \chi \). Use the method in Section B to show that the marginal density for \( T = \frac{x}{\chi} \) has density function

\[ g(T) = e^{-T^2/2} \sum_{r=0}^{\infty} \frac{\delta^r}{r!} \frac{\Gamma \left( \frac{n+r+1}{2} \right)}{\Gamma(1/2)\Gamma(n/2)} \frac{T^r}{1 + \frac{\delta^2}{2} \Gamma(n+r+1)/r} \]

called the noncentral canonical Student(\( n \)) distribution. The rescaled distribution for \( t = \frac{x}{\chi^2/n}^{1/2} \) is called the ordinary Student(\( n \)) distribution with noncentrality \( \delta \); see Figure S.4.

**FIGURE S.4**
Noncentral Student(6) distribution with noncentrality \( \delta^2 \).
3 \textit{Noncentral chi-square(m) distribution with noncentrality \(\delta^2\)}: Consider a normal(\(\delta, 1\)) distribution for \(x\) and an independent standard normal sample \((z_1, \ldots, z_{m-1})\) [or an independent chi-square\((m - 1)\) distribution for \(\chi^2_{z} = z_1^2 + \cdots + z_{m-1}^2\)]. Show that the marginal distribution for \(\chi^2\)

\[\chi^2 = x^2 + \chi^2_{z} = x^2 + z_1^2 + \cdots + z_{m-1}^2\]

has density

\[e^{-x^2/2} \sum_{r=0}^{\infty} \frac{(\delta^2/2)^r}{r!} f(\chi^2:m + 2r),\]

where \(f(\chi^2:m)\) is the chi-square\((m)\) density; apply the convolution formula (3.2.25) term by term to the results in formula (3.6.4). Note the preceding form: a Poisson combination of chi-squares; see Figure S.5.

4 \textit{Noncentral canonical \(F(m, n)\) distribution with noncentrality \(\delta^2\)}: Consider a noncentral chi-square\((m)\) distribution for \(\chi^2_x\) with noncentrality \(\delta^2\) and an independent chi-square\((n)\) distribution for \(\chi^2_y\). Show that the marginal distribution for \(G = \frac{\chi^2_x}{\chi^2_y}\) has density

\[e^{-x^2/2} \sum_{r=0}^{\infty} \frac{(\delta^2/2)^r}{r!} f(G:m + 2r, n),\]

where \(f(G:m, n)\) is the canonical \(F(m, n)\) distribution; apply the convolution procedure in Section B term by term to the results of Problem 3; see Figure S.6.

5 Let \(\chi^2_1, \ldots, \chi^2_k\) be statistically independent and \(\chi^2_1\) have a chi-square\((m_1)\) distribution. Let \(w = \chi^2_1 + \cdots + \chi^2_k\) and \(u_1 = \chi^2_1/w, \ldots, u_{k-1} = \chi^2_{k-1}/w\) as in Problem 3.3.9.

(a) Show that \(w\) is statistically independent of \((u_1, \ldots, u_{k-1})\): make the change of variable from \((\chi^2_1, \ldots, \chi^2_k)\) to \((\chi^2_1, \ldots, \chi^2_{k-1}, w)\) and then to \((u_1, \ldots, u_{k-1}, w)\).

(b) Then deduce that \(w\) is chi-square\((\sum m_i)\) and \((u_1, \ldots, u_{k-1})\) is Dirichlet\((m_1/2, \ldots, m_k/2)\).
D INDEPENDENCE
FOR NORMAL DISTRIBUTIONS:
PROBLEMS

6 Show that \( ay_1^2 + by_1y_2 + cy_2^2 + dy_1 + ey_2 + f \) can be expressed as \( g(y_1) + h(y_2) \) on \( R^2 \) if and only if \( b = 0 \). Show that \( f(y_1, y_2) = \exp \{ ay_1^2 + by_1y_2 + cy_2^2 + dy_1 + ey_2 + f \} \) can be expressed as \( g(y_1)h(y_2) \) on \( R^2 \) if and only if \( b = 0 \).

7 The bivariate normal \((\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)\) distribution is recorded in Problem 2.4.24. Show that \( y_1 \) and \( y_2 \) are statistically independent if and only if \( \rho = 0 \); use Problem 6.

8 Show that \( \Sigma a_i x_i^2 + 2 \Sigma b_{iu} x_i y_u + \Sigma c_u y_u^2 + \Sigma d_i x_i + \Sigma e_y y_u + f \) can be expressed as \( g(x) + h(y) \) on \( R^{k+1} \) if and only if \( b_{iu} = 0 \) for all \( i \) and \( u \).

9 The multivariate normal distribution on \( R^{k+1} \). Let \( y_1 = (y_1, \ldots, y_k)' \), \( y_2 = (y_{k+1}, \ldots, y_{k+1})' \), \( \mu_1 = (\mu_1, \ldots, \mu_k)' \), \( \mu_2 = (\mu_{k+1}, \ldots, \mu_{k+1})' \).

\[
\begin{align*}
\Sigma^{11} &= \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix}, & \Sigma^{22} &= \begin{pmatrix} a_{k+1,k+1} & \cdots & a_{k+1,k+1} \\ \vdots & \ddots & \vdots \\ a_{k+1,k+1} & \cdots & a_{k+1,k+1} \end{pmatrix}, & \Sigma^{12} &= \begin{pmatrix} a_{1k+1} & \cdots & a_{1k+1} \\ \vdots & \ddots & \vdots \\ a_{k+1,k+1} & \cdots & a_{k+1,k+1} \end{pmatrix}, \\
\Sigma^{-1} &= \begin{pmatrix} \Sigma_{11}^{11} & \Sigma_{11}^{12} \\ \Sigma_{21}^{11} & \Sigma_{22}^{11} \end{pmatrix}, & \Sigma &= \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},
\end{align*}
\]

where \( \Sigma \) and \( \Sigma^{-1} \) are defined in Section 2.4G. Suppose that \( y = (y_1', y_2') \) has the multivariate normal \((\mu = (\mu_1, \mu_2)', \Sigma)\) distribution with density

\[
f(y) = (2\pi)^{-k/2} |\Sigma|^{-1/2} \exp \left\{-\frac{1}{2} \begin{pmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \end{pmatrix} \right\}.
\]

Use partitioned-matrix multiplication and show that \( y_1 \) and \( y_2 \) are statistically independent if and only if \( \Sigma^{12} = 0 \) (the zero matrix); if and only if \( \Sigma_{12} = 0 \).

E ORDER STATISTIC DISTRIBUTIONS:
THEORY AND PROBLEMS

We have noted before that the normal distribution is often used to describe the variation in many measurement and response variables. We will see later that it suffices with the normal to
consider the sample average \( \bar{y} = \Sigma y_i / n \) and the sample variance \( s_y^2 = \Sigma (y_i - \bar{y})^2/(n - 1) \). Some of the distribution theory in Sections B and C will be used later for statistical analysis based on \( \bar{y} \) and \( s_y^2 \).

For other continuous distribution forms it is often convenient to examine the distribution of the order statistics. Let \( y = (y_1, \ldots, y_n)' \) designate a point in \( R^n \); and let \( y_{(1)} \) be the smallest of the coordinates \( y_1, y_{(2)} \) the second smallest, \ldots, \( y_{(n)} \) the largest; thus \( y_{(1)} < \cdots < y_{(n)} \). The vector \( y_0 = (y_{(1)}, \ldots, y_{(n)})' \) is called the order statistic and a coordinate \( y_{(i)} \) is called the \( i \)th order statistic.

More formally we can view \( y_0 \) as a function carrying \( y \) in \( R^n \) to \( y_0 \) in \( R^n \), actually the subset of \( R^n \) having \( y_{(1)} < \cdots < y_{(n)} \).

Consider the distribution of a sample \( y = (y_1, \ldots, y_n)' \) from a distribution with density function \( f \). The density for \( y \) is

\[
f(y_1) \cdots f(y_n).
\]

We first derive the marginal density for the order statistic \( y_0 \). For this we follow the pattern in Section 3.2D: the mapping \( y \rightarrow y_0 \) is an \( n! \)-to-1 mapping except at points having two or more equal coordinates; the volume change for any component of the mapping is +1, just a reordering of coordinates; the density function at any component point is \( f(y_{(1)}) \cdots f(y_{(n)}) \) with factors rearranged in a convenient order; it follows that the joint density for \( y_0 \) is

\[
(9) \quad n! f(y_{(1)}) \cdots f(y_{(n)})
\]

for \( y_{(1)} < \cdots < y_{(n)} \) and zero elsewhere.

We now derive the marginal distribution for, say, the largest value \( y_{(n)} \) in the sample. For this we integrate first with respect to \( y_{(1)} \) over the effective range \((-\infty, y_{(2)})\),

\[
\int_{-\infty}^{y_{(2)}} n! f(y_1) \cdots f(y_{(n)}) \, dy_{(1)} = n! F(y_{(2)}) f(y_{(2)}) \cdots f(y_{(n)}).
\]

Next we integrate with respect to \( y_{(2)} \) over the effective range \((-\infty, y_{(3)})\),

\[
\int_{-\infty}^{y_{(3)}} n! F(y_{(2)}) f(y_{(2)}) \cdots f(y_{(n)}) \, dy_{(2)} = \frac{n!}{2} F^2(y_{(3)}) f(y_{(3)}) \cdots f(y_{(n)}),
\]

where \( \int_{-\infty}^{y_{(2)}} F(u) f(u) \, du = \int_{-\infty}^{y_{(3)}} F(u) f(u) \, du = (1/2) F^2(y_{(2)}) \); continuing in this way we obtain the marginal density

\[
(10) \quad n F^{n-1}(y_{(n)}) f(y_{(n)}).
\]

Note that this result can also be obtained from Exercise 3.2.8.

For further marginal distributions we take advantage of the inverse probability integral transformation in Section 2.4D. Consider the distribution of a sample \( (u_1, \ldots, u_n) \) from the uniform \((0, 1)\) distribution; let \( (u_{(1)}, \ldots, u_{(n)}) \) be the order statistic. Now consider a continuous distribution function \( F \) with inverse function \( y = Y(u) = F^{-1}(u) \). Then it follows from Section 2.4D and Proposition 3.3.5 that \( y_1, \ldots, y_n = (Y(u_1), \ldots, Y(u_n)) \) has the distribution of a sample from \( F \); and in addition by the monotonicity of \( Y \) we have \( y_{(1)} = Y(u_{(1)}), \ldots, y_{(n)} = Y(u_{(n)}) \). Thus we can derive order statistic distributions for the uniform and then transform to the general case.

To illustrate the preceding we determine the marginal distribution of \( y_{(n)} = Y(u_{(n)}) \), the \( n \)th smallest order statistic for a sample from a continuous distribution function \( F \); suppose that the distribution is absolutely continuous with density function \( f \). The probability differential for \( u_{(1)}, \ldots, u_{(n)} \) on \((0, 1)^n\) is

\[
(11) \quad n! \, du_{(1)} \cdots du_{(n)}.
\]

We first integrate \( u_{(1)} \) on \((0, u_{(2)}), u_{(2)} \) on \((0, u_{(3)}), \ldots, u_{(n-1)} \) on \((0, u_{(n)})\), and obtain, in the pattern of formula (10),

\[
\frac{n!}{(r-1)!} u_{(r)}^{-1} \, du_{(1)} \cdots du_{(n)}.
\]
We then integrate \( u_\alpha \) on \((u_{\alpha-1}, 1), u_{\alpha-1}\) on \((u_{\alpha-2}, 1), \ldots, u_{\alpha+1}\) on \((u_{\alpha}, 1)\), and obtain
\[
(12) \quad \frac{n!}{(r-1)!(n-r)!} u_\alpha^{n-r}(1 - u_\alpha)^{n-r} du_\alpha
\]
on the range \((0, 1)\). Finally, we apply the inverse probability integral transform: we have \(u_\alpha = F(y_\alpha)\) and \(du = f(y) dy\); thus the probability density for \(y_\alpha\) is
\[
(13) \quad \frac{n!}{(r-1)!(n-r)!} F_r^{-1}(y_\alpha)(1 - F(y_\alpha))^{n-r} f(y_\alpha)
= \frac{\Gamma(n+1)}{\Gamma(r)\Gamma(n-r+1)} F_r^{-1}(y_\alpha)(1 - F(y_\alpha))^{n-r} f(y_\alpha).
\]

10 Consider the distribution of the order statistic \((u_1, \ldots, u_n)\) for a sample from the uniform\((0, 1)\). The amount of probability in the intervals between adjacent order statistics is given by the coverages \(c_1 = u_1, c_2 = u_2 - u_1, \ldots, c_n = u_n - u_{n-1}, c_{n+1} = 1 - u_n\). Show that \(c = (c_1, \ldots, c_n)\) has density equal to \(n!\) in the region \(0 < c_1; \Sigma_j^k c_j < 1\) and equal to 0 elsewhere; the same distribution applies for any of the \(n + 1\) original \(c_i\)'s. Verify that this is Dirichlet\((1, \ldots, 1)\) with \(k = n\).

11 (continuation) Let \(C_1 = c_1 + \cdots + c_1, C_2 = c_1 + c_2, \ldots, C_k = c_1 + c_2 + \cdots + c_k\) where \(\Sigma_j r_j < n + 1\). Show that \((C_1, \ldots, C_k)\) has the Dirichlet\((r_1, \ldots, r_{k+1})\) distribution, where \(r_{k+1} = n + 1 - \Sigma_j^k r_j\); recall the convolution result in Problem 3.2.26.

12 (continuation) Let \(U_1 = u_{r_1}, U_2 = u_{r_1+r_2}, U_3 = u_{r_1+r_2+r_3}, \ldots, U_{k+1} = u_{r_1+ \cdots + r_k}\) be a sample from the distribution \(F\) with density function \(f\). Use Problem 12 to show that the marginal density for \((x_1, \ldots, x_k) = (y_{r_1}, y_{r_1+r_2}, \ldots, y_{r_1+ \cdots + r_k})\) is
\[
(14) \quad \frac{\Gamma(n+1)}{\Gamma(r_1) \cdots \Gamma(r_{k+1})} U_1^{-1}(U_2 - U_1)^{r_2-1} \cdots (U_k - U_{k-1})^{r_k-1}(1 - U_k)^{r_{k+1}-1}
\]
for \(0 < U_1 < \ldots < U_k < 1\) and zero elsewhere.

13 Consider the distribution of the order statistic \((y_{r_1}, \ldots, y_{r_k})\) for a sample from the distribution \(F\) with density function \(f\). Use Problem 12 to show that the marginal density for \((x_1, \ldots, x_k) = (y_{r_1}, y_{r_1+r_2}, \ldots, y_{r_1+ \cdots + r_k})\) is
\[
(15) \quad \frac{\Gamma(n+1)}{\Gamma(r_1) \cdots \Gamma(r_{k+1})} F_1^{-1}(x_1) F_2^{-1}(x_2 - x_1) F_3^{-1}(x_3 - x_2) \cdots f(x_k)
\]
on \(-\infty < x_1 < \cdots < x_k < \infty\) and zero elsewhere; note that \(r_{k+1} = n + 1 - r_1 - \cdots - r_k\).

14 Let \((y_1, \ldots, y_s)\) be a sample from the distribution \(F\) with density \(f\). Show that the order statistic \((y_{r_1}, \ldots, y_{r_k})\) and rank statistic \((r_1, \ldots, r_k)\) are statistically independent; see Exercise 3.2.15.

F PRODUCT SPACES

In Section 3.3E we considered two independent random systems; and we constructed the probability model for the combined system from models for the component systems using independence and the extension theorem. We also indicated how we could continue and combine additional models one by one; as an illustration we constructed the models for a sample from the Bernoulli, the Poisson, and the normal.

Now suppose that we have probability models \((\mathcal{S}_t, \mathcal{G}_t, P_t)\) for each value of \(t\) in an index set \(T\); the index set \(T\) can be finite, countable, or even larger; let \(s_t\) designate a point in \(\mathcal{S}_t\). For example, \(s_t\) could be voltage at time \(t\) and we would be interested in the voltage function \(s_t\) on the time
interval \((0, \infty)\). Certainly, we can consider \(s_t\) as being random: we would, however, have some difficulty in considering duplicate observations at time \(t\); for we would need to consider duplicates of the whole system.

Consider a combined model based on independence. The combined response \(s\) has a coordinate \(s_t\) for each value of the index \(t\): thus \(s\) is a function from \(T\) into \(\bigcup_T S_t\), where the \(t\)th coordinate is restricted to the subset \(S_t\) of \(\bigcup_T S_t\). The space of such responses \(s\) is called the corresponding product space and is designated \(\mathcal{S} = \prod_T S_t\).

We would certainly be interested in an event that involves assertions concerning a finite number of coordinates; consider the rectangles \(\prod_T A_t\), where only a finite number of the \(A_t\) are not equal to the corresponding \(S_t\)'s. We then define the product algebra \(\mathcal{G} = \prod T \mathcal{G}_t\) to be the smallest \(\sigma\)-algebra that includes these rectangles.

On the basis of independence we define \(P\) for any rectangle by \(P(\prod_T A_t) = \prod_T P(A_t)\), where of course only a finite number of the \(P(A_t)\) are different from 1. Then by additivity this produces \(P\) for any finite union of disjoint rectangles. Additivity for this function \(P\) is immediate; continuity at \(\emptyset\) requires some details of proof. It follows then by the Extension Theorem in Section 1.5D that \(P\) is uniquely defined on the \(\sigma\)-algebra \(\mathcal{G}\). We write \(P = \prod_T P_t\).

Thus we have obtained the arbitrary-product space

\[
(\mathcal{S}, \mathcal{G}, P) = \left(\prod_T S_t, \prod_T \mathcal{G}_t, \prod_T P_t\right).
\]

We can very easily, in fact, be much more general. As an alternative to independence, suppose that for each finite subset \(K = \{t_1, \ldots, t_k\}\) of \(T\) we have a probability measure \(P_K\) on the corresponding class \(\prod_K \mathcal{G}_t\) of events; we assume that the probability measures are consistent in the sense that if \(K' \subset K\), then \(P_{K'}\) is the appropriate marginal measure from \(P_K\).

We then proceed as before. We form the space \(\mathcal{S} = \prod_T S_t\), and the algebra \(\mathcal{G} = \prod_T \mathcal{G}_t\). Then for the measure \(P\) on a set \(\prod_T A_t\), we use the measure \(P_K\) for the set \(K\) of coordinates having \(A_t\) not equal to \(S_t\). The consistency of the \(P_K\), then, assures that \(P\) is a measure and the Extension Theorem uniquely determines \(P\) on the \(\sigma\)-algebra \(\mathcal{G}\); for this general case some topological properties are needed for the spaces \(S_t\).

Thus we have obtained a general product space \((\prod_T S_t, \prod_T \mathcal{G}_t, P)\), where \(P\) is derived from mutually consistent probability measures \(P_K\) defined on finite products.

**G  MEASURABLE FUNCTIONS: THEORY AND PROBLEMS**

We now return to the measurable functions of Section 3.1. For this discussion we shall assume that any given space has a \(\sigma\)-algebra associated with it. And, in particular, if the space is \(R\), then the \(\sigma\)-algebra is the Borel class \(\mathcal{B}\); and if the space is \(R^k\), then the \(\sigma\)-algebra is the Borel class \(\mathcal{B}^k\); and if the space is a product space \(S_1 \times S_2\), then the \(\sigma\)-algebra is the smallest \(\sigma\)-algebra containing the product sets \(A_1 \times A_2\), where \(A_t\) is in the \(\sigma\)-algebra \(\mathcal{G}_t\) for the space \(S_t\).

We first record some propositions; the proofs are assembled at the end of the section.

**PROPOSITION 1**

Any simple function is measurable; any discrete function is measurable.

**PROPOSITION 2**

Any continuous function (real- or vector-valued) is measurable.

**PROPOSITION 3**

If \(Y_1\) and \(Y_2\) are measurable functions on the same space, then \((Y_1, Y_2)\) into the product space is measurable.
PROPOSITION 4

If \( X \) is measurable into a space on which a measurable \( Y \) is defined, then the composite function \( YX \) [taking \( s \) into \( Y(X(s)) \)] is measurable.

PROPOSITION 5

If \( Y_1, Y_2, \ldots \) are measurable functions (real- or vector-valued) and if

\[
\lim_{i \to \infty} Y_i(s) = Y(s)
\]

for all \( s \), then \( Y = \lim Y_i \) is measurable.

By applying these in various combinations, we can obtain a great wealth of measurable functions. For example, if \( X \) and \( Y \) are real-valued measurable functions, then \( X + Y, X \times Y, cX, \max(X, Y), \min(X, Y), p(X, Y) \) are measurable, where \( c \) is real and \( p \) is a polynomial. Also, the positive and negative parts of \( Y \), the functions \( Y^+ = \max(Y, 0) \) are \( Y^- = \max(-Y, 0) \), are measurable.

Before recording the proofs it seems reasonable to consider just how remote and strange a measurable function can be. Consider the case of a bounded measurable function \( Y \) with \( -M \leq Y(s) < M \), where \( M \) is an integer. Let \( A_n(i) \) be the preimage of the interval \([i - 1/n, i/n]\) for \( i = -Mn + 1, -Mn + 2, \ldots, Mn \):

\[
A_n(i) = \left\{ s : \frac{i - 1}{n} \leq Y(s) < \frac{i}{n} \right\}.
\]

And let \( X_n \) and \( Y_n \) be simple functions approximating \( Y \) from below and above:

\[
X_n = \sum_{i=-Mn+1}^{Mn} \frac{i - 1}{n} I_{A_n(i)} \quad \text{and} \quad Y_n = \sum_{i=-Mn+1}^{Mn} \frac{i}{n} I_{A_n(i)}.
\]

Then clearly \( X_n \leq Y < Y_n \) and \( Y_n - X_n = 1/n \). And it follows that

\[
\lim_{n \to \infty} X_n = Y, \quad \lim_{n \to \infty} Y_n = Y
\]

and the convergence is uniform. Thus: a bounded measurable function is the uniform limit of a sequence of simple functions.

Proof of Proposition 1 A simple function is a discrete function. A discrete function has the form

\[
Y = \sum_{i=1}^{\infty} a_i I_{A_i},
\]

where \( A_i \) is the preimage of the value \( a_i \) for the function and the \( A_i \) are in the class \( G \). The preimage, then, of any set \( B \) is a finite or countable union of \( A_i \)'s and is thus in the class \( G \).

Proof of Proposition 4 Let \( C \) be a set in the \( \sigma \)-algebra in the image space of \( Y \); then \( Y^{-1}(C) \) is a set in the \( \sigma \)-algebra on the image space of \( X \); then \( X^{-1}Y^{-1}(C) \) is a set in the \( \sigma \)-algebra on the given space for \( X \). But

\[
X^{-1}Y^{-1}(C) = \{ s : Y(X(s)) \in C \} = (YX)^{-1}(C);
\]

thus \( YX \) is measurable.
For the remaining proofs we need the following preliminary lemma:

**Lemma 6**

Let \( Y \) be a function from \( S \) with \( \sigma \)-algebra \( \mathcal{G} \) into \( T \) with \( \sigma \)-algebra \( \mathcal{B} \). If each \( B \in \mathcal{B}_0 \) has \( Y^{-1}(B) \) in \( \mathcal{G} \) and if \( \mathcal{B} \) is the smallest \( \sigma \)-algebra containing \( \mathcal{B}_0 \), then \( Y \) is measurable.

**Proof**

Let \( \mathcal{B}^* \) consist of sets \( B \) whose preimages are in \( \mathcal{G} \):

\[
\mathcal{B}^* = \{ B : Y^{-1}(B) \in \mathcal{G} \},
\]

then \( \mathcal{B}_0 \subseteq \mathcal{B}^* \). From the properties (3.1.9) of the inverse function we have: if \( B \) is in \( \mathcal{B}^* \), then \( B^c \) is in \( \mathcal{B}^* \); if \( B_i \) is in \( \mathcal{B}^* \), then \( \bigcup_i B_i \) is in \( \mathcal{B}^* \). Thus \( \mathcal{B}^* \) is a \( \sigma \)-algebra and it must contain the smallest \( \sigma \)-algebra \( \mathcal{B} \). It follows that \( B \) in \( \mathcal{G} \) implies \( Y^{-1}(B) \) in \( \mathcal{G} \) and hence that \( Y \) is measurable.

**Proof of Proposition 3**

Let \( Y_1 \) map into \( T_1 \) with \( \sigma \)-algebra \( \mathcal{B}_1 \) and \( Y_2 \) map into \( T_2 \) with \( \sigma \)-algebra \( \mathcal{B}_2 \). Then for \( B_1 \) in \( \mathcal{B}_1 \), \( Y_1^{-1}(B_1) \) is in \( \mathcal{G} \); and for \( B_2 \) in \( \mathcal{B}_2 \), \( Y_2^{-1}(B_2) \) is in \( \mathcal{G} \). Thus consider \( (Y_1, Y_2)^{-1}(B_1 \times B_2) = (s : (Y_1(s), Y_2(s)) \in B_1 \times B_2) \)

\[
= \{ s : Y_1(s) \in B_1, Y_2(s) \in B_2 \} = Y_1^{-1}(B_1) \cap Y_2^{-1}(B_2),
\]

which is a set in \( \mathcal{G} \). Thus the preimage of any product set is in \( \mathcal{G} \) and hence, by Lemma 6, \( (Y_1, Y_2) \) is measurable.

**Proof of Proposition 2**

Let \( Y \) be a continuous function from \( R \) into \( R \). And consider the preimage \( Y^{-1}((a, b)) \) of an open interval \( (a, b) \). Let \( s \) be a point in the preimage. Then continuity gives: for any interval \( (Y(s) - \epsilon, Y(s) + \epsilon) \), there is a \( \delta \) such that \( (s - \delta, s + \delta) \) is mapped into \( (Y(s) - \epsilon, Y(s) + \epsilon) \). With \( \epsilon \) sufficiently small, this says that \( Y^{-1}((a, b)) \) contains an open interval about the point \( s \). Thus \( Y^{-1}((a, b)) \) is open. The Borel class is the smallest \( \sigma \)-algebra containing the open intervals, or containing the open sets; hence, by Lemma 6, \( Y \) is measurable. Similarly for \( R^k \) to \( R^l \).

We now consider the proof of Proposition 5 for real-valued functions; the proof for vector-valued functions then follows from Proposition 3. Actually, it is convenient to be somewhat more general and let \( Y \) and even the \( Y_i \)'s be extended real-valued functions taking real values or values \(+\infty\) and \(-\infty\). The Borel sets extend also in an obvious manner, and for measurability it suffices (Lemma 6) to verify that \( Y^{-1}((c, \infty)) \) is in \( \mathcal{G} \) for all real \( c \). The proof needs two preliminary results:

**Lemma 7**

If \( Y_1, Y_2, \ldots \) are measurable, then \( W = \sup Y_i \) is measurable; also \( \inf Y_i \) is measurable.

**Proof**

We have

\[
W(s) = \begin{cases} 
(\sup Y_i(s) \in (c, \infty)] & \text{for some } i \\
(\inf Y_i(s) \in (c, \infty)] & \text{for some } i \\
(\bigcup_i Y_i^{-1}((c, \infty)) & \text{for some } i
\end{cases}
\]

But each \( Y_i^{-1}((c, \infty)) \) is in \( \mathcal{G} \); hence \( \bigcup_i Y_i^{-1}((c, \infty)) \) is in \( \mathcal{G} \); thus \( W \) is measurable.

**Lemma 8**

If \( Y_1, Y_2, \ldots \) are measurable, then \( X = \lim \sup Y_i \) is measurable; also, \( \lim \inf Y_i \) is measurable.

**Proof**

The \( \lim \sup \) or upper limit point of a sequence of real numbers \( a_1, a_2, \ldots \) is a value \( a \) such that any \( a + \delta (\delta > 0) \) is exceeded by a finite number of the \( a_i \), and any \( a - \delta (\delta > 0) \) is
exceeded by a countable number of the $a_i$, provided that such an $a$ exists; otherwise, it is $+\infty$. Formally,

$$\limsup_{i \to \infty} Y_i(s) = \inf_{j=1}^{\infty} \sup \{ Y_j(s), Y_{j+1}(s), \ldots \}.$$ 

The lim sup for a sequence of functions is then given by

$$\limsup_{i \to \infty} Y_i = \inf_{j=1}^{\infty} \sup \{ Y_j, Y_{j+1}, \ldots \}.$$ 

By Lemma 7 we have that $\sup \{ Y_j, Y_{j+1}, \ldots \}$ is measurable. And then using the analogous result for inf, we have that $\limsup Y_i$ is measurable.

**Proof of Proposition 5** If the limit $\lim Y_i$ exists, then $\lim Y_i = \lim \inf Y_i = \limsup Y_i$ and these are measurable by Lemma 8.

15 Let $Y$ be a real-valued measurable function. Prove:
   (a) $Y$ is the uniform limit of a sequence of discrete functions.
   (b) $Y$ is the limit of a sequence of simple functions.

16 Let $Y_1$ and $Y_2$ be measurable functions from $\mathcal{S}$ into $\mathbb{R}$. Use Propositions 1, \ldots, 5 to prove that $Y_1 + Y_2$ is measurable.

17 Let $\mu$ be a measure on the $\sigma$-algebra $\mathcal{G}$ of subsets of $\mathcal{S}$; let $Y$ be a measurable function from $\mathcal{S}$ into $\mathbb{R}$ with class $\mathcal{G}$ of subsets; and let $\mu_Y$ be defined by

$$\mu_Y(\emptyset) = \mu(Y^{-1}(\emptyset)).$$

Now let $\mathcal{S}_0 = \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$; let $\mu$ be the modified cardinality measure:

$$\mu(A) = c(A \cap \mathcal{S}_0)$$

for $A \subset \mathbb{R}^2$; and let $Y$ be the sum function given by $y = Y(x_1, x_2) = x_1 + x_2$. Calculate $\mu_Y$ and compare with $P_Y$ in Exercise 3.1.2.

### 4.3

**CONDITIONAL PROBABILITY**

In science, engineering, and business, there are situations where we want to record a response at a sequence of time points or even record it continuously in time. For this, then, we need not a finite number of coordinates yielding a point in $\mathbb{R}^n$ but a countable number or a continuum of coordinates. The space is more complicated, but the probability model is easily constructed; for some formalities, see Supplement Section 3.6F. We now examine briefly some probability models having a countable number or a continuum of coordinates; conditional probability is used directly or implicitly in the construction of the models. These are examples of stochastic processes and they give an indication of the breadth and importance of this area of probability theory.

For notation we let $s = \{s_t\}$ designate the combined response. And let $s_t$ designate the response at index $t$ or at time $t$; and $\mathcal{S}_t$ designate the corresponding sample space. We call the index time even though in some applications it may refer to space and even take values in $\mathbb{R}^2$, or $\mathbb{R}^3$, or more generally. In accord with this, we call $s_t$ the state of the system at time $t$ and call $\mathcal{S}_t$ the state space at time $t$. A realized response $s = \{s_t\}$ is then a function $s$ which records the state of the system for various times $t$. From Supplement Section 3.6F we have available the probability theory to describe the distribution of such a function $s$. 
A POISSON PROCESSES

Suppose that we start observing a radioactive disintegration and keep a cumulative record of the number of particles observed; let \( n_t \) be the number of particles in the interval \( (0, t] \). We might then obtain a record as illustrated in the first half of Figure S.7; this is an example of a Poisson process. Or in a more general sense, suppose that we keep a record of the number \( n_t \) of people alive in a society at time \( t \). As \( t \) increases, a function \( n_t \) would increase at the time of a birth and decrease at the time of a death, as indicated in the second half of Figure S.7; this is an example of a birth-and-death process. We examine some theory concerning Poisson processes.

Suppose that a certain kind of event can occur uniformly and at random in time. Let \( n_t \) designate the number of occurrences in the time interval \((0, t]\); then \( \{n_t\} \) is a Poisson process. We present this formally:

**Definition 1**

The process \( \{n_t\} \) is a Poisson process if the following hold:

(a) Independent increments: for \( 0 < t_1 < t_2 < \ldots < t_k \), the increments \( n_{t_k} - n_{t_{k-1}} \) are statistically independent.

(b) Homogeneous increments: the distribution of the increment \( n_{t+h} - n_t \) with \( h > 0 \) does not depend on \( t \).

(c) No cascade: let \( p_t(i) \) be the probability \( P(n_t = i) \); then as \( t \to 0 \),

\[
\frac{p_t(1)}{t} \to \lambda, \quad \frac{1 - p_t(0)}{t} \to \lambda, \quad \lambda > 0.
\]

Note that condition (c) says that the probability of one event in an interval is proportional to the length \( t \) of the interval as \( t \to 0 \); and that the probability of one or more event in an interval is the same proportion of the length \( t \) as \( t \to 0 \); that is, the probability of more than one event, as a proportion of \( t \), goes to zero as \( t \to 0 \).

The homogeneity and independence give

\[ n_t \]

\[ w_1 \quad w_2 \quad w_3 \quad w_4 \quad t \]

**FIGURE S.7**
Poisson process; and birth-and-death process.
\( p_{t+h}(0) = p_{h}(0)p_{t}(0). \)

Thus
\[
\frac{p_{t+h}(0) - p_{t}(0)}{h} = \frac{p_{h}(0) - 1}{h} p_{t}(0),
\]

\[
\frac{dp_{t}(0)}{dt} = -\lambda p_{t}(0),
\]

\[
p_{t}(0) = e^{-\lambda t},
\]

where an integration constant has been determined by the necessary \( p_{0}(0) = 1. \) In a similar way we obtain
\[
\frac{p_{t+h}(i) - p_{t}(i)}{h} = \frac{p_{h}(0) - 1}{h} p_{t}(i) + \frac{p_{h}(1)}{h} p_{t}(i-1) + \cdots + \frac{p_{h}(i)}{h} p_{t}(0),
\]

\[
\frac{dp_{t}(i)}{dt} = -\lambda p_{t}(i) + \lambda p_{t}(i-1),
\]

\[
\frac{d}{dt} (p_{t}(i)e^{\lambda t}) = \lambda(p_{t}(i-1)e^{\lambda t}).
\]

We then obtain successively \( p_{t}(1), p_{t}(2), \ldots \) using the necessary \( p_{0}(i) = 0 (i > 0): \)
\[
p_{t}(i) = \frac{(\lambda t)^{i}}{i!} e^{-\lambda t}.
\]

Thus for any particular \( t \) the distribution of \( n_{t} \) is Poisson(\( \lambda t \)).

Now let \( w_{1}, w_{2}, \ldots \) be the time of the first, second, \ldots occurrence of the event. And let \( t_1, t_2, \ldots \) be the interoccurrence times:
\[
t_1 = w_1, \ t_2 = w_2 - w_1, \ t_3 = w_3 - w_2, \ldots
\]

The probability \( P(t_1 > t) \) is the probability that the first occurrence time exceeds \( t \), or equivalently that there are no occurrences in the interval \( (0, t] \); thus
\[
P(t_1 > t) = e^{-\lambda t}; \quad P(t_1 \leq t) = 1 - e^{-\lambda t},
\]

and we see that \( t_1 \) has the exponential(\( \lambda \)) distribution or, equivalently, \( \lambda t_1 \) has the standard exponential distribution.

The homogeneity and independence then show that \( (\lambda t_1, \ldots, \lambda t_r) \) is a sample from the standard exponential; thus
\[
\lambda w_r = \lambda t_1 + \cdots + \lambda t_r
\]
has the gamma(\( \eta \)) distribution: see the reproductive property for the gamma distribution as presented in terms of the chi-square distribution in Exercise 3.3.5.

The probability \( P(w_r > t) \) is the probability that the \( r \)th occurrence time exceeds \( t \) or, equivalently, that there are less than \( r \) occurrences in the interval \( (0, t] \); thus
\[
\int_{t}^{\infty} \frac{1}{\Gamma(\eta)} (\lambda u)^{r-1} e^{-\lambda u} du = P(w_r > t) = P(n_r < t) = \sum_{i=0}^{r-1} \frac{(\lambda t)^{i}}{i!} e^{-\lambda t},
\]

which gives a direct probability verification of the gamma-Poisson relation in Problem 3.5.33.

**B MARKOV PROCESSES**

Consider a system, say an atom, that can be in one of three states \( s = 1, 2, 3 \); let \( s_t \) be the state at time \( t \). And suppose the system is without memory, so that the probabilities for time \( t + h \)
(h > 0) given the state at time t do not depend on the states at times earlier than t. We might then obtain a record as illustrated in the first half of Figure S.8; this is an example of a discrete Markov process.

Or consider the vertical elevation $y_i$ of a particle in a fluid where changes in $y_i$ are due to impacts by adjacent molecules. We might then obtain a record as illustrated in the second half of Figure S.8; this is an example of a Brownian process, a particular kind of continuous Markov process. We examine briefly some theory concerning Markov processes.

Consider a stochastic process $y = \{y_t: t \in J\}$, where the index $t$ can range over the integers $\mathbb{Z}$ or the reals $\mathbb{R}$.

**DEFINITION 2**

The process $\{y_t\}$ is a Markov process if the conditional distribution of $y_{t_k}$ given $(y_{t_1}, \ldots, y_{t_{k-1}})$ with $t_1 < t_2 < \cdots < t_k$ is equal to the conditional distribution given $y_{t_{k-1}}$ only.

In other words, the probability for a future state given the present state is not altered by information concerning past states. A Markov process is called homogeneous if the distribution of $y_{t+h}$ given $y_t$ with $h > 0$ is independent of $t$. The Poisson process and certain birth-and-death processes are examples of Markov processes.

For some theory, consider the special case of a homogeneous Markov process with finite state space $\{1, 2, \ldots, s\}$ and with discrete time space $\{0, 1, 2, \ldots\}$. Let $p_t(i, j)$ be the probability of going to state $j$ at time $t + 1$ from state $i$ at time $t$; this can be presented as the one-step transition matrix

$$P = \begin{pmatrix} p_1(1, 1) & \cdots & p_1(1, s) \\ \vdots & \ddots & \vdots \\ p_1(s, 1) & \cdots & p_1(s, s) \end{pmatrix}$$

And similarly, let $p_n(i, j)$ be the probability of going to state $j$ at time $t + n$ from state $i$ at time $t$; this can be presented as an $n$-step transition matrix

$$P_n = \begin{pmatrix} p_n(1, 1) & \cdots & p_n(1, s) \\ \vdots & \ddots & \vdots \\ p_n(s, 1) & \cdots & p_n(s, s) \end{pmatrix}$$

**FIGURE S.8**

Discrete Markov process; and Brownian process (a kind of Markov process).
Now consider going from state \( i \) at time \( t \) to state \( j \) at time \( t + 2 \). Let \( k \) designate the possible states \( 1, \ldots, s \) at time \( t + 1 \); then

\[
\rho_2(i,j) = \sum_{k=1}^{s} p_1(i,k)p_1(k,j).
\]

Note that this is the summation formula for matrix multiplication; thus

\[
P_2 = P^2.
\]

And by the same method

\[
P_n = P^n;
\]

thus the \( n \)-step transition matrix is the \( n \)th power of the one-step matrix.

In certain cases a transition matrix \( P \) can be factored \( P = ADB \), where \( D \) is a diagonal matrix,

\[
D = \begin{pmatrix}
d_1 & 0 \\
0 & \ddots \\
0 & 0 & d_s
\end{pmatrix}
\]

and \( B \) and \( A \) record the left and right eigenvectors with \( BA = I \). The \( n \)-step transition matrix is then given immediately,

\[
P_n = P^n = ADB \cdot ADB \cdot \ldots \cdot ADB
\]

\[
= A D^n B
\]

by taking the \( n \)th power of the diagonal elements of \( D \).

### C NORMAL PROCESSES

For an electronic system, consider the voltage \( v_i \) at a certain point as a function of time \( t \); this gives a function on the line \( R \). Or consider the height \( h_i \) of the ocean surface as a function of location \( y \) along a meridian at time \( t \); this gives a function over the plane \( R^2 \).

When many small statistically independent components are added together, the resulting sum has a distribution that is approximately normal (this will be examined formally in Section 6.2). For a variety of systems the response has this property of being the resultant of many small component sources of variation, and accordingly the response is found to be approximately normal.

Consider the voltage \( v_i \) at a point in a small electronic system; without stimulation electrons can be discharged from the cathode of a vacuum tube at random in accord with a Poisson process; the passage of an electron to the anode can induce a transient voltage fluctuation; the superposition of many such voltage fluctuations can produce a voltage \( v_i \) such that \( v_i, \ldots, v_{i_n} \) has a multivariate normal distribution. This leads us to examine normal or Gaussian processes.

**DEFINITION 3**

The process \( \{y_i : t \in R\} \) is a normal process if \( \{y_{i_1}, \ldots, y_{i_n}\} \) has a multivariate normal distribution for each \( (t_{i_1}, \ldots, t_{i_n}) \).

A normal process is called homogeneous if the distribution of \( \{y_{i+t}, \ldots, y_{i+n}\} \) is independent of \( h \). A normal process is called a Brownian process if it has

1. Independent increments: for \( t_1 < t_2 < \ldots < t_n \), the increments \( y_{t_2} - y_{t_1}, \ldots, y_{t_n} - y_{t_{n-1}} \) are statistically independent.

2. Homogeneous symmetrical increments: the distribution of \( y_{t+h} - y_t \) with \( h > 0 \) is symmetrical about 0 and is independent of \( t \).
Consider briefly the description for the Brownian process suggested by the second part of Figure S.8. Let $\gamma_0 = 0$ and let $\sigma$ be the scaling of the normal distribution of $\gamma_1$. The distribution of $\gamma_n$ is then the sum of $n$ independent normal$(0, \sigma)$ variables and thus has a normal$(0, \sqrt{n}\sigma)$ distribution; see Problem 3.3.11. It then follows from a continuity argument that $\gamma_t$ has a normal $(0, \sqrt{t}\sigma)$ distribution.

**D THE POLYA URN**

We conclude with a simple model involving colored balls in an urn, the Polya urn. The model has application in the study of occurrences of a contagious disease and in the distribution of plants and animals over area.

An urn contains $b$ black balls and $r$ red balls. The balls are thoroughly mixed, one ball is drawn, its color is observed, and it is replaced in the urn together with $s$ balls of the same color and $d$ balls of a different color. The procedure is repeated successively.

Let $B_i$ and $R_i$ be the occurrence of a black ball and a red ball, respectively, on the $i$th draw. Then,

$$P(B_i) = \frac{b}{b + r}, \quad P(R_i) = \frac{r}{b + r}.$$ 

And, for example, for black on the second draw given the color on the first draw,

$$P(B_2; B_1) = \frac{b + s}{(b + s) + (r + d)}, \quad P(R_2; R_1) = \frac{b + d}{(b + d) + (r + s)}.$$ 

Now let $E_x$ be the event $x$ black balls in $n$ draws from the urn. The calculation of $P(E_x)$ can be very complicated.

Consider a simple case called the Polya urn: $s > 0, d = 0$. Then, for example,

$$P(B_1 \cdots B_{x+1} \cdots R_n)$$

$$= \frac{b(b + s)(b + (x - 1)s)r(r + s)\cdots(r + (n - x + 1)s)}{(b + r)(b + r + s)\cdots(b + r + (n - 1)s)}$$

$$= \frac{b}{s} \left(\frac{b}{s} + 1\right) \cdots \left(\frac{b}{s} + (x - 1)\right) \frac{r}{s} \left(\frac{r}{s} + 1\right) \cdots \left(\frac{r}{s} + n - x - 1\right)$$

$$= \left(\frac{b}{s}\right)^{[x]} \left(\frac{r}{s}\right)^{[n-x]} \left(\frac{b + r}{s}\right)^{[n]}$$

where $M^{[n]} = M(N + 1) \cdots (N + n - 1)$. The probability for any other ordering of $x$ $B$'s and $n - x$ $R$'s involves only a rearrangement of the factors above; thus $P(E_x)$ is $\binom{n}{x}$ times the preceding expression:

$$P(E_x) = \frac{\binom{b}{s}^{[x]} \binom{r}{s}^{[n-x]} \frac{n!}{x!(n-x)!}}{\left(\frac{b + r}{s}\right)^{[n]} \left(\frac{-b/s}{x}\right) (-r/s)^{n-x}} = \frac{\binom{(-b/s)(-r/s)}{x}}{\binom{(-b + r)/s}{n}}.$$ 

These urn models have been used to describe the number of occurrences of a contagious disease: an incidence of the disease can infect others and alter the population.

**E EXERCISE**

1. A player throws a pair of symmetrical dice until he gets 3 points or 7 points; let $n$ be the number of tosses. He wins if he gets 3 points; he loses if he gets 7 points.
(a) Determine the probability that \( n \) is even; that \( n \) is a multiple of 3.
(b) Show that \((\text{Win}, \text{Lose})\) is statistically independent of \( n \).

### F PROBLEMS

2. **The Poisson(\( \lambda \)) process:** Show that the distribution of \( w_1, \ldots, w_n \) is given by

\[
 f(w_1, \ldots, w_n) = \lambda^n \exp\left(-\lambda w_n\right) \quad \text{if} \quad 0 < w_1 < \cdots < w_n
\]

and zero otherwise.

3. **The Poisson(\( \lambda \)) process given the value \( n \) at a specified time \( t \):** Show that the conditional distribution of \( (w_1, \ldots, w_n) \) given \( n = n \) is the distribution of the order statistic \((u_{(1)}, \ldots, u_{(n)})\) of a sample from the uniform(0, \( \theta \)) distribution.

4. **(continuation) The Poisson(\( \lambda \)) process given the time of the \((n + 1)\)st occurrence:** Show that the conditional distribution of \( (w_1, \ldots, w_n) \) given \( w_{n+1} = t \) is the order statistic distribution described in Problem 3.

5. **(continuation) For simplicity take \( \lambda = 1 \). Calculate expressions for**

\[
P(w_r > \rho; w_{n+1} = t) = P(n_p < r; w_{n+1} = t)
\]

and obtain the beta-binomial relation in Problem 3.5.28. Use the order statistic theory in Supplement Section 3.6.E.

6. Calculate the two-step transition matrix for the Markov process with one-step matrix

\[
P = \begin{pmatrix}
0.8 & 0.1 & 0.1 \\
0 & 0.7 & 0.3 \\
0 & 0.6 & 0.4
\end{pmatrix}
\]

Let \( p_i(t) \) be the probability of state \( i \) at time \( t \). If \( (p_0(1), p_0(2), p_0(3)) = (0.4, 0.3, 0.3) \), determine \( (p_1(1), p_1(2), p_1(3)) \) and \( (p_2(1), p_2(2), p_2(3)) \). What will the value of \( p_n(1) \) be for large \( n \)?

7. A symmetric coin is tossed at times \( t = 0, 1, 2, \ldots \). Let \( n_i \) be the number of heads at time \( t \). Treat \( (n_i) \) as a Markov process and calculate the one-step transition probability \( p_i(i, j) \) from state \( i \) at time \( t \) to state \( j \) at time \( t + 1 \). Present this as a matrix with rows 0, 1, 2, \ldots and columns 0, 1, 2, \ldots. Calculate the two-step transition matrix.

8. Consider a Poisson(\( \lambda \)) process \( (n_i) \). Treat \( (n_i) \) as a Markov process and calculate the transition probability \( p_i(i, i+k) \) from state \( i \) at time \( t \) to state \( i+k \) at time \( t+h \). Present this as a transition matrix \( P_h \).

9. Consider a state space consisting of two points; the general homogeneous Markov process can be represented by the one-step transition matrix

\[
P = \begin{pmatrix}
1 - \beta & \beta \\
\alpha & 1 - \alpha
\end{pmatrix}
\]

(a) Show that for \( \alpha + \beta \neq 0, 1 \):

\[
P = \begin{pmatrix}
1 & -\beta \\
\alpha & \alpha
\end{pmatrix} \begin{pmatrix}
1/\alpha + \beta \\
0
\end{pmatrix} \begin{pmatrix}
1/\alpha + \beta & 1 \\
0 & 1/\alpha + \beta
\end{pmatrix} \begin{pmatrix}
1 & \beta \\
0 & 1
\end{pmatrix}
\]

and calculate the \( n \)-step transition matrix \( P_n \).
(b) Determine \( \lim_{n \to \infty} P_n \).

(c) Examine the case \( \alpha + \beta = 1 \), also the case \( \alpha + \beta = 0 \).

10. Let \( t \) be an index with range \( J = \{1, 1/2, 1/2^2, 1/2^3, \ldots\} \). Let \( S_t = (0, 1) \), \( B_t \) be the Borel class on (0, 1), and \( P_t \) be the uniform distribution on (0, 1). Consider \( (\Pi S_t, \Pi B_t, \Pi P_t) \) and set \( s_t = \{s_t \} = \{s_t : t \in J\} \) designate a typical outcome.

(a) Calculate the probability that the function \( s_t \) is less than the function \( e^{-t} \); that is, \( P(s_t < e^{-t} \text{ for all } t) \).
(b) Calculate the probability \( P(s_t < e^{at} \text{ for all } t) \), where \( a < 0 \).

(c) Determine the distribution of \( \nu = \sup (t^{-1} \ln s_t) \).

5.8

MEAN VALUE

We have noted that a moment-generating function uniquely determines a corresponding distribution. This uniqueness or unicity result is a very powerful tool in theoretical analyses. It is also useful for routine calculations with distributions provided the mathematics is tractable—that is, density functions involving powers, exponentials, logarithms, and trigonometric functions. In the first sections of this supplement we establish the uniqueness results for moment-generating functions, characteristic functions, and moment sequences.

The formal theory for mean values needs the concept of the general integral. In Section 5.1 we briefly described the general integral, then in subsequent sections used ordinary summations and integrations. In the later sections of this supplement we develop the basic concepts and properties for the general integral.

A THE UNIQUENESS THEOREM

FOR DISCRETE DISTRIBUTIONS

Consider a discrete distribution for \( x \) on the real line \( R \) and suppose that all the probability is on the set \( Z \) of integers. Let \( \rho \) be the probability function and \( c \) be the corresponding characteristic function. Then

\[
(1) \quad c(t) = E(e^{ixt}) = \sum_{x=-\infty}^{\infty} e^{ixt} \rho(x).
\]

The function \( e^{ixt} = \cos yt + i \sin yt \) is a periodic function of \( t \) for \( y \) an integer; we have

\[
(2) \quad \int_{-\pi}^{\pi} e^{ixt} \, dt = 2\pi \quad \text{if } y = 0
\]

and zero if \( y = \pm 1, \pm 2, \ldots \). Now multiply both sides of (1) by \( e^{-iyt} \) and integrate from \(-\pi\) to \( \pi\):

\[
\int_{-\pi}^{\pi} e^{-iyt} c(t) \, dt = \int_{-\pi}^{\pi} \sum_{x=-\infty}^{\infty} e^{itx - yt} \rho(x) \, dt.
\]

The integrand on the right has \( |e^{itx - yt}| \leq 1 \), and it is integrable with respect to \( x \) on \( Z \) and \( t \) on \( R \). Thus by the Fubini Proposition 5.8.10 we can interchange the order of summation and integration:

\[
(3) \quad \int_{-\pi}^{\pi} e^{-iyt} c(t) \, dt = \sum_{x=-\infty}^{\infty} \int_{-\pi}^{\pi} e^{itx - yt} \rho(x) \, dt = 2\pi \rho(y)
\]

for \( y \) equal to an integer. This gives us the inversion formula,

\[
(4) \quad \rho(y) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-iyt} c(t) \, dt
\]

and from it we can note that a distribution on the integers is uniquely determined by its characteristic function.
THE GENERAL UNIQUENESS THEOREM

Consider a distribution for \( x \) on the real line \( R \). Let \( P \) be the probability measure and \( c \) be the characteristic function:

\[
(5) \quad c(t) = \int_{-\infty}^{\infty} e^{ixt} \, dP(x).
\]

Let \( P^x \) be some probability measure on \( R \) and \( c^x \) be the corresponding characteristic function. We multiply the preceding equation by \( e^{-iyt} \) and integrate using \( P^x \) as the probability measure for \( t \):

\[
\int_{-\infty}^{\infty} e^{-iyt} c(t) \, dP^x(t) = \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} e^{ixy} \, dP(x) \right) \, dP^x(t).
\]

The integrand on the right has \( |e^{ixy}| \leq 1 \) and is integrable with respect to the product distribution for \( x \) and \( t \). Thus by the Fubini Proposition 5.8.10, we can interchange the order of integration:

\[
(6) \quad \int_{-\infty}^{\infty} e^{-iyt} c(t) \, dP^x(t) = \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} \frac{e^{ixy}}{P^x(t)} \, dP(x) \right) \, dP^x(t) = \int_{-\infty}^{\infty} e^{i(x - y)} \, dP(x).
\]

This is called the Parseval relation; a simpler and more symmetrical version is obtained with \( y = 0 \).

Now consider a fruitful choice for \( P^x \). The normal distribution has the remarkable property that its density and its characteristic function have the same functional form. Thus for the central normal with variance \( \sigma^2 \) and the normal with variance \( 1/\sigma^2 \), we have Table S.1. We multiply (6) by \( 1/\sqrt{2\pi\sigma} \) and use the second of these distributions for \( P^x \):

\[
(7) \quad \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iyt} c(t)e^{-\sigma^2 t^2/2} \, dt = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{-tx - \sigma^2 t^2/2 \alpha^2} \, dP(x).
\]

The remarkable property that we noted for the normal has produced a density function in the integrand on the right side. For interpretation think of \( x \) with distribution \( P \) and \( u \) with the normal(0, \( \sigma \)) distribution. The right side then is a slight generalization of the convolution formula \((3.2.25)\) as applied to give the density function for \( y = x + u \); note that \( y \) does have a density even if \( x \) does not. Let \( f_o \) be the density for \( y \), and \( F_o \) be the corresponding distribution function; then

\[
f_o(y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iyt} c(t)e^{-\sigma^2 t^2/2} \, dt,
\]

\[
F_o(y_2) - F_o(y_1) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-iy_2 t} - e^{-iy_1 t}}{-it} \, c(t)e^{-\sigma^2 t^2/2} \, dt.
\]

**TABLE S.1**

<table>
<thead>
<tr>
<th>Density Function ((t))</th>
<th>Characteristic Function ((u))</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{1}{\sqrt{2\pi\sigma}} e^{-t^2/2\sigma^2} )</td>
<td>( e^{-\sigma^2 u^2/2} )</td>
</tr>
<tr>
<td>( \frac{\alpha}{\sqrt{2\pi}} e^{-\alpha^2 t^2/2} )</td>
<td>( e^{-\sigma^2 t^2/2} )</td>
</tr>
</tbody>
</table>
where we have used \( \int_{\mathcal{F}} e^{it} \, d\mu = (e^{it_1} - e^{it_1})/A \) with the Fubini Proposition 5.8.10. Our goal has been to obtain the distribution function for \( x \). In fact, we have obtained the distribution function for \( y = x + u \), where \( u \) is normal(0, \( \sigma \)). If \( \sigma \) is small, the effect of \( u \) should be small. Indeed, our study of limits in Section 6.1 will show that we can obtain the distribution function \( F \) for \( x \) by letting \( \sigma \to 0 \). This gives us the inversion formula,

\[
F(y) - F(y_i) = \lim_{\sigma \to 0} \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-iy_1} - e^{-iy_1}}{-it} c(t) e^{-\frac{t^2}{2}} \, dt
\]

for points \( y_i \) of continuity for the function \( F \); and we can note from this that a distribution on the real line is uniquely determined by its characteristic function.

Now suppose that \( c \) is integrable with respect to length on the real line. Then by the Proposition 5.8.8 we can take the limit inside the integral sign on the right side of (8):

\[
F(y) - F(y_i) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-iy_1} - e^{-iy_1}}{-it} c(t) \, dt.
\]

The density function \( f \) is then given (use Proposition 5.8.10) by the inversion formula,

\[
f(y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iyt} c(t) \, dt.
\]

Let \( \Phi \) be the set of all probability measures \( P \) on the real line. And let \( \mathcal{C} \) be the set of all characteristic functions of distributions on the real line. The defining equation (5) gives a mapping \( \Phi \to \mathcal{C} \); the equation (8) shows that the mapping is one to one.

**C THE UNIQUENESS THEOREM FOR MOMENT-GENERATING FUNCTIONS**

Now consider a distribution for \( x \) on the real line \( \mathbb{R} \) and suppose the moment-generating function exists. Let \( P \) be the probability measure and \( m \) be the moment-generating function.

\[
m(t) = \int_{-\infty}^{\infty} e^{xt} \, dP(x);
\]

by Definition 5.7.1 the integral exists for an interval \((t_1, t_2)\) about the origin. The analogous function of a complex variable \( w = t + iu \).

\[
m(w) = \int_{-\infty}^{\infty} e^{xt+iwu} \, dP(x)
\]

exists if the real part of \( w \) lies in the interval \((t_1, t_2)\); this follows by noting that

\[
|e^{zt}| = |e^{zt} e^{i\theta u}| = e^{zt}
\]

and using results from Section 1. Complex function theory shows that the function \( m(w) \) is determined by its values along a segment such as the interval \((t_1, t_2)\) on the real axis. Thus the moment-generating function \( m \) determines the function (11), which in turn determines the characteristic function

\[
c(w) = m(iw),
\]

and hence determines the distribution \( P \). Thus, a distribution on the real line is uniquely determined by the moment-generating function.

**D THE UNIQUENESS THEOREM FOR MOMENT SEQUENCES**

Consider a distribution for \( x \) on the real line \( \mathbb{R} \). Let \( P \) be the probability measure and \( m \) be the moment sequence:
\[ m = (E(x^0), E(x^1), E(x^2), \ldots)' = (\mu_0, \mu_1, \mu_2, \ldots)'. \]

If the series
\[ \sum_{r=0}^{\infty} \frac{\mu_r t^r}{r!} \]
converges for some interval containing the origin, then a detailed analysis shows that the characteristic function is determined and thus shows that the distribution is uniquely determined by its moment sequence. Note that the series (12) has the series properties of a moment-generating function; indeed, analysis shows that it is the moment-generating function.

It is possible to construct different distributions that have the same moment sequence. A necessary and sufficient condition for uniqueness of the distribution has been found—that the series
\[ \sum_{r=0}^{\infty} \mu_{2r}^{1/(2r)} \]
diverge.

**E UNIQUENESS ON \( \mathbb{R}^k \)**

Consider a distribution for \( x \) on \( \mathbb{R}^k \). The characteristic function \( c \) given by
\[ c(t) = E(e^{itx}) \]
always exists. A generalization of the methods in Section B show that the distribution is uniquely determined by its characteristic function.

The moment-generating function \( m \) given by
\[ m(t) = E(e^{tx}) \]
is said to exist if the mean exists for an open rectangle containing the origin. A generalization of the methods in Section C shows that the distribution is uniquely determined by its moment-generating function.

The moment array has the \((r_1, \ldots, r_k)\)th moment
\[ \mu_{r_1, \ldots, r_k} = E(x_1^{r_1} \cdots x_k^{r_k}), \quad r_1 = 0, 1, \ldots \]
If the moment-generating type of series (12) converges for some rectangle containing the origin, it can be shown that the distribution is uniquely determined by the moment array.

**F THE INTEGRAL OF A NONNEGATIVE FUNCTION**

The discussion in Section 5.1A lead us to think of the mean
\[ E(Y) = \int Y(s) \, dP(s) \]
as the summation of values of the function \( Y \) weighted by corresponding probabilities. It is easy to be much more general and think of the summation of values of the function \( Y \) weighted by some measure \( \mu \).

Consider a sample space \( \mathcal{S} \), a class \( \mathcal{G} \) of events, and a measure \( \mu \). And let \( g \) be a nonnegative function mapping \( \mathcal{S} \) into \([0, \infty]\). For the moment we are allowing the value \(+\infty\); this, in fact, leads to simpler initial results.

If \( g \) is a simple function, then we define the integral of \( g \) with respect to \( \mu \) by
\[ \int g(s) \, d\mu(s) = \sum a_i \mu(g^{-1}(a_i)) = \sum a_i \mu(a_i), \]

**Sec. 5.8: Mean Value 547**

\[ m = (E(x^0), E(x^1), E(x^2), \ldots)' = (\mu_0, \mu_1, \mu_2, \ldots)'. \]
where the \( a_i \) are the values of the function \( g \) and where \( \mu_0 \) in the final expression is the marginal measure for \( g \); see Problem 3.6.17. This integral is really just a summation; it was presented earlier in Problem 1.5.21 together with two basic properties.

Now consider some general function \( g \) mapping \( \mathbb{S} \) into \([0, \infty]\). The integral of \( g \) with respect to \( \mu \) is defined by

\[
(14) \quad \int g(s) \, d\mu(s) = \sup_{Y \leq g} \int Y(s) \, d\mu(s),
\]

where the supremum is taken over all simple functions \( Y \) having \( 0 \leq Y(s) \leq g(s) \). Note that the integral has a value in \([0, \infty]\) and that we are, in effect, approximating \( g \) from below by simple functions. If we think of \( g \) plotted over the space \( \mathbb{S} \), then the integral can be thought of as the area under \( g \) with respect to a length measure \( \mu \).

In the preceding definition the supremum is taken over all simple functions \( Y \) having \( 0 \leq Y(s) \leq g(s) \). Actually, it suffices to consider simple functions such as \( X(g) \) having \( 0 \leq X(g(s)) \leq g(s) \). For if we have any simple function \( Y \), then we can find a bigger function \( X(g) \).

\[
X(g(s)) = \max \{a_i : a_i \leq g(s)\},
\]

where the \( a_i \) are the values of the function \( Y \). It follows, then, that the integral

\[
(15) \quad \int g(s) \, d\mu(s) = \int_0^\infty Y \, d\mu_0(y)
\]

depends only on values of \( g \) and on the marginal measure for those values; this justifies the right-hand expression.

From the definition of the integral note that if \( g_1 \leq g_2 \), then

\[
\int g_1(s) \, d\mu(s) \leq \int g_2(s) \, d\mu(s).
\]

The integral of \( g \) over a set \( A \) is obtained by the use of indicator functions:

\[
(16) \quad \int_A g(s) \, d\mu(s) = \int I_A(s) g(s) \, d\mu(s).
\]

We, of course, assume that the functions \( g \) are measurable (Definition 3.1.2) and that sets \( A \) are measurable (belong to the \( \sigma \)-algebra \( \mathcal{G} \)).

**G PROPERTIES OF THE INTEGRAL FOR NONNEGATIVE FUNCTIONS**

The integral of a nonnegative function has a very fundamental continuity property:

**THEOREM 1**

**Monotone convergence theorem:** If \( \{g_n\} \) is a monotone sequence of measurable functions \( 0 \leq g_1(s) \leq g_2(s) \leq \cdots \), then

\[
(17) \quad \lim_{n \to \infty} \int g_n(s) \, d\mu(s) = \int \lim_{n \to \infty} g_n(s) \, d\mu(s).
\]

**Proof:** Let \( g(s) = \lim g_n(s) \); the limit exists because of the monotonicity and the inclusion of the value \( +\infty \); \( g \) is measurable by Proposition 3.6.5. Consider the integrals

\[
\int g_n(s) \, d\mu(s) \quad \text{and} \quad \int g(s) \, d\mu(s).
\]

Any simple function that approximates \( g_n \) from below also approximates \( g \) from below; thus the left integral is less than or equal to the right integral, and

\[
\lim_{n \to \infty} \int g_n(s) \, d\mu(s) \leq \int g(s) \, d\mu(s).
\]
Now consider a simple function $Y$ that approximates $g$ from below. The function $g_n$ is above $cY$ $(0 < c < 1)$ on the set

$$A_n = \{s: g_n(s) \geq cY(s)\},$$

and of course

$$\int_{A_n} g_n(s) \, d\mu(s) \geq \int_{A_n} g_n(s) \, d\mu(s) \geq c \int_{A_n} Y(s) \, d\mu(s).$$

We now note that $\lim A_n = S$. For certainly, the sequence $A_1, A_2, \ldots$ is monotone increasing; and if $Y(s) = 0$, then $s \in A_1$ and hence $s \in A_n$; and if $Y(s) > 0$, then $g_n(s) \rightarrow g(s)$ implies that $g_n(s) \geq cY(s)$ for some $n$ and hence $s \in A_n$.

By Problem 1.5.21, the right-hand integral can be viewed as a measure of $A_n$: the monotone property for measures then allows us to take the limit on both sides as $n \rightarrow \infty$:

$$\lim_{n \rightarrow \infty} \int_{A_n} g_n(s) \, d\mu(s) \geq c \int_S Y(s) \, d\mu(s) = c \int Y(s) \, d\mu(s).$$

With arbitrary $c < 1$ and then with arbitrary $Y \leq g$, we obtain

$$\lim_{n \rightarrow \infty} \int g_n(s) \, d\mu(s) \geq \int g(s) \, d\mu(s).$$

This together with the first paragraph gives the equality stated in the theorem.

A function $f$ mapping $S$ into $[0, \infty]$ can be used as a density function for a measure just as such a function has been used as a density for probability.

**PROPOSITION 2**

A measurable function $f$ from $S$ into $[0, \infty]$ produces a new measure $\nu$ from a given measure $\mu$:

$$\nu(A) = \int_A f(s) \, d\mu(s);$$

for brevity we write $d\nu = f \, d\mu$.

---

For a simple function this was presented in Problem 1.5.21. For a general measurable function, the proof uses the Monotone Convergence Theorem; see Problem 6.

**H**

**THE INTEGRAL OF A REAL-VALUED FUNCTION**

Consider a sample space $S$, a class $\mathcal{G}$ of events, and a measure $\mu$. And let $g$ be a measurable function from $S$ into $R$. The function $g$ can be decomposed into positive and negative components

$$g^+(s) = \max \{0, g(s)\}, \quad g^-(s) = \max \{0, -g(s)\},$$

and we have $g = g^+ - g^-$. We then define the integral of $g$ with respect to $\mu$ by

$$\int g(s) \, d\mu(s) = \int g^+(s) \, d\mu(s) - \int g^-(s) \, d\mu(s)$$

provided both terms on the right are finite, and we then say that $g$ is integrable; otherwise we say the integral does not exist or diverges. Note that we no longer allow the value $+\infty$ for an integral. Also note that if $g$ is integrable, then any function $f$ having $|f| \leq |g|$ is also integrable; we say that $f$ is dominated by $g$. 
We can use the result (15) to see that the integral depends only on values of the function $g$ and the marginal measure for those values:

\begin{equation}
\int g(s) \, d\mu(s) = \int_{-\infty}^{\infty} y \, d\mu(y).
\end{equation}

The integral of $g$ over a set $A$ is obtained by the use of indicator functions:

\begin{equation}
\int_{A} g(s) \, d\mu(s) = \int I_A(s) g(s) \, d\mu(s).
\end{equation}

If the measure $\mu$ gives the ordinary length measure on $R$, and if $g$ is continuous, then the integral (19) agrees with the ordinary Riemann integral.

I \hspace{1cm} \textbf{PROPERTIES OF THE INTEGRAL FOR REAL-VALUED FUNCTIONS}

Many of the ordinary properties of the Riemann integral also hold for the general integral. In addition, there are important properties involving limits as indicated in part by the monotone convergence theorem. We now record some of these properties, properties that are useful in probability and statistics: some proofs are included, others are organized as problems in Section J.

\textbf{PROPOSITION 3}

\textit{Linearity in the integrand:} If $g_1$ and $g_2$ are integrable over the set $A$, then

\begin{equation}
\int_{A} (a g_1(s) + b g_2(s)) \, d\mu(s) = a \int_{A} g_1(s) \, d\mu(s) + b \int_{A} g_2(s) \, d\mu(s)
\end{equation}

for real $a$ and $b$.

\textbf{PROPOSITION 4}

\textit{$\sigma$-additive in the domain:} If $g$ is integrable on $\mathfrak{S}$, then

\begin{equation}
\int_{\bigcup_{i=1}^{n} A_i} g(s) \, d\mu(s) = \int_{A} g(s) \, d\mu(s),
\end{equation}

where \{ $A_1, A_2, \ldots$ \} is a partition of $A$; also,

\begin{equation}
\int_{B} g(s) \, d\mu(s) = 0
\end{equation}

if $\mu(B) = 0$.

Thus the integral acts like a generalized measure that takes positive and negative values. The generalized measure depends essentially on $g$; for if

\begin{equation}
\int_{B} g(s) \, d\mu(s) = 0
\end{equation}

for all $B$ in $\mathfrak{S}$, then $g = 0$ except perhaps on a set having $\mu$ measure 0.

\textbf{PROPOSITION 5}

\textit{Linear in the measure:} If $g$ is integrable with respect to $\mu_1$ and $\mu_2$, then
(25) \[ \int_A g(s) \, d(b \mu_1(s) + c \mu_2(s)) = b \int_A g(s) \, d\mu_1(s) + c \int_A g(s) \, d\mu_2(s). \]

where \( b, c \geq 0. \)

**Proof:** The linear combination of measures is clearly a measure. For \( g \geq 0 \) let \( Y \) be a simple function approximating \( g \) from below:

\[ \sum a_i (b \mu_1(Y^{-1}(a_i)) + c \mu_2(Y^{-1}(a_i))) = b \sum a_i \mu_1(Y^{-1}(a_i)) + c \sum a_i \mu_2(Y^{-1}(a_i)). \]

It follows then that (25) holds for \( g \geq 0 \). And then by the decomposition, it holds for general \( g \).

**PROPOSITION 6**

**Absolute value:** For any integrable \( g \)

(26) \[ \left| \int_A g(s) \, d\mu(s) \right| \leq \int_A |g(s)| \, d\mu(s). \]

**Proof:** The left and right sides are given respectively by

\[ \left| \int_A g^+(s) \, d\mu(s) - \int_A g^-(s) \, d\mu(s) \right|, \quad \int_A g^+(s) \, d\mu(s) + \int_A g^-(s) \, d\mu(s). \]

The result (26) then follows from the absolute-value property for real numbers.

**PROPOSITION 7**

A measurable function \( f \) from \( \mathbb{R} \) into \([0, \infty]\) produces a new measure \( \nu \),

\[ \nu(A) = \int_A f(s) \, d\mu(s) \]

and

(27) \[ \int_A g(s) \, d\nu(s) = \int_A g(s) f(s) \, d\mu(s) \]

for \( g \) integrable with respect to \( \nu \).

**PROPOSITION 8**

**Dominated convergence theorem:** If \( g_1, g_2, \ldots \) is a convergent sequence of measurable functions, and if \( |g_n| \leq h \) where \( h \) is integrable with respect to \( \mu \), then

(28) \[ \lim_{n \to \infty} \int_A g_n(s) \, d\mu(s) = \int_A \lim_{n \to \infty} g_n(s) \, d\mu(s). \]

**PROPOSITION 9**

**Dominated derivative theorem:** If \( g(\cdot, t_0) \) is integrable with respect to \( \mu \) and if

\[ \left| \frac{\partial}{\partial t} g(s, t) \right| \leq h(s) \]

for all \( t \) in the interval \((t_0 - \delta, t_0 + \delta)\) for some \( \delta > 0 \) and if \( h \) is integrable with respect to \( \mu \), then

(29) \[ \frac{d}{dt} \int_A g(s, t) \, d\mu(s) = \int_A \frac{\partial}{\partial t} g(s, t) \, d\mu(s) \]

at \( t = t_0 \).
Proof: Apply Proposition 8 to the function 
\[
\frac{g(s, t_0 + \theta) - g(s, t_0)}{\theta} = \frac{\partial}{\partial t} g(s, t) \bigg|_{t_0 + \theta = 0},
\]
where \( \theta \) is given by the mean value theorem.

PROPOSITION 10

Fubini theorem: If \( S_1 \) and \( S_2 \) are sample spaces with classes \( \mathcal{G}_1 \) and \( \mathcal{G}_2 \) of events and with measures \( \mu_1 \) and \( \mu_2 \) and if \( g \) is integrable on \( S_1 \times S_2 \) with respect to \( \mu = \mu_1 \times \mu_2 \) (definitions as in Section 3.6F), then

\[
\int_{S_1 \times S_2} g(s_1, s_2) \, d\mu(s_1, s_2) = \int_{S_1} \left[ \int_{S_2} g(s_1, s_2) \, d\mu_2(s_2) \right] \, d\mu_1(s_1),
\]

where the term in brackets may diverge but only for \( s_1 \) values in a set of \( \mu_1 \)-measure zero. In particular, if \( g \) factors so that the variables separate, then

\[
\int_{S_1 \times S_2} g_1(s_1)g_2(s_2) \, d\mu(s_1, s_2) = \int_{S_1} g_1(s_1) \, d\mu_1(s_1) \cdot \int_{S_2} g_2(s_2) \, d\mu_2(s_2).
\]

Proof: See textbooks on measure theory.

J PROBLEMS CONCERNING THE PROPOSITIONS
IN SECTION I

1. If \( g_1 \) and \( g_2 \) are measurable functions from \( S \) into \([0, \infty]\), show that

\[
\int_S (g_1(s) + g_2(s)) \, d\mu(s) = \int_S g_1(s) \, d\mu(s) + \int_S g_2(s) \, d\mu(s).
\]

Method: Let \( Y_{11}, Y_{12}, \ldots \) be a monotone increasing sequence of simple functions such that \( \lim Y_{1i} = g_1 \) (as in Section 5.1A); and let \( Y_{21}, Y_{22}, \ldots \) be a monotone increasing sequence of simple functions such that \( \lim Y_{2i} = g_2 \); use Problem 1.5.21 and the monotone convergence Theorem 1.

2. (continuation) If \( g(s) = \sum_{n=1}^{\infty} g_n(s) \) is a sum of measurable functions from \( S \) into \([0, \infty]\), show that

\[
\int g(s) \, d\mu(s) = \sum_{n=1}^{\infty} \int g_n(s) \, d\mu(s).
\]

3. (continuation) If \( g(s) = \sum_{n=1}^{\infty} g_n(s) \) is a convergent series of measurable functions from \( S \) into \([0, \infty]\), show that

\[
\int g(s) \, d\mu(s) = \sum_{n=1}^{\infty} \int g_n(s) \, d\mu(s).
\]

Method: Monotone Convergence Theorem 1.

4. (continuation) If \( g_n \) is a sequence of measurable functions from \( S \) into \([0, \infty]\), show that

\[
\int \lim_{n \to \infty} \inf g_n(s) \, d\mu(s) = \lim_{n \to \infty} \int \inf \{ g_n(s), g_{n+1}(s), \ldots \} \, d\mu(s).
\]

5. Fatou lemma: If \( g_n \) is a sequence of measurable functions from \( S \) into \([0, \infty]\), show that

\[
\int \lim_{n \to \infty} \inf g_n(s) \, d\mu(s) \leq \lim_{n \to \infty} \int g_n(s) \, d\mu(s).
\]
6 **Proof of Proposition 2:** Let \( f \) be a measurable function from \( \mathbb{S} \) into \([0, \infty)\). If \( A_1, A_2, \ldots \) is a sequence of disjoint subsets of \( \mathbb{S} \), show that

\[
\int_{\bigcup_{i} A_i} f(s) \, d\mu(s) = \sum_{i} \int_{A_i} f(s) \, d\mu(s).
\]

Method: With integrals over \( \mathbb{S} \), use \( \int_{\bigcup_{i} A_i} \mu = \int_{A_i} \mu \) and Problem 3.

7 If a measurable function \( f \geq 0 \) changes \( \mu \) into \( \nu \) as in Proposition 2, show that

\[
\int_{\mathbb{S}} g(s) \, d\nu(s) = \int_{\mathbb{S}} g(s)f(s) \, d\mu(s)
\]

for any measurable function \( g \geq 0 \). Method: With \( g = I_A \), the equation follows from Proposition 2; prove for a simple function \( g \); then use the Monotone Convergence Theorem to prove for a general \( g \).

8 **Proof of part of Proposition 3:** If \( g \) is integrable over \( A \), show that

\[
\int_{A} ag(s) \, d\mu(s) = a \int_{A} g(s) \, d\mu(s)
\]

for any real \( a \). Method: Examine \( a = 0, >0, <0 \); use the decomposition \( g = g^+ - g^- \).

9 **Proof of part of Proposition 3:** If \( g_1 \) and \( g_2 \) are integrable over \( A \), show that

\[
\int_{A} (g_1(s) + g_2(s)) \, d\mu(s) = \int_{A} g_1(s) \, d\mu(s) + \int_{A} g_2(s) \, d\mu(s).
\]

Method: Let \( g = g_1 + g_2 \), and note that \( g^+ + g_1^+ + g_2^- = g^- + g_1^- + g_2^+ \).

10 **Proof of Proposition 4:** If \( g \) is integrable on \( \mathbb{S} \), then prove that

\[
\int_{\bigcup_{i} A_i} g(s) \, d\mu(s) = \sum_{i} \int_{A_i} g(s) \, d\mu(s),
\]

where the \( A_i \) are disjoint. Method: Use Problem 6 for \( g^+ \) and for \( g^- \).

11 **Proof of the addendum to Proposition 4:** If

\[
\int_{A} g(s) \, d\mu(s) = 0
\]

for all \( A \) in \( \mathcal{A} \), then prove that \( g = 0 \) almost everywhere with respect to \( \mu \). Method: Let \( A_n = \{ s: g(s) > 1/n \} \) and \( A_+ = \{ s: g(s) > 0 \} \); show that \( \mu(A_n) = 0 \) and \( \mu(A_+) = 0 \). Similarly, \( \mu(A) = 0 \).

12 **Proof of Proposition 7:** If a measurable function \( f \geq 0 \) changes \( \mu \) into \( \nu \) as in Proposition 2, show that

\[
\int_{\mathbb{S}} g(s) \, d\nu(s) = \int_{\mathbb{S}} g(s)f(s) \, d\mu(s)
\]

for any measurable function \( g \). Method: Use Problem 7 for \( g^+ \) and for \( g^- \).

13 **Proof of Proposition 8:** If \( g_1, g_2, \ldots \) is a convergent sequence of measurable functions and if \( |g_n| < h \), where \( h \) is integrable with respect to \( \mu \), show that

\[
\lim_{n \to \infty} \int_{\mathbb{S}} g_n(s) \, d\mu(s) = \int_{\mathbb{S}} g(s) \, d\mu(s).
\]

Method: (a) \( |g_n^+| \) and \( |g_n^-| \leq h \); hence \( g_n \) is integrable.

(b) Let \( g = \lim g_n \) and apply Problem 5 to \( 2h - |g_n - g| \), obtaining

\[
\lim_{n \to \infty} \int_{\mathbb{S}} |g_n - g| \, d\mu \leq 0
\]

and hence
\[ \lim_{n \to \infty} \int_S |g_n - g| \, d\mu = 0. \]

(c) Apply Proposition 6.

14 Use the properties of the general integral to prove the properties of the mean as cited in Section 5.1C.

**K EXERCISES AND PROBLEMS**

15 Let \( n \) be the counting measure, where \( n(A) \) is the number of integers in \( A \); see the measure \( \mu_1 \) in Section 1.3B with \( S_0 = \mathbb{Z} \), the set of integers.

(a) Show that the Poisson(\( \lambda \)) distribution is given by

\[ P(A) = \int_A \frac{\lambda^x e^{-\lambda}}{x!(x + 1)} \, dx. \]

(b) Show that the binomial(\( n, \rho \)) distribution is given by

\[ P(A) = \int_A \binom{n}{x} \rho^x (1 - \rho)^{n-x} \, dx. \]

16 If \( X_1, \ldots, X_k \) are integrable with respect to \( \mu \) and if \( |Y| \leq \max \{|X_1|, \ldots, |X_k|\} \), show that \( Y \) is integrable with respect to \( \mu \).

17 **Schwartz inequality:** Let \( X \) and \( Y \) be real-valued measurable functions and suppose that \( X^2 \) and \( Y^2 \) are integrable with respect to \( \mu \).

(a) Show that \( |XY| \leq X^2 + Y^2 \) and hence that \( XY \) is integrable.

(b) Show that

\[ \left( \int X(s)Y(s) \, d\mu(s) \right)^2 \leq \int X^2(s) \, d\mu(s) \cdot \int Y^2(s) \, d\mu(s) \]

with equality if and only if \( c_1X = c_2Y \) (for some \( c_1 \) and \( c_2 \) not both zero) almost everywhere with respect to \( \mu \). Method: Consider the quadratic expression in \( c_1, c_2 \in \mathbb{R} \) given by

\[ \int (c_1X(s) + c_2Y(s))^2 \, d\mu(s). \]

18 Consider a distribution for \( y \) on the real line with distribution function \( F \). Show that

\[ E(y) = \int_0^\infty (1 - F(y)) \, dy - \int_{-\infty}^0 F(y) \, dy. \]

**Method:** Apply the Fubini Proposition 5.8.10 to the following expression:

\[ E(y) = \int_{[0, \infty)} \left( \int_{[0,y]} \, dt \right) dP(u) - \int_{(-\infty,0]} \left( \int_{[y,0]} \, dt \right) dP(u). \]

Recall the simpler version in Problem 5.2.11.

**6.3 LIMITING FUNCTIONS**

In preceding sections of Chapter 6 we have been considering a sequence of distributions on the real line and have examined how the shape of a distribution can converge to a limit. In this section we consider a sequence of functions and examine various definitions for the convergence
of the sequence of functions. We then record the proofs for two theorems in Section 6.1. And in conclusion we examine the Strong Law of Large Numbers.

A LIMITING FUNCTIONS

In Section 6.1 we considered the Poisson($\lambda$) distribution for $\gamma$ and showed that the distribution of $2(\sqrt{\gamma} - \sqrt{\lambda})$ converges to the standard normal distribution. This example has a distribution on $\mathbb{R}$ for each value of $\lambda$ and the shape of the distribution converges to the standard normal as $\lambda \to \infty$; this is convergence in the sense of Definition 6.1.

In Section 6.2 we considered a distribution on $\mathbb{R}^n$ for a sample from a distribution with mean $\mu$ and variance $\sigma^2$ and we showed that the shape of the distribution of $(\sum_{i=1}^n y_i - n\mu)/\sqrt{n\sigma}$ converges to the standard normal as $n \to \infty$.

This second example can be viewed somewhat differently. We could consider a distribution on $\mathbb{R}^n$ for $(y_1, y_2, \ldots)$, where the coordinates are independent and identically distributed with mean $\mu$ and variance $\sigma^2$. The function $(\sum_{i=1}^n y_i - n\mu)/\sqrt{n\sigma}$ is a function of the first $n$ coordinates but can be viewed as a function on $\mathbb{R}^n$. As $n$ varies we obtain a sequence of functions and the shape of the distribution converges to the standard normal as $n \to \infty$.

We now consider several types of convergence for a sequence $(Y_1, Y_2, \ldots)$ of real-valued functions. The functions $Y_i$ are measurable functions defined on a space $\mathcal{S}$ with $\sigma$-algebra $\mathcal{G}$ and probability measure $P$.

B CONVERGENCE IN DISTRIBUTION

The first definition refers only to the marginal distributions of the functions.

DEFINITION 1

A sequence $Y_1, Y_2, \ldots$ converges in distribution to $Y$ designated $Y_n \xrightarrow{D} Y$ if $F_n \to F$ as defined by Definition 6.1.1, where $F_n$ and $F$ are the marginal distribution functions for $Y_n$ and $Y$, respectively.

This definition does not say that a value $Y_n(s)$ is close to the corresponding value $Y(s)$; only that the distribution for $Y_n(s)$ is close to that for $Y(s)$, as discussed in Section 6.1.

C CONVERGENCE IN PROBABILITY

We now consider a definition that does say that a value $Y_n(s)$ is close to $Y(s)$, at least for "most" of the values for $s$.

DEFINITION 2

A sequence $Y_1, Y_2, \ldots$ converges in probability to $Y$ designated $Y_n \xrightarrow{P} Y$ if

\[ \lim_{n \to \infty} \mathbb{P}(|Y_n(s) - Y(s)| \leq \epsilon) = 1 \]

for each $\epsilon > 0$.

An equivalent definition for $Y_n \xrightarrow{P} Y$ is that $Y_n - Y \xrightarrow{D} 0$; this follows from the alternative expression (6.1.10) for convergence to a constant.
Note that if \( Y_n \xrightarrow{d} Y \), then we cannot conclude that \( Y_n - y \xrightarrow{d} 0 \) and thus cannot conclude that \( Y_n \xrightarrow{p} y \).

Convergence in probability is stronger than convergence in distribution:

**Lemma 3**

If \( Y_n \xrightarrow{p} Y \), then \( Y_n \xrightarrow{d} Y \).

**Proof:** Let \( W_n = Y_n - Y \) give the deviation from the limit function; then from (1) we obtain that \( W_n \xrightarrow{d} 0 \). We can then write

\[
Y_n = Y + W_n
\]

and use Theorem 6.1.4 to show that \( Y_n \) and \( Y \) have the same limiting distribution; that is, \( Y_n \xrightarrow{d} Y \).

### D CONVERGENCE IN MEAN(r)

We now consider a definition that says the mean deviation goes to zero.

**Definition 4**

A sequence \( Y_1, Y_2, \ldots \) converges in mean of order \( r (r \geq 1) \) to \( Y \) designated \( Y_n \xrightarrow{(r)} Y \) if

(2) \[ E[|Y_n(s) - Y(s)|^r] \rightarrow 0 \]

as \( n \rightarrow \infty \).

Convergence in mean of order \( r \) is stronger than convergence in probability:

**Lemma 5**

If \( Y_n \xrightarrow{(r)} Y \), then \( Y_n \xrightarrow{p} Y \).

**Proof:** The Chebyshev inequality in Problem 5.2.12 gives

(3) \[ P(|Y_n(s) - Y(s)| > t) \leq \frac{E[|Y_n(s) - Y(s)|^r]}{t^r} \]

for \( t > 0 \). The right side goes to zero, hence the left side also, thus giving convergence in probability.

Also note that convergence in mean of order \( r \) is stronger than convergence in mean of order \( s \) with \( 1 \leq s \leq r \).

### E CONVERGENCE ALMOST SURELY

As a final definition we consider convergence at all sample points except perhaps on a set having probability zero.

**Definition 6**

A sequence \( Y_1, Y_2, \ldots \) converges almost surely to \( Y \), designated \( Y_n \xrightarrow{a.s.} Y \) if
The sequence \( Y_1(s), Y_2(s), \ldots \) may not converge to \( Y(s) \), but the points where it does not converge form a set having probability zero.

To appreciate the definition of almost-sure convergence we should take a very close look at the concept of a limit. Consider sample points that have \( \epsilon \) closeness for the \( i \)th function:

\[
C_i(\epsilon) = \{ s : |Y_i(s) - Y(s)| \leq \epsilon \}.
\]

Now consider sample points that have \( \epsilon \) closeness from the \( n \)th function onward:

\[
C_n(\epsilon) = \{ s : \sup_{j \geq n} |Y_j(s) - Y(s)| \leq \epsilon \} = \bigcap_{j=n}^{\infty} C_j(\epsilon).
\]

Note that \( C_{i1}(\epsilon), C_{i2}(\epsilon), \ldots \) is a monotone increasing sequence of sets. Now consider sample points at which the sequence in the limit is \( \epsilon \) close:

\[
C(\epsilon) = \bigcup_{n=1}^{\infty} C_n(\epsilon) = \lim_{n \to \infty} C_n(\epsilon).
\]

Note that \( C(1), C(1/2), C(1/3), \ldots \) is a monotone decreasing sequence of sets. Now consider sample points at which the sequence converges:

\[
C = \bigcap_{i=1}^{\infty} C(1/i) = \lim_{i \to \infty} C(1/i).
\]

The preceding analysis shows that \( Y_n \xrightarrow{\text{a.s.}} Y \) if and only if \( P(C) = 1 \). Note that \( C \) is defined by countable unions and intersections and thus is measurable; recall Problem 1.5.19.

**F CRITERIA FOR ALMOST-SURE CONVERGENCE**

We now examine several ways of testing for almost sure convergence.

**LEMMA 7**

\( Y_n \xrightarrow{\text{a.s.}} Y \) if and only if

\[
P \left( \sup_{j \geq n} |Y_j(s) - Y(s)| > \epsilon \right) \to 0
\]

as \( n \to \infty \) for any \( \epsilon > 0 \).

**Proof:** From the analysis at the end of Section E we have that \( Y_n \xrightarrow{\text{a.s.}} Y \) if and only if

\[
P(C) = P \left( \lim_{i \to \infty} C(1/i) \right) = \lim_{i \to \infty} P(C(1/i)) = 1.
\]

The sequence \( C(1/1), C(1/2), C(1/3) \) is monotone decreasing; thus \( P(C) = 1 \) if and only if

\[
P(C(1/i)) = 1
\]

for each \( i \). The sequence \( C_{i1}(\epsilon), C_{i2}(\epsilon), C_{i3}(\epsilon), \ldots \) is monotone increasing; thus \( P(C(1/i)) = 1 \) for each \( i \) if and only if

\[
\lim_{n \to \infty} P(C_{i}(1/i)) = 1
\]

for each \( i \); but this is equivalent to condition (9) in the lemma.
We can use the preceding lemma to show that almost-sure convergence implies convergence in probability.

**COROLLARY 8**

If $Y_n \xrightarrow{a.s.} Y$, then $Y_n \xrightarrow{P} Y$.

**Proof:** Note that

\[
C_{\alpha}(\epsilon) = \left\{ s : \sup_{j=\alpha}^{\infty} |Y_j(s) - Y(s)| \leq \epsilon \right\} \subset \left\{ s : |Y_n(s) - Y(s)| \leq \epsilon \right\} = C_{\alpha}(\epsilon).
\]

Almost-sure convergence means that $P(C_{\alpha}(\epsilon)) \rightarrow 1$, while convergence in probability means that $P(C_{\alpha}(\epsilon)) \rightarrow 1$; the lemma follows from the inequality (11).

Almost-sure convergence is more difficult to work with than convergence in probability. The source of this difficulty can be found in the supremum operation in formula (9). One approach to this difficulty lies in the Kolmogorov inequality; see Problems 5.5.8 and 5.5.9 and the Strong Law later in this section. Another possibility lies in the following lemma:

**LEMMA 9**

If $\sum_{i=1}^{\infty} P(|Y_i(s) - Y(s)| > \epsilon)$ is finite, for each $\epsilon > 0$, then $Y_n \xrightarrow{a.s.} Y$.

**Proof:** We can use Lemma 7:

\[
P\left( \sup_{j=\alpha}^{\infty} |Y_j(s) - Y(s)| > \epsilon \right) \leq \sum_{k=\alpha}^{\infty} P(|Y_k(s) - Y(s)| > \epsilon);
\]

the right side is a tail sum of a convergent series and thus goes to zero as $n \rightarrow \infty$. Thus the criterion in Lemma 7 is satisfied.

**COROLLARY 10**

If $\sum_{i=1}^{\infty} E(|Y_i(s) - Y(s)|^r)$ is finite for some $r > 0$, then $Y_n \xrightarrow{a.s.} Y$.

**Proof:** Use the Chebyshev inequality in Problem 5.2.12.

---

**G PROBLEMS**

1. If $Y_n \xrightarrow{(\alpha)} Y$, show that $Y_n \xrightarrow{(\beta)} Y$ for $1 \leq s \leq r$.
2. Let $Y_n(w) = w/n$, where $w$ is standard Cauchy. Show that $Y_n \xrightarrow{a.s.} Y$ does not imply that $Y_n \xrightarrow{(\alpha)} Y$.
3. Let $Y_{2^k,i}(u) = 1$ if $(c - 1)/2^k < u \leq c/2^k$ and $= 0$ otherwise ($c = 1, \ldots, 2^k$), and let $u$ be uniform(0, 1). Show that $Y_n \xrightarrow{(\alpha)} Y$ does not imply that $Y_n \xrightarrow{a.s.} Y$.

**H PROOF OF THE CONTINUITY THEOREM**

The following lemma is needed for the Continuity Theorem 6.1.3.
If $F_n \to F$ in accordance with Definition 6.1.1, then

$$\int_{-\infty}^{\infty} g(y) \, dP_n(y) \to \int_{-\infty}^{\infty} g(y) \, dP(y)$$

as $n \to \infty$ for any bounded continuous function $g$ (the integrations are with respect to the measures $P_n$ and $P$ corresponding to $F_n$ and $F$).

Proof: By using a scale factor we see it is sufficient to prove the theorem for a function having $|g(y)| \leq 1$.

Consider a continuity interval $B$ for the limiting distribution such that there is at most $\varepsilon$ probability in $B^c$. And suppose that we choose $n$ sufficiently large that the distribution $F_n$ assigns at most $2\varepsilon$ probability to $B^c$.

Now subdivide $B$ into component continuity intervals $B_j$ so that the variation of $g$ is less than $\varepsilon$ on each $B_j$. Let $h(y)$ be an approximating discrete function based on the subdivision, and let $h(y) = 0$ on $B^c$. Consider the following integrals:

$$\int_{-\infty}^{\infty} h(y) \, dP_n(y), \quad \int_{-\infty}^{\infty} h(y) \, dP(y).$$

The left (right) integral is within $\varepsilon + 2\varepsilon (\varepsilon + \epsilon)$ of the left (right) integral in (12). But the left integral is a linear combination of the $P_n(B_j)$; and the right integral is the same linear combination of the $P(B_j)$; accordingly, the left integral converges to the right integral. With $\varepsilon$ arbitrary this proves the lemma.

We can extend the lemma and allow $F_n$ to converge to a function $F$ without requiring $F$ to be a distribution function; of course, $F$ will be monotone nondecreasing with values in $[0, 1]$. With this extension, however, we must assume that $g(-\infty) = g(+\infty) = 0$.

We now record the proof for the Continuity Theorem 6.1.3.

Proof: First assume that $F_n \to F$. We have

$$c_n(t) = \int_{-\infty}^{\infty} e^{iyt} \, dF_n(y).$$

The integrand $e^{iyt} = \cos yt + i \sin yt$ is a bounded continuous function; hence by Lemma 11 we have

$$\lim_{n \to \infty} c_n(t) = \int_{-\infty}^{\infty} e^{iyt} \, dF(y) = c(t).$$

Now assume that $c_n(t) \to c(t)$ for real $t$. Given any bounded sequence of real numbers we can extract a subsequence that is convergent. Now consider the sequence $\{F_n\}$ of distribution functions. For a rational number $r_1$ we can extract a subsequence $\{F_{n_k}\}$ that is convergent at $r_1$; for a second rational number $r_2$ we can extract a further subsequence that is also convergent at $r_2$; and so on. We then form a special subsequence by taking the first function in the original sequence, the second in the second sequence, the third in the third, and so on following this diagonalization procedure. We then have a subsequence $\{F_{n_{in}}\}$ which is convergent at the rationals. Let $F$ be the limit function.

The Parseval relation (5.8.7) gives

$$\int_{-\infty}^{\infty} e^{-\sigma^2/2} \, c_{n_{in}}(t) \, dt = \int_{-\infty}^{\infty} e^{-(x-y)^2/2a^2} \, dF_{n_{in}}(x).$$

Let $n \to \infty$ and use the Dominated Convergence Theorem 5.8.8 for the left side and Lemma 11 as extended for the right side.
\[ \int_{-\infty}^{\infty} e^{-is^{2}/2} c(t) \frac{e^{-\sigma t^{2}/2}}{\sqrt{2\pi}} dt = \int_{-\infty}^{\infty} e^{-t^{2}/2a^{2}} F(x). \]

Now let \( a \to \infty \). We obtain

\[ c(0) = F(+\infty) - F(-\infty). \]

But \( c \) is a characteristic function; hence \( c(0) = 1 \) and we obtain that \( F(+\infty) = 1, F(-\infty) = 0. \)

We now use the Uniqueness Theorem in Section 5.8B to show that there can be only one such function \( F \). By the first part of the present theorem, any such limiting \( F \) has the characteristic function \( c \). The Uniqueness Theorem, however, says there is only one such \( F \). Thus all convergent subsequences have the same limit. But this means that the original sequence is convergent to \( F \); for otherwise we could construct a subsequence that converged to something different from \( F \).

I PROOF OF THEOREM 6.1.4

Consider the sample space \( R^{2} \) for \((x_{n}, y_{n})\). Let \( c \) be a point of continuity of \( F \); then for any \( \epsilon > 0 \), there are points of continuity \( c - \delta, c + \delta \) such that

\[ F(c) - F(c - \delta) < \epsilon, \quad F(c + \delta) - F(c) < \epsilon. \]

Consider:

\[ P(y_{n} \leq c - \delta, |x_{n}| < \delta) \leq P(x_{n} + y_{n} \leq c) \leq P(y_{n} \leq c + \delta \quad \text{or} \quad \|x_{n}\| \geq \delta). \]

As \( n \to \infty \), \( P(\|x_{n}\| < \delta) \to 1 \); it follows that the left side has limit \( F(c - \delta) \) and the right side has limit \( F(c + \delta) \). Thus the limiting values for \( P(x_{n} + y_{n} \leq c) \) are in the range \( F(c) \pm \epsilon \).

With arbitrary \( \epsilon > 0 \), this proves that \( \lim P(x_{n} + y_{n} \leq c) = F(c) \).

J THE STRONG LAW OF LARGE NUMBERS

For the Weak Law of Large Numbers we considered a distribution for \( y = (y_{1}, \ldots, y_{n}) \) on \( R^{n} \), where the coordinates \( y_{i} \) are statistically independent and identically distributed with mean \( \mu \). The law says that the marginal distribution of \( \overline{y}_{n} = \Sigma_{i=1}^{n} y_{i}/n \) piles up at the mean \( \mu \) as \( n \to \infty \); that is, \( y_{n} \) converges in distribution to the constant \( \mu \).

In this section we examine a conceptually more incisive question. We envisage sampling from a distribution with mean \( \mu \) obtaining \((y_{1}, y_{2}, \ldots)\); we contemplate the successive sample means \( \overline{y}_{1}, \overline{y}_{2}, \ldots, \) where \( \overline{y}_{n} = \Sigma_{i=1}^{n} y_{i}/n \) and we question whether the sequence \( \overline{y}_{1}, \overline{y}_{2}, \ldots \) has \( \mu \) at its mathematical limit. The Strong Law of Large Numbers says that this mathematical convergence occurs with probability 1. This result was developed from Borel (1909) and Cantelli (1917) to Kolmogorov (1930), giving the Kolmogorov Theorem of the Strong Law of Large Numbers:

THEOREM 12

**Strong law of large numbers:** If \((y_{1}, y_{2}, \ldots)\) is a sequence with coordinates that are statistically independent and identically distributed with mean \( E(y) = \mu \), then \( \overline{y}_{n} = \Sigma_{i=1}^{n} y_{i}/n \) converges to \( \mu \) almost surely; that is, \( \overline{y}_{n} \overset{a.s.}{\to} \mu \).

The Strong Law is proved in three stages: Section K, the Borel–Cantelli lemma, which is a basic means of examining a condition on a countable number of coordinates; Section L, the Kolmogorov lemma, which is essentially a proof of the Strong Law with finite variance; and Section M, the main part of the proof of the Kolmogorov theorem.
**K BOREL–CANTONELLI LEMMA**

Consider a sample space $\mathcal{S}$ with $\sigma$-algebra $\mathcal{G}$ and probability measure $P$.

**Lemma 13**

**Borel–Cantelli lemma:** If $\sum_{n=1}^{\infty} P(A_n)$ is finite for events $A_1, A_2, \ldots$ in $\mathcal{G}$, then the probability that an infinity of the $A_n$ occur is zero, that is,

$$P(\lim \sup A_n) = 0.$$  

**Proof:** The definition of $\lim \sup A_n$ is given in Problem 1.5.16:

$$\lim \sup A_n = \bigcap_{n=1}^{\infty} \bigcup_{j=n}^{\infty} A_j = \lim \bigcup_{n=1}^{\infty} A_n.$$  

Note that $\bigcup_{j=n}^{\infty} A_j$ is monotone decreasing as $n \to \infty$, thus justifying the use of the limit in the expression on the right side. We calculate:

$$P\left(\lim_{n \to \infty} \bigcup_{j=n}^{\infty} A_j\right) \leq P(\bigcup_{j=n}^{\infty} A_j) \leq \sum_{j=n}^{\infty} P(A_j);$$

the right side is a tail sum of a convergent series and thus goes to zero as $n \to \infty$.

**L KOLMOGOROV LEMMA**

Consider a distribution for $(y_1, y_2, \ldots)$ on $\mathbb{R}^n$ with statistically independent coordinates; we do not require that the coordinates have a common distribution.

**Lemma 14**

**Kolmogorov lemma:** If $(y_1, y_2, \ldots)$ is a sequence with statistically independent coordinates and with $E(y_n) = 0$, $\text{var}(y_n) = \sigma_n^2$, and if

$$\sum_{n=1}^{\infty} \frac{\sigma_n^2}{n^2} < \infty,$$  

is finite, then $\overline{y}_n = \sum_{i=1}^{n} y_i / n$ converges to 0 almost surely; that is, $\overline{y}_n \xrightarrow{a.s.} 0$.

**Proof:** To use the criterion for almost-sure convergence in Lemma 7, we must show that

$$P\left(\sup_{j=n}^{\infty} |\overline{y}_n| > \epsilon\right) \to 0$$

for each $\epsilon > 0$ as $n \to \infty$. The procedure is to separate the sequence into batches as follows: $\overline{y}_1, \overline{y}_2, \overline{y}_3, \overline{y}_4, \overline{y}_5, \overline{y}_6, \overline{y}_7, \ldots$, and use the Kolmogorov inequality (Problem 5.5.9) within batches, and the Borel–Cantelli lemma from batch to batch.

Let $S_n = \sum_{i=1}^{n} y_i$; thus $\overline{y}_n = S_n / n$. Let $A_k$ be the event that an $\epsilon$ discrepancy occurs somewhere in the $k$th batch:

$$A_k = \left\{ \left| \frac{S_n}{n} \right| > \epsilon \text{ for some } n \text{ with } 2^{k-1} \leq n < 2^k \right\}$$

$$\subset \left\{ |S_n| > 2^{k-1} \epsilon \text{ for some } n \text{ with } 2^{k-1} \leq n < 2^k \right\}.$$  

We now apply the Kolmogorov inequality (Problem 5.5.9) to the $S_n$ with $2^{k-1} \leq n < 2^k$, and note the trivial $E(S_{2^{k+1}}) \leq E(S_{2^k})$:
We now use the Borel–Cantelli lemma to see if more than a finite number of the $A_k$ occur, that is, to see if the event on the left side of (15) occurs for arbitrarily large values of $n$. For this we check whether the following series is convergent:

$$
\sum_{k=1}^{\infty} P(A_k) \leq \frac{4}{\varepsilon^2} \sum_{k=1}^{\infty} \sum_{n=1}^{2^k} \frac{\sigma_n^2}{2^{2k}} = \frac{4}{\varepsilon^2} \sum_{k=1}^{\infty} \sum_{n \in 2^k} \frac{\sigma_n^2}{2^{2k}}
$$

$$
\leq \frac{4}{\varepsilon^2} \sum_{n=1}^{\infty} \frac{\sigma_n^2}{n^2} \cdot \frac{4}{3} \frac{1}{n^2} = \frac{16}{3\varepsilon^2} \sum_{n=1}^{\infty} \frac{\sigma_n^2}{n^2} < \infty;
$$

the first step uses (17); the second step involves a change in the order of summation; the third step uses the sum of a geometric series with ratio $1/4$; the fourth step involves a rearrangement; and the final step uses (14). Thus by the Borel–Cantelli lemma the condition (15) is satisfied and we have convergence almost surely.

**M KOLMOGOROV THEOREM**

We now prove the Kolmogorov Theorem 12 as presented in Section J. We must show that $\bar{Y}_n \overset{a.s.}{\longrightarrow} \mu$ or $\bar{Y}_n - \mu = 2 \frac{1}{n} \frac{n}{n} \overset{a.s.}{\longrightarrow} 0$. It thus suffices to examine the case with $\mu = 0$.

The method of proof is to use the Kolmogorov Lemma 14. This requires finite variance, so we truncate the initial distributions in accord with position in the sequence; we define

$$
\omega_n = Y_n \quad \text{if} \quad -n < Y_n < n
$$

$$
\omega_n = 0 \quad \text{otherwise};
$$

Let $\ell_n$ be the event $|x_n| > n$ or equivalently the event $\omega_n \neq Y_n$. Note the mean and variance of $\omega_n$:

$$
E(\omega_n) = \mu_n = \int_{-n}^{n} y \, dP(y) \overset{\mu \rightarrow \mu}{\longrightarrow} \mu = 0.
$$

$$
\text{var}(\omega_n) \leq \int_{|y| < n} y^2 \, dP(y) \leq n \int_{j-1 \leq |y| < j} |y| \, dP(y)
$$

$$
= \sum_{j=1}^{n} ja_j,
$$

where $a_j$ is the integral in the second-to-last expression.

We check whether the truncated sequence $(\omega_1, \omega_2, \ldots)$ satisfies the criterion in the Kolmogorov Lemma 14:

$$
\sum_{n=1}^{\infty} \frac{\text{var}(\omega_n)}{n^2} \leq \sum_{n=1}^{\infty} \sum_{j=1}^{n} \frac{ja_j}{n^2} = \sum_{j=1}^{\infty} ja_j \frac{1}{n^2}
$$

$$
< \sum_{j=1}^{\infty} ja_j \frac{2}{j} = 2E(|\mu|) < \infty;
$$

the first step uses (20); the second step involves a change in the order of summation; the third step uses a bound for $\sum_{j} n^{-2}$ obtained from $\int_{y}^{y} y^{-2} \, dy$; the fourth uses the definition of $a_j$; and the fifth step uses the existence of $E(\mu)$. Thus
\[ P \left( \lim \left( \sum_{i=1}^{n} w_i/n - \sum_{i=1}^{n} \mu_i/n \right) = 0 \right) = 1. \]

But \( \mu_i \to 0 \) and thus \( \sum_{i=1}^{n} \mu_i/n \to 0 \): it follows that

\[ P \left( \lim \sum_{i=1}^{n} w_i/n = 0 \right) = 1. \]

We now show that with probability 1 just a finite number of the \( w_i \) are not equal to the corresponding \( y_i \); that is, just a finite number of the events \( L_i \) occur. For this we use the Borel–Cantelli lemma:

\[
\sum_{n=1}^{\infty} P(L_n) = \sum_{n=1}^{\infty} \int_{|y| \geq n} dP(y) \leq \sum_{n=1}^{\infty} \sum_{k=n}^{\infty} \frac{a_{k+1}}{k} = \sum_{n=1}^{\infty} \frac{a_{k+1}}{k} \sum_{k=n}^{\infty} 1 \leq E(|y|) < \infty.
\]

If only a finite number of the \( w_i \) are not equal to the corresponding \( y_i \), then \( \lim \overline{w}_n = \lim \overline{y}_n \), provided the second limit exists. It follows that

\[ P(\lim \overline{y}_n = 0) = 1, \]

which completes the proof.

\section{Limits of Proportions and Averages}

We have now come full circle. In Chapter 1 we interpreted the probability of an event as the empirical limit of the proportion as the number of repetitions becomes very large. And in Chapter 5 we interpreted a mean as the empirical limit of the average value as the number of repetitions becomes large.

We have now been able to build a probability model on a countable product space to describe a continuing sequence of repetitions. The model describes the sample proportion and shows that with probability 1 it converges to the probability. And more generally the model describes the sample average and shows that with probability 1 it converges to the mean. Thus in a consistent way the model describes and describes correctly the process on which, in fact, the model was based. This is a measure of the coherence of the model.

\section{Problems}

4 (continuation of Exercises 6.2.3 and 6.2.5) Show that \( m_n \xrightarrow{a.s.} \mu \); that \( g(m_n) \xrightarrow{a.s.} g(\mu) \) with \( g \) continuous.

5 (continuation of Problems 6.2.7 and 6.2.8) Show that \( \sigma_2 \xrightarrow{a.s.} \sigma^2 \); that \( s_1 \xrightarrow{a.s.} \sigma_{12} \).

6 (continuation of Problem 6.2.9) Show that \( \overline{m}_n \xrightarrow{a.s.} \overline{\mu} \).

\section{Testing Randomness}

We have been investigating statistical inference under the assumption that the system being examined is random. The alternative, of course, is that the system is not random—that conditions
are changing, thus causing response fluctuations, or that various determining factors are not under control, thus causing corresponding response fluctuations.

In this section we examine several simple tests of randomness, of whether we have a sample from a distribution. With scientific and sample survey investigations the randomness is often formally introduced by random sampling methods and by randomization of treatments to experimental units. For some of these investigations a simple test for randomness can be of interest as a routine check that conditions are, in fact, being kept essentially constant. We present one test of this kind in Sections C and D; it illustrates some probability theory and some combinatorial methods and is easy to apply.

With industrial applications the question of randomness is of much more specific interest—is some factor not under control, thus causing corresponding fluctuations in the response? For such cases there is a need for procedures that help choose among various factors for an assignable cause for any unusual variation found in the response values. In Sections A and B we examine the quality-control chart. The use of such charts has been an important aid in controlling quality in industrial production.

Our approach to inference in this chapter uses simple results directly available from probability theory in preceding chapters. At the same time the approach is consistent with our view that a first approach to inference should be quite general and not immediately involve a highly structured framework concerning the form of the distribution or concerning the production of the response. In succeeding chapters we shall examine inference with more structured models.

A QUALITY-CONTROL CHARTS

Consider a manufacturing or industrial process and suppose that we have a variable or characteristic \( v \) that is of special importance. For example, consider the production of bearings or moldings: we might take a sample of 5 each hour, record the values \( y_1, \ldots, y_6 \) of a critical dimension, and obtain the characteristic \( v = (y_1 + \cdots + y_6)/5 \), representing average performance at that time; alternatively, we might calculate the sample variance and obtain the characteristic \( v = s^2 \) representing the variability at that time. Or, for example, we might observe an interval of time and obtain a Poisson variable \( v \) recording the number of occurrences of some simple event such as an error or an accident. Or, for example, we might take a sample of 10 each half-day and record the binomial variable \( v \), the number of items that fail a go–no go test.

The quality-control method involves sampling routinely and systematically and recording the resulting values \( v_1, v_2, \ldots, v_N \) on a chart plotted against time. Limits are drawn that embrace a reasonable range if production is random: a central value \( CL \) is obtained as the average

\[
(1) \quad CL = \frac{1}{N} \sum_{i=1}^{N} v_i / N;
\]

an upper control limit UCL and lower control limit LCL are then formed as

\[
(2) \quad CL \pm 3\hat{\sigma}_v,
\]

where \( \hat{\sigma}_v \) is an estimated standard deviation obtained in ways to be described in Section B. Values \( v_{N+1}, v_{N+2}, \ldots \) are then plotted continually against time and the eye can check easily for values outside the control limits or for any unusual pattern in the sequence of values.

Consider an example involving the mass production of a plastic molding. A sample of 5 was taken each hour, and each unit in the sample was measured with respect to a critical dimension. The data are recorded in Table S.2 and the sample averages \( \bar{v}_i \) are plotted against time in Figure S.9.

If the production is random or in statistical control, we can estimate a central value and then estimate upper and lower control limits based on a conventional 3 standard deviations. For the central value \( CL \) we use the average of the 27 sample averages as an overall estimate of the mean \( \mu \) for the process:
TABLE S.2  
Plastic molding data

<table>
<thead>
<tr>
<th>( i )</th>
<th>( y_{i1} )</th>
<th>( y_{i2} )</th>
<th>( y_{i3} )</th>
<th>( y_{i4} )</th>
<th>( y_{i5} )</th>
<th>( \bar{y}_i = \bar{y}_i )</th>
<th>( R_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>140</td>
<td>143</td>
<td>137</td>
<td>134</td>
<td>135</td>
<td>137.8</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>138</td>
<td>143</td>
<td>143</td>
<td>145</td>
<td>146</td>
<td>143.0</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>139</td>
<td>133</td>
<td>147</td>
<td>148</td>
<td>139</td>
<td>141.2</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>143</td>
<td>141</td>
<td>137</td>
<td>138</td>
<td>140</td>
<td>139.8</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>142</td>
<td>142</td>
<td>145</td>
<td>135</td>
<td>136</td>
<td>140.0</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>136</td>
<td>144</td>
<td>143</td>
<td>136</td>
<td>137</td>
<td>139.2</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>142</td>
<td>147</td>
<td>137</td>
<td>142</td>
<td>138</td>
<td>141.2</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>143</td>
<td>137</td>
<td>145</td>
<td>137</td>
<td>138</td>
<td>140.0</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>141</td>
<td>142</td>
<td>147</td>
<td>140</td>
<td>140</td>
<td>142.0</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>142</td>
<td>137</td>
<td>145</td>
<td>140</td>
<td>132</td>
<td>139.2</td>
<td>13</td>
</tr>
<tr>
<td>11</td>
<td>137</td>
<td>147</td>
<td>142</td>
<td>137</td>
<td>135</td>
<td>139.6</td>
<td>12</td>
</tr>
<tr>
<td>12</td>
<td>137</td>
<td>146</td>
<td>142</td>
<td>142</td>
<td>140</td>
<td>141.4</td>
<td>9</td>
</tr>
<tr>
<td>13</td>
<td>142</td>
<td>142</td>
<td>139</td>
<td>141</td>
<td>142</td>
<td>141.2</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>137</td>
<td>145</td>
<td>144</td>
<td>137</td>
<td>140</td>
<td>140.6</td>
<td>8</td>
</tr>
<tr>
<td>15</td>
<td>144</td>
<td>142</td>
<td>143</td>
<td>135</td>
<td>144</td>
<td>141.5</td>
<td>9</td>
</tr>
<tr>
<td>16</td>
<td>140</td>
<td>132</td>
<td>144</td>
<td>145</td>
<td>141</td>
<td>140.4</td>
<td>13</td>
</tr>
<tr>
<td>17</td>
<td>137</td>
<td>137</td>
<td>142</td>
<td>143</td>
<td>141</td>
<td>140.0</td>
<td>6</td>
</tr>
<tr>
<td>18</td>
<td>137</td>
<td>142</td>
<td>142</td>
<td>145</td>
<td>143</td>
<td>141.8</td>
<td>8</td>
</tr>
<tr>
<td>19</td>
<td>142</td>
<td>142</td>
<td>143</td>
<td>140</td>
<td>135</td>
<td>140.4</td>
<td>8</td>
</tr>
<tr>
<td>20</td>
<td>136</td>
<td>142</td>
<td>140</td>
<td>139</td>
<td>137</td>
<td>138.8</td>
<td>6</td>
</tr>
<tr>
<td>21</td>
<td>142</td>
<td>144</td>
<td>140</td>
<td>138</td>
<td>143</td>
<td>141.4</td>
<td>6</td>
</tr>
<tr>
<td>22</td>
<td>139</td>
<td>146</td>
<td>143</td>
<td>140</td>
<td>139</td>
<td>141.4</td>
<td>7</td>
</tr>
<tr>
<td>23</td>
<td>140</td>
<td>145</td>
<td>142</td>
<td>139</td>
<td>137</td>
<td>140.6</td>
<td>8</td>
</tr>
<tr>
<td>24</td>
<td>134</td>
<td>147</td>
<td>143</td>
<td>141</td>
<td>142</td>
<td>141.4</td>
<td>13</td>
</tr>
<tr>
<td>25</td>
<td>138</td>
<td>145</td>
<td>141</td>
<td>137</td>
<td>141</td>
<td>140.4</td>
<td>8</td>
</tr>
<tr>
<td>26</td>
<td>140</td>
<td>145</td>
<td>143</td>
<td>144</td>
<td>138</td>
<td>142.0</td>
<td>7</td>
</tr>
<tr>
<td>27</td>
<td>145</td>
<td>145</td>
<td>137</td>
<td>138</td>
<td>140</td>
<td>141.0</td>
<td>8</td>
</tr>
</tbody>
</table>

(3) \( CL = \hat{\mu} = \bar{y} = 140.64 \).

For the standard deviation we can proceed in a variety of ways that are discussed in Section B; for the present data we use the estimate \( \hat{\sigma}_e = 1.657 \). For the upper and lower control limits (UCL and LCL), we use 3 standard deviations on either side of the central value CL:

\[
UCL = \hat{\mu} + 3\hat{\sigma}_y = 140.64 + 3 \times 1.657 = 145.62.
\]

(4) \( LCL = \hat{\mu} - 3\hat{\sigma}_y = 140.64 - 3 \times 1.657 = 135.67 \).

These values are plotted as horizontal lines in Figure S.9.

If the estimated control limits are reasonably close to the true limits, then values outside the limits will occur less than 1 percent of the time (normal case) and certainly less than 1/9 of the time (the extreme case under Chebyshev).

The usual practice is to continue plotting sample averages \( y_i = \bar{y}_i \) based on the systematic sampling procedure and to watch for anomalous patterns or occurrences. If a value goes outside the limits, the engineer would be called to search for an assignable cause for the departure. It is frequently found that the variation in a manufactured item is gradually reduced as assignable causes of variation are discovered and corrected.

The chart can also provide early warnings for progressive departures, often long before the effects would be discovered in the routine use of the manufactured units. One type of departure that is easily observed is a succession of high values or a succession of low values; as a rule of
FIGURE S.9
Sample averages plotted against time: center line, CL; upper control limit, UCL; lower control limit, LCL.

thumb, a succession of 6 or 7 values above the median or a succession of 6 or 7 values below the median can reasonably be viewed as an anomalous result.

B ESTIMATING THE STANDARD DEVIATION

Consider an estimate \( \hat{\sigma}_v \) for the standard deviation for use in the quality-control chart. First, consider the case where \( v_i = \bar{y}_i \) is based on samples of \( n \). We can estimate the variance \( \sigma^2 \) for an individual response by calculating the individual sample variances,

\[
(5) \quad s_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2
\]

and then forming an overall variance estimate

\[
(6) \quad s^2 = \frac{1}{N} \sum_{i=1}^{N} s_i^2 = \frac{1}{N(n-1)} \sum_{ij} (y_{ij} - \bar{y}_i)^2.
\]

For the standard deviation \( \sigma_v = \sigma_v \) we would then have

\[
(7) \quad \hat{\sigma}_v = \frac{s}{\sqrt{n}}.
\]

An alternative and numerically easier method is to calculate the range for the individual samples

\[
R_i = \max_{j=1}^{n} y_{ij} - \min_{j=1}^{n} y_{ij},
\]

and then form the average range
Sec. 7.5: Testing Randomness

\[ R = \frac{1}{N} \sum R_i. \]

Consider the range \( R \) for a sample of \( n \) from the standard normal; let \( f_n \) be its density (use Problem 3.6.13); then

\[ E(R) = \int_0^\infty R f_n(R) \, dR = I_n \]

gives the factor by which the range misestimates the standard deviation for the case of the normal. Accordingly, \( R/I_n \) is an estimate of \( \sigma \) and

\[ k_n = \frac{1}{\sqrt{n} I_n} \]

is an estimate of \( \sigma^{-2} \). We record some values for \( k_n \):

<table>
<thead>
<tr>
<th>( n )</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k_n )</td>
<td>0.627</td>
<td>0.341</td>
<td>0.243</td>
<td>0.192</td>
<td>0.161</td>
<td>0.140</td>
<td>0.124</td>
<td>0.112</td>
<td>0.103</td>
</tr>
</tbody>
</table>

The standard deviation \( \sigma^{-2} \) in Section A was estimated as

\[ \sigma^{-2} = \frac{0.192}{27} = 1.657. \]

Now consider the case where \( \nu \) is a Poisson(\( \theta \)) variable. The central value \( \text{CL} = \bar{\nu} = \frac{\sum \nu_i}{N} \) gives an estimate of \( \theta \). Accordingly, \( \sqrt{\bar{\nu}} \) is an estimate of the standard deviation \( \sqrt{\theta} \) and the control limits would be

\[ \text{UCL} = \bar{\nu} + 3\sqrt{\bar{\nu}}, \]

\[ \text{LCL} = \bar{\nu} - 3\sqrt{\bar{\nu}}. \]

Now consider the case where \( \nu \) is a binomial(\( n, p \)) variable. The central value \( \text{CL} = \bar{\nu} = \frac{\sum \nu_i}{N} \) gives an estimate of the mean number \( np \) of events. Accordingly, the standard deviation \( \sigma = (npq)^{1/2} \) can then be estimated as \( (\bar{\nu}(1 - \bar{\nu}/n))^{1/2} \), and the control limits would be

\[ \text{UCL} = \bar{\nu} + 3(\bar{\nu}(1 - \bar{\nu}/n))^{1/2}, \]

\[ \text{LCL} = \bar{\nu} - 3(\bar{\nu}(1 - \bar{\nu}/n))^{1/2}. \]

C THE RUN TEST

Now consider a sequence of response observations on some system and suppose that we want to test whether the system is random or in statistical control. A simple test can be made by examining the sequence in terms of values above (A) or below (B) the median value of the sequence. If there were a trend or a change in level somewhere along the sequence then there would be a tendency for A's to be clumped together and for B's to be clumped together—with the result that the number of runs of A's and of B's would tend to be smaller than under randomness. The run test involves counting the number of runs and comparing the total with the distribution for such under the hypothesis of randomness.

Consider an example. A manufacturing process mass produces an electronic component. A sample is taken every 20 minutes and a critical property is measured. The resulting sequence of 15 measurements is

\[ 0.212 \quad 0.211 \quad 0.213 \quad 0.212 \quad 0.213 \]

\[ B \quad B \quad B \quad B \quad B \]
The median value is 0.214 and the values are denoted B for below the median and A for above the median. Let $U$ be the number of runs of A's and B's in the sequence of A and B. There is one run of B's followed by one run of A's; thus $U = 2$.

The distribution of $U$ under randomness will be derived in Section D below; some percentage points are recorded in Table S.3. The table is calculated so that $U_{99\%}$ and $U_{95\%}$ are the largest values satisfying

$$P(U \leq U_{99\%}) \leq 1\%, \quad P(U \leq U_{95\%}) \leq 5\%;$$

and $U_{5\%}$ and $U_{1\%}$ are the smallest values satisfying

$$P(U \geq U_{5\%}) \leq 5\%, \quad P(U \geq 1\%) \leq 1\%.$$

For the example there are $n_1 = 7$ values above the median and $n_2 = 7$ values below the median. The value $U = 2$ is the smallest value possible; it is below the 95 percentage point 4 and below the 99 percentage point 3. It provides moderately strong evidence against randomness. An extremely large value of $U$ could also be of concern—perhaps suggesting an oscillation or periodic effect underlying the observations.

### Table S.3

**Distributions of $U$**

<table>
<thead>
<tr>
<th>$n_1 = n_2$</th>
<th>$U_{99%}$</th>
<th>$U_{95%}$</th>
<th>$U_{5%}$</th>
<th>$U_{1%}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2</td>
<td>3</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>3</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>4</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>5</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>6</td>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>6</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>11</td>
<td>6</td>
<td>7</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>8</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>13</td>
<td>7</td>
<td>9</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>14</td>
<td>8</td>
<td>10</td>
<td>19</td>
<td>21</td>
</tr>
<tr>
<td>15</td>
<td>9</td>
<td>11</td>
<td>20</td>
<td>22</td>
</tr>
<tr>
<td>16</td>
<td>10</td>
<td>11</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td>17</td>
<td>10</td>
<td>12</td>
<td>23</td>
<td>25</td>
</tr>
<tr>
<td>18</td>
<td>11</td>
<td>13</td>
<td>24</td>
<td>26</td>
</tr>
<tr>
<td>19</td>
<td>12</td>
<td>14</td>
<td>25</td>
<td>27</td>
</tr>
<tr>
<td>20</td>
<td>13</td>
<td>15</td>
<td>26</td>
<td>28</td>
</tr>
<tr>
<td>21</td>
<td>14</td>
<td>16</td>
<td>27</td>
<td>29</td>
</tr>
<tr>
<td>22</td>
<td>14</td>
<td>17</td>
<td>28</td>
<td>31</td>
</tr>
<tr>
<td>23</td>
<td>15</td>
<td>17</td>
<td>30</td>
<td>32</td>
</tr>
<tr>
<td>24</td>
<td>16</td>
<td>18</td>
<td>31</td>
<td>33</td>
</tr>
<tr>
<td>25</td>
<td>17</td>
<td>19</td>
<td>32</td>
<td>34</td>
</tr>
<tr>
<td>26</td>
<td>18</td>
<td>20</td>
<td>33</td>
<td>35</td>
</tr>
<tr>
<td>27</td>
<td>19</td>
<td>21</td>
<td>34</td>
<td>36</td>
</tr>
<tr>
<td>28</td>
<td>19</td>
<td>22</td>
<td>35</td>
<td>38</td>
</tr>
<tr>
<td>29</td>
<td>20</td>
<td>23</td>
<td>36</td>
<td>39</td>
</tr>
<tr>
<td>30</td>
<td>21</td>
<td>24</td>
<td>37</td>
<td>40</td>
</tr>
</tbody>
</table>
Sec. 7.5: Testing Randomness

Many tests can be designed and developed with the object of obtaining sensitivity to various patterns such as trends up or down or cyclic effects. The run test, however, is a convenient and simple test, and it is useful for illustrating some probability methods and calculations.

D THE DISTRIBUTION OF RUNS UNDER RANDOMNESS

Consider the distribution of the number of runs $U$ under randomness—that is, the sequence is a sample from some distribution. For this we can easily examine a generalization in which we have $n_1$ values above a certain percentile and $n_2$ values below the percentile. The symmetry under randomness gives equal probability

$$\frac{1}{\binom{n_1 + n_2}{n_1}}$$

to each possible sequence of $n_1$ A's and $n_2$ B's.

Let $r_1$ be the number of runs of A's and $r_2$ be the number of runs of B's. For example, we might have

$$\text{AAA_BBB_ABBAA_AB}, \quad r_1 = 3, \quad r_2 = 3,$$

with $U = r_1 + r_2 = 6$. Note that $|r_1 - r_2| \leq 1$; also note that if $|r_1 - r_2| = 1$, there is only one way in which an A sequence with $r_1$ runs can be interlaced with a B sequence with $r_2$ runs; and that if $|r_1 - r_2| = 0$, there are two ways in which an A sequence with $r_1$ runs can be interlaced with a B sequence with $r_2$ runs. For this let

$$\delta(r_1, r_2) = 2, 1, 0 \quad \text{according as } |r_2 - r_1| = 0, 1, \text{ other}.$$

Now consider the number of different ways in which a sequence of $n_1$ A's can be separated into $r_1$ runs of A's: the number of ways of choosing $r_1 - 1$ separation points from the $n_1 - 1$ spaces between A's is $\binom{n_1 - 1}{r_1 - 1}$. Similarly, the number of different ways in which $r_2$ B runs can be formed is $\binom{n_2 - 1}{r_2 - 1}$. It follows, then, that the probability function for $(r_1, r_2)$ is

$$p(r_1, r_2) = \frac{\binom{n_1 - 1}{r_1 - 1} \binom{n_2 - 1}{r_2 - 1}}{\binom{n_1 + n_2}{n_1}} \delta(r_1, r_2).$$

The probabilities for $U$ can then be calculated easily:

$$P(U = 2r + 1) = \frac{\binom{n_1 - 1}{r} \binom{n_2 - 1}{r - 1} + \binom{n_1 - 1}{r - 1} \binom{n_2 - 1}{r}}{\binom{n_1 + n_2}{n_1}},$$

$$P(U = 2r) = \frac{2 \binom{n_1 - 1}{r - 1} \binom{n_2 - 1}{r - 1}}{\binom{n_1 + n_2}{n_1}}.$$

The mean of $U$ is fairly easy to calculate; the variance is substantially more complicated. The mathematically direct calculation is concerned with the general factorial moment for $(r_1, r_2)$.

For this we rewrite the hypergeometric identity [Problem 3.4.7 or implicit in formula (3.5.3)]
as
\[ \sum_i \binom{m_1}{d + i} \binom{m_2}{i} = \sum_i \binom{m_1 - d - i}{m_2} \binom{m_2}{i} = \binom{m_1 + m_2}{m_1 - d} = \binom{m_1 + m_2}{d + m_2} \]
and use it twice in the following calculation:
\[ \sum_{r_1, r_2} \binom{n_1 - 1}{r_1 + d - 1} \binom{n_2 - 1}{r_2 - 1} \delta(r_1, r_2) \]
\[ = \sum_r \binom{m_1 - 1}{r + d} \binom{m_2 - 1}{r - 1} + 2 \sum_r \binom{n_1 - 1}{r + d - 1} \binom{n_2 - 1}{r - 1} + \sum_r \binom{n_1 - 1}{r + d - 2} \binom{n_2 - 1}{r - 1} \]
\[ = \binom{n_1 + n_2 - 2}{d + n_2} + 2 \binom{n_1 + n_2 - 2}{d + n_2 - 1} + \binom{n_1 + n_2 - 2}{d + n_2 - 2} = \binom{n_1 + n_2}{d + n_2}. \]
Now consider the joint factorial-type moment:
\[ E(r_1 - 1)^{h_1}(r_2 - 1)^{h_2} = \sum_{r_1, r_2} (r_1 - 1)^{h_1}(r_2 - 1)^{h_2} \delta(r_1, r_2) \]
\[ = (n_1 - 1)^{h_1}(n_2 - 1)^{h_2} \sum_{r_1, r_2} \frac{\binom{n_1 - h_1 - 1}{r_1 - h_1 - 1} \binom{n_2 - h_2 - 1}{r_2 - h_2 - 1}}{\binom{n_1 + n_2}{n_1}} \delta(r_1, r_2) \]
\[ = (n_1 - 1)^{h_1}(n_2 - 1)^{h_2} \frac{\binom{n_1 + n_2 - h_1}{h_2 - h_1 + n_2 - h_2}}{\binom{n_1 + n_2}{n_1}} \]
\[ = (n_1 - 1)^{h_1}(n_2 - 1)^{h_2} \frac{\binom{n_1 + n_2 - h_1}{n_2 - h_1}}{\binom{n_1 + n_2}{n_1}} \]
\[ = \binom{n_1 + n_2}{n_1} \binom{n_1 + n_2 - h_1}{n_2 - h_1}. \]
The means, variances, and covariances of \( r_1 \) and \( r_2 \) can then be calculated:
\[ E(U) = E(r_1) + E(r_2) = \frac{2n_1 n_2}{n_1 + n_2} + 1, \]
\[ \text{var}(U) = \text{var}(r_1) + \text{var}(r_2) + 2 \text{cov}(r_1, r_2) = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}. \]
The calculations for the distribution of \( U \) with \( n_1 \) and \( n_2 \) large can be excessive. Fortunately, the distribution approaches normal form rapidly and the approximation is generally adequate for \( n_1 \) and \( n_2 \geq 10 \): treat
\[ Z = \frac{U - E(U)}{SD(U)} \]
as standard normal (see Problem 4).

E EXERCISES

1. Twenty-four samples of four dispensers were taken at regular intervals from a bottling process. The weight of the contents was determined for each dispenser and the average and range for each sample are as follows:
Calculate the upper and lower control limits for the sample averages. Plot the quality-control chart and comment on the results obtained. The specifications are that individual charges should have a mean of 454 and a standard deviation of 9 or less; comment on the production process.

(continuation) Control charts are sometimes used for the variation as given by say the sample range. The values of $R$ are plotted as were the values of $\overline{y}$; the overall average $\overline{R} = \Sigma R_i / N$ is calculated and lower and upper control limits are obtained as $LCL = D_1 \overline{R}$ and $UCL = D_2 \overline{R}$.

The values of $D_1$ and $D_2$ are obtained as 3 standard deviation limits based on the distribution of the range for normal sampling:

<table>
<thead>
<tr>
<th>$n$</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_1$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.08</td>
<td>0.14</td>
<td>0.18</td>
<td>0.22</td>
</tr>
<tr>
<td>$D_2$</td>
<td>3.27</td>
<td>2.58</td>
<td>2.28</td>
<td>2.12</td>
<td>2.00</td>
<td>1.92</td>
<td>1.86</td>
<td>1.82</td>
<td>1.78</td>
</tr>
</tbody>
</table>

Calculate the upper and lower control limits for the sample ranges. Plot the quality-control chart and comment on the results.

3 Millikan in 1930 obtained the following sequence of measurements on the charge on an electron (column by column):

4.782  4.778  4.808  4.740  4.790
4.769  4.791  4.771  4.775  4.747

Make a run test for randomness. Use the tables as far as possible. Calculate the observed level of significance using the normal approximation.

F PROBLEMS

4 Consider the run test as $N = n_1 + n_2 \to \infty$ with $n_1 = \alpha_1 N$ and $n_2 = \alpha_2 N$. Let $p_N(U)$ be the probability function and consider
\[ f_n(z) = p_N(E(U) + z\sigma_U)\sigma_U \]

as \( N \to \infty \). Use Stirling’s formula as described in Problem 7.4.24. Thus verify the limiting normality.

5 Consider the two-sample problem as described in Problem 7.4.14. The two samples are combined and then ordered from smallest to largest. The values in this sequence are then replaced by A’s if they originated from the first sample and by B’s if they originated from the second sample. Let \( U \) be the number of runs of A’s and B’s.

(a) Show that the distribution of \( U \) in this section is appropriate with \( n_1 = m \) and \( n_2 = n \).

(b) For the data in Exercise 7.2.5 and Problem 7.2.7, test the hypothesis that the location shift \( \theta = 0 \).

9.6

BAYES ESTIMATION

A biophysicist is investigating radiation effects on mice. A particular black mouse can be homozygous black \( BB \) or heterozygous Black-brown \( Bb \) with probabilities \( 1/3 \) and \( 2/3 \), respectively. The distribution of a certain characteristic being investigated is different for the two-gene combinations. How can the investigator use the prior probabilities \( 1/3 \) and \( 2/3 \) for the gene combinations \( BB \) and \( Bb \)?

In this section we examine some methods of inference for the case where the parameter value in whole or in part has arisen from a performance of some antecedent random system with known probabilities.

A THE ANALYSIS OF THE BAYESIAN MODEL

For notation let \( y \) be the response under investigation with sample space \( S \), and let \( \theta \) be the parameter with parameter space \( \Omega \). We assume that the statistical model

\[ \{f(\cdot|\theta): \theta \in \Omega\} \]

is given by density functions as discussed in Section 8.1.

We now suppose that the true value of the parameter \( \theta \) is a realization from some separate random system with known probability properties. Often in such cases the true value of \( \theta \) is a characteristic of some individual or experimental unit and the individual or experimental unit has been obtained from a finite population by random sampling (Section 7.3) or from some random system such as occurs with the combination of genes in mating. Typically, then, the response \( y \) is a measurement or experimental result on the sampled individual or unit. Let \( p(\cdot) \) be the density function for the characteristic \( \theta \) relative to some volume or counting measure on the parameter space \( \Omega \).

We thus have a prior probability model,

\[ p(\cdot), \]

describing the source of the true parameter value, and we have a statistical model,

\[ \{f(\cdot|\theta): \theta \in \Omega\}, \]

describing the production of the response \( y \) for any given value for \( \theta \).

The Bayesian approach is to enlarge the viewpoint and examine the combined model describing the generation of the \( \theta \) value and the production of the response value \( y \) based on that \( \theta \) value. The enlarged sample space is \( \Omega \times S \); the class of events is the appropriate product class as discussed in Section 3.3E; and the distribution is given by the joint density \( p(\theta)f(y|\theta) \) relative
to the volume or counting measure on \( \Omega \times \mathcal{G} \). We thus obtain a probability model describing the combined outcome \((\theta, y)\).

The analysis of the combined model uses basic probability theory. Consider an observed response \( y \). We then have a realization \((\theta, y)\) from the combined probability model and the value of the second coordinate is known. The principles of conditional probability in Chapter 4 then give the conditional distribution of \( \theta \) given \( y \) as the appropriate description of the unknown \( \theta \).

The joint probability model has probability differential

\[
1. \quad p(\theta)f(y|\theta) \, d\theta \, dy.
\]

The conditional distribution for \( \theta \) given \( y \) is

\[
2. \quad q(\theta | y) \, d\theta = k^{-1}(y)p(\theta)f(y|\theta) \, d\theta,
\]

where the normalizing constant \( k(y) \) is the marginal density for \( y \) and is obtained by integration,

\[
3. \quad k(y) = \int p(\theta)f(y|\theta) \, d\theta.
\]

This conditional distribution (2) is the appropriate description for \( \theta \) based on the observed \( y \) and the combined model (1).

The distribution \( p(\theta) \, d\theta \) is called the prior distribution; it describes the true \( \theta \) before any measurement or experimental investigation of it. The conditional distribution \( q(\theta | y) \, d\theta \) is called the posterior distribution; it describes the true \( \theta \) after the measurement or experimental investigation has produced the value \( y \). Note the simple form of the posterior density: prior density times likelihood,

\[
q(\theta | y) \propto p(\theta)l(y|\theta),
\]

with a normalizing constant obtained by integration; see Figure S.10.

The inference procedure is essentially solved with the production of the posterior distribution itself. Of course, certain characteristics of the posterior distribution may be of particular interest. As examples, consider the mean, the quartiles (the 25 percent point and the 75 percent point), or say a central 95 percent probability interval \((a_1, a_2)\), where

\[
\int_{a_1}^{a_2} q(\theta | y) \, d\theta = 95\%.
\]

The particular choice of characteristic, however, would be a matter for the judgment of a particular user at a particular time.

**B EXAMPLES**

Consider examples: involving an individual sampled randomly from a population; and involving a characteristic obtained from random gene combination in mating.

**EXAMPLE 1**

Let \( \theta \) be an intelligence characteristic and suppose a measuring procedure is available which gives a response \( y \) that is normally distributed about the true value \( \theta \) with standard deviation 5. Also suppose that an individual is randomly sampled from a population in which the intelligence characteristic is known to be normally distributed with mean 100 and standard deviation 10, and that the measuring procedure applied to the individual gives \( y = 120 \).

The prior density, the likelihood function, and the posterior density are as follows:

\[
p(\theta) = \frac{1}{\sqrt{2\pi \cdot 10}} \exp \left\{ -\frac{1}{2 \cdot 10^2} (\theta - 100)^2 \right\},
\]
\[ L(y|\theta) = c \exp \left\{ - \frac{1}{2 \cdot 5^2} (y - \theta)^2 \right\}. \]

\[ q(\theta|y) = k^{-1}(y)p(\theta)f(y|\theta) \]
\[ = c(y) \exp \left\{ - \frac{1}{2 \left( \frac{1}{10^2} + \frac{1}{5^2} \right)} \theta^2 + \left( \frac{100}{10^2} + \frac{y}{5^2} \right) \theta + \cdots \right\} \]
\[ = \frac{1}{\sqrt{2\pi\tau}} \exp \left\{ - \frac{1}{2\tau^2} (\theta - \nu(y))^2 \right\}. \]

Thus the posterior distribution for \( \theta \) is normal with mean
\[ \nu(y) = \left( \frac{100}{10^2} + \frac{y}{5^2} \right) \left( \frac{1}{10^2} + \frac{1}{5^2} \right)^{-1} \]
and precision (reciprocal variance)
\[ \frac{1}{\tau^2} = \frac{1}{10^2} + \frac{1}{5^2} = \frac{1}{20}. \]

For the observed \( y = 120 \) the posterior distribution is normal \((116, 4.47)\). See Figure S.10.

Note that the posterior mean is the weighted average of 100 and \( y \) with weights propor-

**FIGURE S.10**
Prior is normal\((100, 10)\); likelihood is normal\((120, 5)\); posterior is normal\((116, 4.47)\).
tional to precision; see the procedure in Section 9.3C. Also note that the posterior variance is the variance of such an average; see Section 9.3C.

The measuring procedure alone would indicate the value 120 with a scaling unit of 5. The effect of enlarging the model and using the prior distribution is to displace the indicated value toward the population mean 100 and to reduce the scaling unit slightly.

EXAMPLE 2

A classical example of Bayesian analysis in genetics is concerned with a black mouse that is an offspring of a $Bb$ mouse mated with a $Bb$ mouse. Capital $B$ denotes a dominant gene for the color black and small $b$ denotes a recessive gene for the color brown. An offspring receives one gene from each parent, and it could be either of the parent’s genes with equal probability. Thus the prior distribution for an offspring from $Bb$ to $Bb$ is

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>$BB$</th>
<th>$Bb$</th>
<th>$bb$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>1/4</td>
<td>1/2</td>
<td>1/4</td>
</tr>
</tbody>
</table>

An offspring is black if it receives a dominant gene $B$, and it is brown otherwise. Thus the information that the mouse is black identifies the mouse as being $BB$ or $Bb$. It follows that the posterior distribution given the mouse color is

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>$BB$</th>
<th>$Bb$</th>
<th>$bb$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>1/3</td>
<td>2/3</td>
<td>0</td>
</tr>
</tbody>
</table>

Now suppose that the particular mouse is tested by mating it with a brown mouse $bb$. And suppose that the seven offspring are all black. The probability of a black offspring from $\theta = BB$ mated with $bb$ is 1 (offspring are necessarily $Bb$); and the probability that 7 are black is 1. The probability of a black offspring from $\theta = Bb$ mated with $bb$ is 1/2; and the probability that 7 are black is $1/2^7 = 1/128$. Thus

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>$BB$</th>
<th>$Bb$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f(y</td>
<td>\theta)$</td>
<td>1</td>
</tr>
</tbody>
</table>

Multiplying the two functions together (modulating one by the other) gives

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>$BB$</th>
<th>$Bb$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Posterior $\propto$</td>
<td>1/3</td>
<td>1/192</td>
</tr>
<tr>
<td>Posterior</td>
<td>64/65</td>
<td>1/65</td>
</tr>
</tbody>
</table>

The likelihood function gives a ratio 128:1 in favor of homozygous $BB$ to heterozygous $Bb$. Combining this with the probability information concerning the source of the mouse produces the modified ratio 64:1. Note that the result from the testing operation alone is the likelihood ratio 128:1; that the result from the prior process alone is the ratio 1:2; and that the result from the testing operation combined with the source information is the modified ratio 64:1.
C BAYES SOLUTION

Consider the combined model (1) as discussed in Section A. We now investigate some theoretical aspects of choosing an estimator \( t(y) \) for the parameter \( \theta \). We do this for theoretical interest on the basis of the combined model (1). Of course from the principles of probability theory we do not, in fact, have this initial option: the conditional distribution (2) is the proper description of \( \theta \) and choosing an estimate must be based directly on it.

Suppose that we have a loss function \( l(t, \theta) \), which presents the loss to us if we present the estimate \( t \) when the parameter has, in fact, the value \( \theta \). Accordingly, if we use an estimator \( t(y) \) our mean loss or risk is given by

\[
\mathcal{R} = E(l(t(y), \theta)) = \int l(t(y), \theta)p(\theta|y)\,d\theta\,dy.
\]

It is reasonable from some points of view to consider choosing the function \( t \) in such a way as to minimize the preceding risk \( \mathcal{R} \). The risk \( \mathcal{R} \) can be written

\[
\mathcal{R} = \int \left[ \int l(t(y), \theta)q(\theta|y)\,d\theta \right] k(y)\,dy;
\]

with the conditional risk given \( y \) recorded in the square brackets.

Now suppose that for each value of \( y \) we choose \( t(y) \) to minimize the conditional risk

\[
\int l(t(y), \theta)q(\theta|y)\,d\theta.
\]

If we choose the function \( t \) in this way, we obtain an integrand \( \cdots \) for (5) that is smaller \((\leq)\) than any other integrand; accordingly, such a function \( t \) minimizes the risk (5). In short: we minimize the risk generally if we minimize the conditional risk given \( y \). This agrees with the general result we mentioned at the beginning.

Now for a real parameter \( \theta \), suppose that we use squared error

\[
l(t, \theta) = (t - \theta)^2
\]

as the loss in estimating \( t \) when the parameter is \( \theta \). The minimum-risk estimate \( t(y) \) is obtained by minimizing

\[
\int (t(y) - \theta)^2q(\theta|y)\,d\theta;
\]

thus the minimum mean-square-error estimate is given by the conditional mean,

\[
t(y) = \int \theta q(\theta|y)\,d\theta,
\]

of the posterior distribution; see Problem 5.3.29.

An estimator obtained by minimizing the risk with the combined model discussed in this section is called a Bayes solution.

D SOME OTHER USES OF BAYESIAN ANALYSIS

Consider a random system having a response \( y \) with sample space \( S \) and a parameter \( \theta \) with parameter space \( \Omega \). We assume that the statistical model

\[
\{l(\cdot|\theta): \theta \in \Omega\}
\]

is given by density functions. For an application recall that a model presents the various possibilities for response performances as identified by earlier experience with the same or similar systems.
Now suppose that the true $\theta$ is some physical characteristic of the system and that there is no random system as an identifiable source for the true $\theta$. Some statisticians recommend the following approaches to the analysis of an observed value with this model:

1. There is no reason to prefer one $\theta$ value to another; hence assign a uniform density ($p(\theta) = \text{constant}$) to the parameter space and perform a Bayesian analysis as in Section A. The resultant is the likelihood function but normed and treated as a posterior density. If $\Omega$ is finite, then $p(\theta)$ is defined; if $\Omega$ is continuous, a uniform density relative to one metric in $\theta$ will typically not be a uniform density with respect to another metric; and if $\Omega$ is unbounded, a constant density must be truncated in order to be integrable. This method is called the method of insufficient reason.

2. The vague and general information concerning $\theta$ does not sharply differentiate among the possible values in a reasonable central range embraced by the likelihood function. Hence assign an approximately constant prior and perform a Bayesian analysis. The result is the likelihood function but treated as a posterior density. This is the method of the diffuse prior.

3. The investigator consults his feelings and impressions concerning the various $\theta$ values and assigns a density that represents the shadings up and down from one $\theta$ value to another. He then performs a Bayesian analysis as in Section A. This is the method of the personal prior.

Whatever the justifications, these extended Bayesian methods do produce solutions, as many solutions as there are priors. This is in rather sharp contrast to a somewhat limited range of results from the other theory in this chapter and in Chapter 10, and is certainly an attractive feature of the methods. A solution, however, is at most as good as the arguments that support it, and the prior distribution can be a very personal ingredient of the investigator. Certainly solutions can be assessed pragmatically—in terms of how they work in practice. But, of course, from this point of view it suffices to obtain solutions from any plausible source.

If the extended Bayesian methods are to be favored because they produce solutions, it should be noted that simpler methods produce as many solutions, in fact more solutions. We can start with the likelihood function; apply weight functions for any plausible reason; integrate it; section it; profile it; use any device associated with the Bayesian or other approaches. No probability arguments are needed, and there is perhaps less chance, then, that a result will be labeled and hence taken as a probability. These alternative methods produce as many solutions; in fact, they produce more because the likelihood function itself can be obtained and it provides a point evaluation rather than the density evaluation of the Bayesian approach.

An argument sometimes advanced for the extended Bayesian methods notes that subjective elements are involved in the selection of a model and in the choice of methods for analysis; and it thus recommends that subjective assessments also be made of one parameter value against another. This argument overlooks some important distinctions. The choice of a model properly is based on objective evidence concerning the system being investigated, and a model provides the set of possibilities for the system being investigated; the subjective prior, on the other hand, is placing shadings or preferences on the various possibilities. It is one thing to allow a set of possibilities for a system and it is a different thing to introduce preferences and shadings among the possibilities. The extended Bayesian approach provides adjusted preferences or shadings. It does not produce subjective assessments from the data and model alone.

The principles of the experimental method in science have long been concerned with excluding any effects from an investigator's preferences or beliefs. The extended Bayesian approach treats these subjective elements as the primary input, and in doing this it runs counter to the basic principles of experimentation.

Another argument for the extended Bayesian approach comes from decision theory—that optimum decisions should be based on a prior distribution and a Bayesian analysis using a utility function. In any observational or experimental investigation a first objective can always be to assemble the information concerning the set of possibilities presented by the model. The Bayesian
approach says that this information should be combined with a prior; the decision-theory argument proposes this as a necessity for optimum decisions. Of course, an investigation can have a multitude of different end users, and this argument may perhaps be appropriate for certain individual end users. The scientific approach, however, would say that the presentation of results from an investigation should be objective and not involve prior shadings and preferences on the set of possibilities.

The question remains concerning some end user and how he is to use the information from the observational or experimental investigation. The extended Bayesian use of a prior and a utility function provides one approach; the justification for this approach derives from the Bayesian acceptance of prior probabilities. Some recent theoretical results indicate, however, that probabilities are not available in general to describe unknowns; this contradicts the usual justification for extended Bayesian methods. Of course, an end user can examine separately the information from the investigation and his prior impressions concerning the unknown of the system, and can then make his judgments and decisions. This may be a safer route than the examination of a formal combination of the information from the investigation and the prior impressions concerning the unknown.

E EXERCISES

1 A certain capability  \( \theta \) is approximately normal(100, 20) in a relatively homogeneous population. A measuring procedure for  \( \theta \) is normal(\( \theta \), 10). An individual randomly sampled from the population has the measurement \( y = 110 \).
   (a) Determine the posterior distribution for  \( \theta \).
   (b) Record the estimate of  \( \theta \) using the mean of the posterior distribution.

2 A certain tissue characteristic  \( \theta \) in a species of fish is normal(8, 1.5). A rapid method of measuring  \( \theta \) gives a normal(\( \theta \), 0.5) distribution for the measurement  \( y \). A fish randomly sampled from the population has the measurement  \( y = 10.5 \); estimate the value of  \( \theta \) using the mean of the posterior distribution.

3 Suppose that the mouse in Example 2 was tested by mating it with a brown mouse  \( bb \) and the offspring were 7 black and 1 brown. Determine the posterior distribution for the genetic type of the mouse.

F PROBLEMS

4 A particular population consists of two racial groups  \( A_1 \) and  \( A_2 \) in the proportion 1 to 5. The distribution of a capability  \( \theta \) in racial group  \( A_1 \) is normal(85, 10) and in  \( A_2 \) is normal(115, 10). A measuring procedure for  \( \theta \) is normal(\( \theta \), 10). An individual randomly sampled from the population obtains  \( y = 110 \) without record of the racial group.
   (a) Determine the posterior distribution for  \( \theta \).
   (b) Estimate of  \( \theta \) using the mean of the posterior distribution.
   (c) Calculate the posterior probability for racial group  \( A_1 \).
   (d) If the individual is of racial group  \( A_1 \), calculate the mean of the posterior distribution of  \( \theta \).

Note that the posterior distribution here is a normal likelihood function located at 110 with scale 10 multiplied by a factor that decreases very rapidly from left to right; the factor is the right tail of the normal(85, 10). Thus the posterior is shifted sharply to the left in comparison with the likelihood. The validity, of course, depends heavily on how well the tail of the normal describes the population; ordinarily the normal is not viewed as a good approximation out on the tails.

5 Let  \( y = (y_1, \ldots, y_n) \) be a sample from a normal distribution  \( (\theta, \sigma) \) with  \( \theta \) in  \( R \). And suppose that the true  \( \theta \) is a realized value from a normal distribution  \( (\theta_0, \sigma_0) \). Determine the posterior distribution for  \( \theta \) given  \( y \).
(continuation) (a) Determine the estimator \( t(y) \) for \( \theta \) that minimizes the mean-squared error \( R = E((t - \theta)^2) \). (b) Determine the estimator \( t(y) \) for \( \theta \) that minimizes the mean absolute error \( R = E(|t - \theta|) \).

7. Let \( x = (x_1, \ldots, x_n)' \) be a sample from the Bernoulli(\( p \)) distribution with \( p \in [0, 1] \). And suppose that the true \( p \) is a realized value from a beta distribution(\( \alpha_0, \beta_0 \)). Determine the posterior distribution for \( p \).

8. Let \( y = (y_1, \ldots, y_n) \) be a sample from the Poisson(\( \theta \)) distribution with \( \theta \) in \( \mathbb{R}^+ \). And suppose that the true \( \theta \) is a realized value from the distribution for \( \beta_0 u \), where \( u \) is gamma(\( \rho_0 \)). Determine the posterior distribution of \( \theta \).

9. Let \( y \) have a multivariate normal(\( \mu; \Sigma_0 \)) distribution with \( \mu \) in \( \mathbb{R}^k \). And suppose that the true \( \theta \) is a realized value from the multivariate normal(\( \mu_0; \Sigma_0 \)). Determine the posterior distribution of \( \theta \).

10.5

CONFIDENCE INTERVALS AND CONFIDENCE REGIONS

Confidence intervals are one of the basic methods of statistical inference. Consider some examples from preceding chapters. For this let \((y_1, \ldots, y_n)\) be a sample from a distribution on the real line.

1. If the distribution has finite mean \( \mu \) and finite variance, then

\[
(\bar{y} \pm z_{n/2}\sigma_y/\sqrt{n})
\]

is an approximate (large-sample) 1 - \( \alpha \) confidence interval for \( \mu \).

2. If, in addition, the distribution form is normal, then

\[
(\bar{y} \pm t_{n/2}\sigma_y/\sqrt{n})
\]

is an exact 1 - \( \alpha \) confidence interval for \( \mu \) [based on the Student(\( n - 1 \)) distribution].

3. If, in addition, the variance is known to be \( \sigma^2_y \), then

\[
(\bar{y} \pm z_{n/2}\sigma/\sqrt{n})
\]

is an exact 1 - \( \alpha \) confidence interval for \( \mu \).

4. If the distribution has a density function \( f(\cdot|\theta) \) satisfying the regularity discussed in Section 8.6, then

\[
(\bar{y} \pm z_{n/2}/\sqrt{n(f(\bar{y}))})
\]

is an approximate (large-sample) confidence interval for \( \theta \).

In this section we discuss some theory and some methods for confidence intervals and confidence regions.

A GENERAL THEORY OF CONFIDENCE REGIONS

Consider a statistical model with a response \( y \) taking values in a sample space \( \mathcal{S} \) and a parameter \( \theta \) taking values in a parameter space \( \Omega \). A 1 - \( \alpha \) confidence interval or region is an interval or set on \( \Omega \) that depends on the response \( y \); for each \( y \) we obtain a set in \( \Omega \). Thus formally a confidence region is a function from the sample space \( \mathcal{S} \) into a class of subsets on \( \Omega \); the essential probability property is recorded in the following definition.
DEFINITION 1

A function \( B(\cdot) \) carrying a point \( y \) in \( S \) to a set \( B(y) \) on \( \Omega \) is a \( 1 - \alpha \) confidence region for the parameter \( \theta \) if

\[
P(\theta \in B(y) | \theta) \geq 1 - \alpha
\]
for each \( \theta \) in \( \Omega \).

Confidence regions are very closely connected with hypothesis-testing methods. To show this connection, let \( C(\theta_*) \) be a size \( \alpha \) critical region for testing the hypothesis \( \theta = \theta_* \); that is,

\[
P(C(\theta_*) | \theta_*) \leq \alpha.
\]

And let \( A(\theta_*) = \Omega - C(\theta_*) \) be the corresponding acceptance region:

\[
P(A(\theta_*) | \theta_*) \geq 1 - \alpha.
\]

We then establish a correspondence between confidence regions and acceptance regions. Given the acceptance regions \( A(\theta) \) for \( \theta \) in \( \Omega \), we examine an observed response \( y \) and collect together the parameter values that are acceptable for that response:

\[
B(y) = \{ \theta : y \in A(\theta) \}.
\]

Clearly we have the following logical equivalence:

\[
\theta \in B(y) \iff y \in A(\theta).
\]

For we can view \( A(\theta) \) and \( B(y) \) as the first- and second-axis sections through a set \( D \) on \( S \times \Omega \); see Figure S.11. From formula (4) we then have

\[
P(\theta \in B(y) | \theta) = P(y \in A(\theta) | \theta) \text{ for } \theta \in \Omega,
\]

which establishes the following lemma.

LEMMA 2

\( B(\cdot) \) is a \( 1 - \alpha \) confidence region for \( \theta \) if and only if, for each \( \theta_* \), \( A(\theta_*) \) is a \( 1 - \alpha \) acceptance region for testing \( \theta = \theta_* \).

---

**FIGURE S.11**

The \( \theta \) distribution can be viewed as a distribution along the \( \theta \)-section of \( S \times \Omega \). \( A(\theta) \) is a \( 1 - \alpha \) acceptance region for the value \( \theta \); \( B(y) \) is the \( 1 - \alpha \) confidence region obtained from a value \( y \).
We thus have a formal verification of the procedure we followed in Chapters 7 and 8: test each of the parameter values and collect together those that are acceptable when assessed against the observed response.

**B CONFIDENCE INTERVAL FOR THE BINOMIAL PROBABILITY \( p \)**

Consider a sample \( \{x_1, \ldots, x_n\} \) from the Bernoulli(\( p \)) distribution with \( p \) in \([0, 1]\). We apply the general approach discussed in the preceding section. For each value \( p \), a \( 1 - \alpha \) acceptance interval \( A(p) = [a_1(p), a_2(p)] \) can be formed so that

\[
\sum_{a_1(p)}^{a_2(p)} \binom{n}{y} p^y (1 - p)^{n-y} \geq 1 - \alpha,
\]

where the interval is an interval of values for the likelihood statistic \( y = \sum x_i \). Clearly we would want to come as close to \( 1 - \alpha \) as possible, and we do have some freedom in determining an interval that is approximately central to the distribution.

For the confidence level \( 1 - \alpha = 95 \) percent and for various sample sizes, the acceptance regions are plotted in Figure S.12 (using \( \hat{p} = y/n \) rather than \( y \) itself). A 95 percent confidence region is then obtained from the appropriate vertical section: find the observed \( \hat{p} \) on the horizontal scale and on the vertical section obtain the interval between the curves for the particular sample size.

**C THE USE OF PIVOTAL QUANTITIES**

A somewhat more direct approach to confidence regions is by means of pivotal quantities:

**DEFINITION 3**

A function

\[
t = q(y, \theta)
\]

defined on \( \mathbb{S} \times \Omega \) is a pivotal quantity if it has a fixed distribution (independent of \( \theta \)) as derived from the \( \theta \)-distribution for \( y \).

For an illustration consider a sample \( \{y_1, \ldots, y_n\} \) from a normal \( (\mu, \sigma) \) with \( (\mu, \sigma) \) in \( R \times R^+ \). Then the familiar \( t \) quantity

\[
t = \frac{\bar{y} - \mu}{s_y / \sqrt{n}}
\]

has a fixed distribution, the Student\((n - 1)\) distribution. Note that the pivotal quantity involves \( \mu \) and not \( \sigma^2 \). Also recall that we used \( t \) in Section 8.3C to determine a confidence interval for \( \mu \).

Now, in general, let \( T \) be a set having probability \( 1 - \alpha \) according to the fixed distribution of the pivotal quantity \( t = q(y, \theta) \) in formula (7):

\[
A(\theta) = \{ y : q(y, \theta) \in T \}
\]

is a \( 1 - \alpha \) acceptance region for testing the value \( \theta \). And

\[
B(\theta) = \{ \theta : q(y, \theta) \in T \}
\]

is a \( 1 - \alpha \) confidence region for the parameter \( \theta \). Note that this definition (9) for \( B(\theta) \) is the general equivalent of "rearranging the inequalities" as, say, in Section 7.2C.
We would naturally choose a pivotal quantity that is a function of the likelihood statistic; this derives from properties of the likelihood statistic in Chapter 8.

D THE POWER OF A CONFIDENCE REGION

Let $A(\theta_*) = \mathcal{S} - C(\theta_*)$ be a $1 - \alpha$ acceptance region for testing the parameter value $\theta_*$. And let $B(\gamma)$ be the corresponding $1 - \alpha$ confidence region for the parameter $\theta$.

The power function for testing the value $\theta_*$ is given by

$$P_{\theta_*}(\theta) = P(C(\theta_*) | \theta) = P(\mathcal{S} - A(\theta_*) | \theta) = P(\theta^* \notin B(\gamma) | \theta).$$

Thus the power function gives the probability that the confidence region does not cover the value $\theta^*$ when the true value is $\theta$. For all $\theta$ values except $\theta_*$, this is the probability of not covering a wrong parameter value. Thus in certain problems it is possible to obtain a uniformly most powerful confidence interval. See Problem 10.2.7 concerning the parameter $\theta$ of the uniform$(0, \theta)$ distribution.
And more generally it is possible to obtain uniformly most powerful unbiased confidence intervals; see examples in Section 10.3A concerning the parameter \( \mu \) with the normal\((\mu, \sigma)\) and the normal\((\mu, \sigma)\).

**E CONFIDENCE REGIONS FOR COMPONENT PARAMETERS**

Our general discussion in Section A is concerned with the full parameter \( \theta \) of a statistical model. Often, however, we are concerned with a coordinate of \( \theta \) or some function, say \( \beta(\theta) \) of \( \theta \). For example, with the normal\((\mu, \sigma)\) we may be concerned primarily with \( \mu \) and obtain a confidence interval as mentioned in Section C.

We now discuss briefly the theory of confidence regions for component parameters such as \( \beta(\theta) \). Let \( \beta(\theta) \) take values in a parameter space \( \Lambda \):

**DEFINITION 4**

A function \( B(\cdot) \) carrying a point \( y \) in \( \Omega \) into a set \( B(y) \) on \( \Lambda \) is a \( 1 - \alpha \) confidence region for the parameter \( \beta(\theta) \) if

\[
P(\beta(\theta) \in B(y) | \theta) \geq 1 - \alpha
\]

for each \( \theta \) in \( \Omega \).

Let \( A(\beta_{\star}) \) be an acceptance region for testing the hypothesis: \( \beta(\theta) = \beta_{\star} \) at level \( \alpha \):

\[
P(A(\beta_{\star}) | \theta) \geq 1 - \alpha
\]

for all \( \theta \) having \( \beta(\theta) = \beta_{\star} \).

Then the definition

\[
B(y) = \{ \beta : y \in A(\beta) \},
\]

gives the logical equivalence

\[
\beta \in B(y) \iff y \in A(\beta).
\]

From this formula (12) we then have

\[
P(\beta(\theta) \in B(y) | \theta) = P(y \in A(\beta(\theta)) | \theta)
\]

for \( \theta \in \Omega \), which established the following lemma.

**LEMMA 5**

\( B(\cdot) \) is a \( 1 - \alpha \) confidence region for \( \beta(\theta) \) if and only if, for each \( \beta_{\star} \) in \( \Lambda \), \( A(\beta_{\star}) \) is a \( 1 - \alpha \) acceptance region for testing \( \beta(\theta) = \beta_{\star} \).

**F EXERCISES**

Confidence intervals were derived for the normal location parameter in Section 8.3. The present Section D, together with the examples in Section 10.3A, show that those confidence intervals are uniformly most powerful unbiased; thus the location-parameter confidence intervals in the exercises in Section 8.3D are examples of UMP unbiased confidence intervals.

1 A sample of 5 was obtained on a nitrogen determination known to be normally distributed: 12.7, 13.3, 12.9, 13.0, and 13.1. On the assumption that \( \sigma^2 = 0.06 \), determine the UMP unbiased 95 percent confidence interval for the mean \( \mu \).
2 (continuation) With no information concerning \( \sigma \), determine the UMP unbiased 95 percent confidence interval for the mean \( \mu \).

Consider samples from the normal \((\mu_1, \sigma)\) and normal \((\mu_2, \sigma)\) with common variance \(\sigma^2\). A central 1 - \(\alpha\) confidence interval for \(\mu_2 - \mu_1\) was proposed in Problem 8.3.9; Theorem 10.3.6 with Section D shows that the confidence interval is a UMP unbiased confidence interval. A 1 - \(\alpha\) confidence interval for \(\sigma^2\) was proposed in Problem 8.3.7; the amount of probability on each tail of the chi-square distribution can, in fact, be chosen to produce an unbiased test and an unbiased confidence interval, which are then UMP unbiased by Theorem 10.3.6; usually, however, \(\alpha/2\) is taken on each tail rather than bothering with this unbiasedness adjustment.

3 Strength determinations for two different materials are known to be normally distributed with means \(\mu_1\) and \(\mu_2\) and common variance \(\sigma^2\). A sample of 5 on the first material gave \(\bar{y}_1 = 0.275\), \(s_1^2 = 0.00045\), and a sample of 6 on the second material gave \(\bar{y}_2 = 0.295\), \(s_2^2 = 0.00039\). Determine the UMP unbiased 95 percent confidence interval for \(\mu_2 - \mu_1\).

4 (continuation) Determine a central 95 percent confidence interval for the common variance \(\sigma^2\).

---

**G PROBLEMS**

5 Let \((y_1, \ldots, y_n)\) be a sample from the uniform\((0, \theta)\) with \(\theta\) in \(R^+\). Determine the UMP size 1 - \(\alpha\) confidence interval for \(\theta\); see Problem 10.2.7 and use the distribution results from Section 3.6E.

6 Let \((y_1, \ldots, y_n)\) be a sample from the uniform \((\theta - 1/2, \theta + 1/2)\) with \(\theta\) in \(R\). Give an expression for the confidence level using \((Y_{11}, Y_{n1})\) as a confidence interval for \(\theta\); some distribution results are available in Section 3.6E and Problem 3.6.13.

7 (continuation) Suppose that an investigator receives a sample having \(Y_{11} - Y_{n1} = r > 1/2\). What can he assert concerning the interval covering or not covering \(\theta\)? Use the distribution results in Problem 3.6.13 to determine the probability that \((Y_{11}, Y_{n1})\) brackets \(\theta\) given that \(Y_{11} - Y_{n1}\) has an observed value \(r\).

8 Let \(s_1^2\) be an unbiased estimate of \(\sigma_1^2\) with a scaled chi-square\((m)\) distribution, and \(s_2^2\) be an independent estimate of \(\sigma_2^2\) with a scaled chi-square\((n)\) distribution. Then show that

\[
\left( \frac{s_1^2}{s_2^2}, \frac{s_2^2}{s_1^2} \right) \sim F_{m, n} \]

is a 95 percent confidence interval for \(\sigma_1^2/\sigma_2^2\), where \(F_{m, n}\) is the \(\alpha\) point on the tail of the \(F(m, n)\) distribution.

---

**12.3 THREE-FACTOR DESIGN**

We now examine the fully randomized three-factor design. The analogous design with randomized blocks can be obtained using the methods in Section 12.4.

**A THE THREE-FACTOR DESIGN**

Consider a random system with three design factors \(A, B, C\); let the chosen levels for the factors be as follows:
Suppose that \( n \) response observations are made at each treatment \( A_i B_j C_k \) for a total of \( N = abcn \) observations. And suppose that the \( N \) observations have been fully randomized against experimental units, time, or any uncontrolled factor.

Let \( y_{ijk} \) be the \( s \)th observation at the treatment \( A_i B_j C_k \). A matrix array for \( AB \) as in Section 12.2 can be presented for each level of the factor \( C \); this can be useful for some of the calculations. In a similar manner, a matrix array for \( AC \) can be presented for each level of the factor \( B \), and a matrix array for \( BC \) can be presented for each level of the factor \( A \), these also can be useful for some of the calculations.

Let \( \mathbf{y} \) designate the vector in \( R^N \). And let \( \mathbf{1}_i \) be a vector with 1’s in the \( A_i \) positions and 0’s elsewhere; similarly for \( \mathbf{1}_j \) and \( \mathbf{1}_k \) with respect to \( B_j \) and \( C_k \). And let \( \mathbf{1}_{ij} \) be a vector with 1’s in the \( A_i B_j \) positions and 0’s elsewhere; similarly for \( \mathbf{1}_{ij} \) and \( \mathbf{1}_{jk} \). And let \( \mathbf{1}_{ijk} \) be a vector with 1’s in the \( A_i B_j C_k \) positions and 0’s elsewhere. Note the identities such as \( \sum_k 1_{ijk} = 1_{ij} \).

Various projections and squared lengths can be separated out using the two-factor design in \( A \) and \( B \) with \( cn \) observations, in \( A \) and \( C \) with \( bn \) observations, and in \( B \) and \( C \) with \( an \) observations. In particular, for the mean we obtain the projection and squared length:

\[
p_1 = \bar{y},
\]

\[
SS_1 = N\bar{y}^2 = T_{..}^2/N.
\]

For \( A \) effects, \( B \) effects, and \( C \) effects we obtain from Section 12.2 the following projections and squared lengths:

\[
p_2 = \sum_i (\bar{y}_{i..} - \bar{y})\mathbf{1}_{i..} - p_1 = \sum_i (\bar{y}_{i..} - \bar{y})\mathbf{1}_{i..}
\]

\[
SS_2 = bcn \sum_i (\bar{y}_{i..} - \bar{y})^2 = \sum_i T_{i..}^2/bcn - SS_1.
\]

\[
p_3 = \sum_j (\bar{y}_{.j.} - \bar{y})\mathbf{1}_{.j.} - p_1 = \sum_j (\bar{y}_{.j.} - \bar{y})\mathbf{1}_{.j.}
\]

\[
SS_3 = acn \sum_j (\bar{y}_{.j.} - \bar{y})^2 = \sum_j T_{.j.}^2/acn - SS_1.
\]

\[
p_4 = \sum_k (\bar{y}_{..k} - \bar{y})\mathbf{1}_{..k} - p_1 = \sum_k (\bar{y}_{..k} - \bar{y})\mathbf{1}_{..k}
\]

\[
SS_4 = abn \sum_k (\bar{y}_{..k} - \bar{y})^2 = \sum_k T_{..k}^2/abn - SS_1.
\]

These projection vectors are orthogonal vectors and they are obtained as projections of \( \mathbf{y} \) into orthogonal subspaces of dimensions \( a-1 \), \( b-1 \), and \( c-1 \); of course, each of these subspaces is orthogonal to \( \mathcal{L}(1) \).

The interactions \( AB \), \( BC \), and \( AC \) can also be separated out using the analysis of the two-factor design. The projection and squared length for the \( AB \) interaction, then \( BC \), then \( AC \) are:

\[
p_5 = \sum_{ij} (\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y})\mathbf{1}_{ij} - p_1 - p_2 - p_3 = \sum_{ij} (\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y})\mathbf{1}_{ij}.
\]
SS_5 = c n \sum \limits_{ij} (\bar{y}_{ij} - \bar{y}_{..} - \bar{y}_{.j} + \bar{y})^2 = \sum \limits_{ij} T^2_{ij}/cn - SS_1 - SS_2 - SS_3.

p_6 = \sum \limits_{jk} \bar{y}_{jk} 1_{jk} - p_1 - p_3 - p_4 = \sum \limits_{jk} (\bar{y}_{jk} - \bar{y}_{.j} - \bar{y}_{..k} + \bar{y}) 1_{jk}.

SS_6 = a n \sum \limits_{jk} (\bar{y}_{jk} - \bar{y}_{.j} - \bar{y}_{..k} + \bar{y})^2 = \sum \limits_{jk} T^2_{jk}/an - SS_1 - SS_3 - SS_4.

p_7 = \sum \limits_{ik} \bar{y}_{i-k} 1_{i-k} - p_1 - p_2 - p_4 = \sum \limits_{ik} (\bar{y}_{i-k} - \bar{y}_{..i} - \bar{y}_{..-k} + \bar{y}) 1_{i-k}.

SS_7 = b n \sum \limits_{ik} (\bar{y}_{i-k} - \bar{y}_{..i} - \bar{y}_{..-k} + \bar{y})^2 = \sum \limits_{ik} T^2_{i-k}/bn - SS_1 - SS_2 - SS_4.

Again we have orthogonality, a benefit of the factorial design. For consider a vector in the subspace for the AB interaction,
\[ \sum \limits_i c_{ij} 1_{ij} \text{ where } \sum \limits_i c_{ij} = 0 = \sum \limits_j c_{ij} \]
and a vector in the subspace for the AC interaction,
\[ \sum \limits_i b_{ik} 1_{i-k} \text{ where } \sum \limits_i b_{ik} = 0 = \sum \limits_k b_{ik}. \]

The inner product of 1_{ij} and 1_{i-k} is n; thus the inner product of \( \Sigma c_{ij} 1_{ij} \) and \( \Sigma b_{ik} 1_{i-k} \) is
\[ \sum \limits_{ijk} n c_{ij} b_{ik} = n \sum \limits_{ij} c_{ij} \sum \limits_k b_{ik} = 0. \]

Also it is easily checked, for example, that an AB interaction vector is orthogonal to a C-effect vector. Thus the two-way interactions are obtained as projections of y into orthogonal subspaces of dimensions \((a - 1)(b - 1), (b - 1)(c - 1), (a - 1)(c - 1)\), and each of these subspaces is orthogonal to the subspaces for the mean and for the main effects of the factors A, B, and C.

Now consider the ABC interaction. A vector giving the level \( \mu_{ijk} \) at the treatment \( A_i B_j C_k \) is
\[ \sum \limits_{ijk} \mu_{ijk} 1_{ijk}. \]

The vectors 1_{ijk} are orthogonal; hence the vectors \( \Sigma \mu_{ijk} 1_{ijk} \) form an abc-dimensional subspace of \( \mathbb{R}^N \). The projection of y into this subspace and the squared lengths are
\[ \sum \bar{y}_{ijk} 1_{ijk}, \]
\[ n \sum \bar{y}_{ijk}^2 = \sum \limits_{ijk} T^2_{ijk}/n. \]

Note the following vector relations:
\[ \sum \limits_k 1_{ijk} = 1_{ij}, \quad \sum \limits_k 1_{ijk} = 1_{i..}, \quad \sum \limits_{ijk} 1_{ijk} = 1. \]

Hence the subspaces for the mean, for the main effects, and for the two-way interactions are component subspaces of \( \mathbb{L}(1_{ijk}: \text{all } i, j, k) \). The three-way interaction ABC is the projection of y into
the orthogonal complement of these components within the space \( \mathcal{L}(1_{ijk}: \text{all } i, j, k) \); thus the projection and squared length are

\[
P_B = \sum \bar{y}_{ijk} 1_{ijk} - p_1 - p_2 - p_3 - p_4 - p_5 - p_6 - p_7
\]

\[
= \sum (\bar{y}_{ijk} - \bar{y}_{ij} - \bar{y}_{vk} - \bar{y}_{jk} + \bar{y}_{..} + \bar{y}_{..j} + \bar{y}_{..k} - \bar{y}) 1_{ijk},
\]

\[
SS_B = n \sum (\bar{y}_{ijk} - \bar{y}_{ij} - \bar{y}_{vk} - \bar{y}_{jk} + \bar{y}_{..} + \bar{y}_{..j} + \bar{y}_{..k} - \bar{y})^2
\]

\[
= \sum_{ijk} T_{ijk}^2 / n - SS_1 - SS_2 - SS_3 - SS_4 - SS_5 - SS_6 - SS_7.
\]

Note the substantial identity obtained from the two ways of deriving the sum of squares \( SS_B \).

The residual is obtained as the projection into \( \mathcal{L}(1_{ijk}: \text{all } i, j, k) \); the projection and squared length are

\[
P_2 = \bar{y} - \sum \bar{y}_{ijk} 1_{ijk},
\]

\[
SS_\varepsilon = \sum_{ijks} (y_{ijks} - \bar{y}_{ijk})^2 = \sum_{ijks} y_{ijks}^2 - \sum_{ijk} T_{ijk}^2 / n.
\]

The cumulative table recording projections and squared lengths is as shown in Table S.4; note that for the projections a typical \( ijks \) coordinate is recorded inside the parentheses indicating the vector. The omitted elements are available by direct analogy.

The ordinary table describing projections into orthogonal subspaces is then obtained by the differencing described earlier. Thus we obtain Table S.5, giving the separation of \( \bar{y} \) and \( \sum y_{ijks}^2 \) into components corresponding to the various dependencies on the factors \( A, B, \) and \( C \). The omitted elements are available by analogy. For computation the \( SS \) entries are best calculated by differencing using the second expression recorded in each case.

**TABLE S.4**
Cumulative table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Projection</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1</td>
<td>( \bar{y} )</td>
<td>( N\bar{y}^2 = \sum T_{ij..}^2 / N )</td>
</tr>
<tr>
<td>( A )</td>
<td>( a )</td>
<td>( \bar{y}_{ij..} )</td>
<td>( bcn \sum_i \bar{y}<em>{ij..}^2 = \sum_i T</em>{ij..}^2 / bcn )</td>
</tr>
<tr>
<td>( B )</td>
<td>( b )</td>
<td>( \bar{y}_{..j} )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( C )</td>
<td>( c )</td>
<td>( \bar{y}_{..k} )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( AB )</td>
<td>( ab )</td>
<td>( \bar{y}_{ij} )</td>
<td>( cn \sum_{ij} \bar{y}<em>{ij}^2 = \sum</em>{ij} T_{ij..}^2 / cn )</td>
</tr>
<tr>
<td>( BC )</td>
<td>( bc )</td>
<td>( \bar{y}_{ik} )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( AC )</td>
<td>( ac )</td>
<td>( \bar{y}_{i..k} )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( ABC )</td>
<td>( abc )</td>
<td>( \bar{y}_{ijk} )</td>
<td>( n \sum_{ijk} \bar{y}<em>{ijk}^2 = \sum T</em>{ijk}^2 / n )</td>
</tr>
<tr>
<td>Response</td>
<td>( abcn )</td>
<td>( y_{ijks} )</td>
<td>( \sum_{ijks} y_{ijks}^2 )</td>
</tr>
</tbody>
</table>
TABLE S.5
Analysis-of-variance table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1</td>
<td>(N\bar{y}^2 = \frac{T^2}{N})</td>
</tr>
<tr>
<td>A</td>
<td>(a - 1)</td>
<td>(bcn \sum (\bar{y}<em>{ij} - \bar{y})^2 = \sum T^2</em>{ij}/bcn - SS_1)</td>
</tr>
<tr>
<td>B</td>
<td>(b - 1)</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>C</td>
<td>(c - 1)</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>AB</td>
<td>((a - 1)(b - 1))</td>
<td>(cn \sum (\bar{y}<em>{ij} - \bar{y}</em>{..} - \bar{y}<em>{j} + \bar{y})^2 = \frac{T^2</em>{ijk}}{cn} - SS_1 - SS_2 - SS_3)</td>
</tr>
<tr>
<td>BC</td>
<td>((b - 1)(c - 1))</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>AC</td>
<td>((a - 1)(c - 1))</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>ABC</td>
<td>((a - 1)(b - 1)(c - 1))</td>
<td>(n \sum (\bar{y}<em>{ijk} - \bar{y}</em>{ij} - \bar{y}<em>{j} - \bar{y}</em>{i} + \bar{y}<em>{..} + \bar{y}</em>{j..} - \bar{y} - \bar{y}<em>{..k} + \bar{y}</em>{..k} - \bar{y})^2 = \sum T^2_{ijk}/n - \sum SS_a)</td>
</tr>
<tr>
<td>Residual</td>
<td>(abc(n - 1))</td>
<td>(\sum (\bar{y}<em>{ijkl} - \bar{y}</em>{ikl})^2 = \sum y^2_{ijkl} - \sum SS_a)</td>
</tr>
</tbody>
</table>

B  THE ANALYSIS

We consider the analysis for the case of approximately normal variation. For this we assume that the response \(y\) has variation \(\sigma z\), where \(z\) is standard normal. By Section 11.4, in the pattern discussed in Section 12.1C, we note that an SS entry has the distribution of \(\sigma^2 \chi^2\), where \(\chi^2\) has the chi-square distribution with degrees of freedom given by the corresponding DF entry; the chi-square distribution is ordinary chi-square if the mean vector has zero component in the particular subspace, and it is noncentral chi-square with values that tend to be larger if the mean vector has a nonzero component in the subspace.

The first test for dependence on the factors \(A\), \(B\), and \(C\) is made by comparing the mean square for the \(ABC\) interaction with the mean square for the residuals. Significance would indicate that the response level depended on the factors \(A\), \(B\), and \(C\); some more detailed analysis perhaps using regression on input variables could investigate the form of the dependence.

On the assumption that there is no \(ABC\) interaction, then the various two-factor interactions can be tested against the residual. A significant value would indicate dependence, an interactive dependence, on the corresponding pair of factors.

On the assumption that there are no two-way interactions, then the various main effects can be tested against the residual. A significant value would indicate dependence, a direct additive dependence, on the corresponding factor.

The form of the dependence on the factors \(A\), \(B\), and \(C\) can be investigated in part by separating out individual dimensions in the analysis of variance. This can be done by examining specific comparisons between groups of levels as in Section 12.2C. Or with input variables corresponding to the various factors, it can be done by separating out specific linear and quadratic dependencies following again the pattern in Section 12.2C.

C  THE 2^3 DESIGN

The analysis of the three-factor design is particularly simple if each factor is examined at just two levels: the 2^3 design. The same simplicity carries over to the 2^r design with \(r\) factors each at two levels.
TABLE S.6
Analysis-of-variance table

<table>
<thead>
<tr>
<th>Source</th>
<th>Dimension</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1</td>
<td>$u^2$</td>
</tr>
<tr>
<td>$A$</td>
<td>1</td>
<td>$u_A^2$</td>
</tr>
<tr>
<td>$B$</td>
<td>1</td>
<td>$u_B^2$</td>
</tr>
<tr>
<td>$C$</td>
<td>1</td>
<td>$u_C^2$</td>
</tr>
<tr>
<td>$AB$</td>
<td>1</td>
<td>$u_{AB}^2$</td>
</tr>
<tr>
<td>$BC$</td>
<td>1</td>
<td>$u_{BC}^2$</td>
</tr>
<tr>
<td>$AC$</td>
<td>1</td>
<td>$u_{AC}^2$</td>
</tr>
<tr>
<td>$ABC$</td>
<td>1</td>
<td>$u_{ABC}^2$</td>
</tr>
<tr>
<td>$8$</td>
<td></td>
<td>$\sum y^2$</td>
</tr>
</tbody>
</table>

For simplicity consider the case with $n = 1$ and suppose that we know that there is no three-factor interaction so that the corresponding SS entry can be used to provide an estimate of the error variance. The more general case with $n > 1$ follows easily from the analysis here and earlier in this section.

For notation let $y_{ab}$, for example, be the response at the upper level of the first and second factors, $A$ and $B$, and the lower level of the third factor $C$. And consider the following orthogonal transformation to a new coordinate vector $u$:

$\begin{pmatrix} u \\ u_A \\ u_B \\ u_C \\ u_{AB} \\ u_{BC} \\ u_{AC} \\ u_{ABC} \end{pmatrix} = \frac{1}{\sqrt{8}} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \end{pmatrix} \begin{pmatrix} y_{abc} \\ y_{ab} \\ y_{ac} \\ y \\ y_{bc} \\ y_{b} \\ y_{c} \\ y \end{pmatrix}$

Note, for example, that $\sqrt{8} u_A$ has +1 for each observation at the upper level of $A$ and -1 for each observation at the lower level of $A$. And note that $\sqrt{8} u_{AB}$ has coefficients that are the product of those for $A$ and for $B$; it examines the A-factor difference as it changes from lower to upper level for the factor $B$. Thus each of the $u$'s corresponds to a particular entry of the analysis-of-variance table. But we have only one dimension for each entry; hence the table is as shown in Table S.6. For precise assessments this design needs to be replicated so as to have an estimate for error separate from the main effects and interactions.

**D EXERCISE**

1 A three-factor design was used to examine a spinning band laboratory fractionating column. The factors involved were $A$, band clearance; $B$, boil-up rate; and $C$, band speed; the response was the number of equivalent theoretical plates based on a refractive-index calculation. The data are as follows:

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$C$</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>1</td>
<td>750</td>
<td>1500</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>750</td>
<td>1500</td>
</tr>
<tr>
<td>0.10</td>
<td>1</td>
<td>750</td>
<td>1500</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>750</td>
<td>1500</td>
</tr>
</tbody>
</table>
Calculate the analysis-of-variance table and consider appropriate tests.

12.4

RANDOMIZED BLOCKS

At the beginning of Chapter 12 we briefly discussed the need to randomly assign treatments to experimental units. This randomization changes the inevitable variation among units into the dominant variation affecting response values. The ability, then, of a design to detect treatment effects depends fundamentally on the magnitude of treatment effects in comparison with the variation among units. Any reduction in this variation among units thus produces a corresponding increase in the ability of a design to detect treatment effects.

Now suppose that units or times or locations can be grouped into blocks on the basis of some accessible characteristic so that in general there is less variation between units from the same block than between units from different blocks. If the number of units in a block is equal to the number of treatments under consideration and if the treatments are randomly assigned within each block, then the design is called the randomized block design. The response variation that is appropriate to judging treatment comparisons is then the variation within blocks. Effective grouping to form blocks can thus produce a substantial increase in the sensitivity of the design. In this section we discuss some aspects of randomized blocks.

If an experimental unit has some numerical variable or characteristic that measures the effective variation between units, then such a concomitant variable can account for a substantial part of the response variation. In the next section we examine how the analysis of covariance can be used to remove the effects of such variation and thus increase the sensitivity of the design.

A UNIFORMITY TRIALS

In agricultural experimentation a response observation will typically require a full growing season from the time of implementation. Plots of land will often differ substantially in their effects on growing material. These and many other things place tremendous emphasis on the need for careful design and proper analysis. As part of the investigation of experimental procedures, plots of land are sometimes seeded and grown without the application of any treatments. Such investigations are called uniformity trials and they can give some indications concerning the effectiveness of techniques such as blocking and covariance analysis.

Several researchers conducted uniformity trials on the root weights for marigolds. The data obtained for five blocks, each consisting of four plots, are shown in Table S.7. The total sum of squares for the 20 observations is 2,276,781.

If the data are analyzed as a fully randomized design, Table S.8 is obtained. The estimated

<table>
<thead>
<tr>
<th>B_1</th>
<th>B_2</th>
<th>B_3</th>
<th>B_4</th>
<th>B_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>376</td>
<td>316</td>
<td>326</td>
<td>317</td>
<td>321</td>
</tr>
<tr>
<td>371</td>
<td>338</td>
<td>326</td>
<td>343</td>
<td>332</td>
</tr>
<tr>
<td>355</td>
<td>336</td>
<td>335</td>
<td>330</td>
<td>317</td>
</tr>
<tr>
<td>356</td>
<td>356</td>
<td>343</td>
<td>327</td>
<td>318</td>
</tr>
<tr>
<td>1458</td>
<td>1346</td>
<td>1330</td>
<td>1317</td>
<td>1288</td>
</tr>
</tbody>
</table>

2,274,953.25  2,270,706.05
TABLE S.8
Analysis-of-variance table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1</td>
<td>2,270,706.05</td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>19</td>
<td>6,074.95</td>
<td>319.73</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>2,276,781.00</td>
<td></td>
</tr>
</tbody>
</table>

variance for the variation from plot to plot is 319.73 and the standard deviation is 17.88.

Now if the data are analyzed as a randomized block design that recognizes the five distinct blocks of land, Table S.9 is obtained. The estimated variance from plot to plot within blocks is 121.85 and the standard deviation is 11.04. Thus the recognition of the blocks has produced a substantial reduction in the variation that would affect any treatment comparisons that could have been under investigation.

TABLE S.9
Analysis-of-variance table; separation of block components

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1</td>
<td>2,270,706.05</td>
<td></td>
</tr>
<tr>
<td>Blocks</td>
<td>4</td>
<td>4,247.20</td>
<td>1061.8</td>
</tr>
<tr>
<td>Residual</td>
<td>15</td>
<td>1,827.75</td>
<td>121.85</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>2,276,781.00</td>
<td></td>
</tr>
</tbody>
</table>

B THE RANDOMIZED BLOCK DESIGN

Now consider treatments \( T_1, \ldots, T_t \), and suppose that we have \( b \) blocks each consisting of \( t \) plots. The randomized block design is obtained by assigning treatments to units randomly within each block. The resulting response values can be presented in an array as follows:

\[
\begin{array}{cc}
B_1 & B_2 \\
T_1 & \begin{bmatrix} y_{11} & \cdots & y_{1b} \\ \vdots & \vdots \\ y_{t1} & \cdots & y_{tb} \end{bmatrix} \\
T_t & \\
\end{array}
\]

The analysis is based on separating out the mean accounting for one dimension, then the block differences accounting for \( b - 1 \) dimensions; the remaining \( tb - b = (t - 1)b \) dimensions refer to differences within blocks. The analysis then separates out \( t - 1 \) dimensions for the treatment differences, leaving \((t - 1)(b - 1)\) dimensions referring to other differences within blocks. This separation is the standard separation for the two-factor design in Section 12.2 and we obtain Table S.10. The test for treatment effects based on normal variation uses the \( F \) for treatment MS divided by residual MS.

C EXAMPLE

Suppose that we have two types of fertilizer, \( F_1 \) and \( F_2 \), and two types of ground preparation, \( G_1 \) and \( G_2 \). The two-factor design then involves four treatments \( T_1 = F_1G_1, T_2 = F_1G_2, T_3 = F_2G_1 \),
TABLE S.10
Analysis-of-variance table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1</td>
<td>$T_0^2/\text{tb}$</td>
</tr>
<tr>
<td>Blocks</td>
<td>$b - 1$</td>
<td>$\sum T_i^2/t - T_0^2/\text{tb}$</td>
</tr>
<tr>
<td>Treatments</td>
<td>$t - 1$</td>
<td>$\sum T_i^2/b - T_0^2/\text{tb}$</td>
</tr>
<tr>
<td>Residual</td>
<td>$(t - 1)(b - 1)$</td>
<td>$\sum y_{ij}^2/tb$</td>
</tr>
</tbody>
</table>

and $T_4 = F_2G_2$. And suppose that these four treatments are randomly assigned to four units in each of five blocks.

The data in Section A can be used to construct an example here. Four "treatment effects" were chosen and then randomly assigned within each block to the numbers as given in Section A; the following data were obtained:

<table>
<thead>
<tr>
<th>$T_1$</th>
<th>$T_2$</th>
<th>$T_3$</th>
<th>$T_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>345</td>
<td>365</td>
<td>366</td>
<td>382</td>
</tr>
<tr>
<td>327</td>
<td>305</td>
<td>347</td>
<td>367</td>
</tr>
<tr>
<td>324</td>
<td>315</td>
<td>337</td>
<td>354</td>
</tr>
<tr>
<td>306</td>
<td>332</td>
<td>338</td>
<td>341</td>
</tr>
<tr>
<td>321</td>
<td>306</td>
<td>329</td>
<td>332</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1623</td>
<td>1623</td>
<td>1717</td>
<td>1776</td>
</tr>
<tr>
<td>2,274,953.25</td>
<td>2,270,706.05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The total sum of squares is 2,279,795.

The analysis-of-variance table for this randomized block design is shown in Table S.11. The estimated variance from plot to plot within blocks is 120.27 and the standard deviation is 10.97.

If we ignore the rather special nature of the four treatments, we could test in general for treatment effects by calculating $F$ for the treatment MS divided by the residual MS. The observed value is 9.42 and the 5 and 1 percent points for $F$ on 3 over 12 degrees of freedom are 3.49 and 5.95; the observed value is extreme, giving strong evidence for treatments effects.

The analysis, however, must take account of the two-factor design for the treatments. The totals and some related two-factor calculations are

<table>
<thead>
<tr>
<th>$G_1$</th>
<th>$G_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$</td>
<td>1623</td>
</tr>
<tr>
<td>$F_2$</td>
<td>1717</td>
</tr>
<tr>
<td></td>
<td>3340</td>
</tr>
</tbody>
</table>

The three dimensions for treatments can be separated into one dimension for $F$, one dimension for $G$, and one dimension for the $F \times G$ interaction (Table S.12).

The first test for dependence on the factors $F$ and $G$ is made by calculating $F$ for the interactive effect $F \times G$; the observed value is 1.45, and the 5 and 1 percent values on 1 over 12 degrees of freedom are 4.75 and 9.33.

On the assumption that there is no interactive effect, tests for $F$ effects and $G$ effects can be
made by examining the corresponding MS's divided by the residual MS: for the factor G there is no significance; for the factor F the observed value of the function F is 25.36, and the 5 and 1 percent values are 4.74 and 9.33. The test for dependence on the factor F is highly significant, providing very substantial evidence for effects due to fertilizers.

The data for this example were constructed from the uniformity trial data by using the effects +11 for $F_2$ observations and $-11$ for $F_1$ observations and no other effects; the treatment effects were, of course, randomized to plots within each block.

D EXERCISE

1 Consider an experiment to test the effect of temperature on the yield of a chemical process. Six samples of raw material are obtained from each of 3 batches of material; each is used as an input for the process thus giving 18 runs. The temperatures are assigned randomly within batches, each temperature is used for 2 runs within each batch.

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>500</td>
<td>27.3</td>
</tr>
<tr>
<td></td>
<td>27.4</td>
</tr>
<tr>
<td>550</td>
<td>30.4</td>
</tr>
<tr>
<td></td>
<td>30.4</td>
</tr>
<tr>
<td>600</td>
<td>29.5</td>
</tr>
<tr>
<td></td>
<td>29.3</td>
</tr>
</tbody>
</table>

Construct the analysis-of-variance table separating Mean, Blocks, Temperature, $B \times T$, Determinations. Make appropriate tests; the experimental error for testing temperature is given by $B \times T$. 

12.5
ANALYSIS OF COVARIANCE

In the preceding section we considered how experimental units can sometimes be organized into blocks with less variation within blocks than between blocks. The restricted randomization of treatments-to-units-within-blocks then gives a design that typically has greater precision and greater sensitivity.

In a variety of problems, however, there may be one or several variables that indicate or measure the fluctuations attributable to the units. The use of such concomitant variables in addition to randomization and perhaps blocking can frequently reduce the relevant variation from unit to unit and thus increase the precision and sensitivity. The analysis, however, involves the essential use of the regression methods from Sections 11.2 and 11.3. And it goes outside the basic experimental methods of manipulating factors, controlling factors, and randomizing against possible effects of uncontrolled factors, and it does this primarily because these variables cannot be manipulated, just recorded.

Consider some examples. The effect of several diets on the weight of a certain kind of animal; a possible concomitant variable is the initial weight of an animal. The effect of various treatments on the yield of a particular crop; a possible concomitant variable is the preceding year’s yield if uniformity trials were performed. The effects of several treatments on the yield of rubber trees; a possible concomitant variable is some average yield from preceding years if the yields were obtained without treatments.

The computational methods for the factorial designs are organized to take advantage of orthogonality within sets of 1-vectors and between certain subspaces. By contrast the direct regression methods of Section 11.2 used a step-by-step procedure that calculates inner products of residual vectors. If we now introduce concomitant variables to a factorial design, we have an essential need for the step-by-step procedure. But—we can benefit computationally by using the orthogonality as far as possible and just terminally using the step-by-step procedure (which usually destroys orthogonality relations). We recall from Section 11.2 that the projection into an orthogonal complement and its squared length do not depend on the order in which components are removed.

A THE FULLY RANDOMIZED DESIGN

Consider a fully randomized design with response \( y \) and treatments \( T_1, \ldots, T_t \). Let \( x \) be a concomitant variable that can be observed for each experimental unit, and suppose for simplicity here that the response level without treatments is at most linear in \( x \), linearity providing in many cases a reasonable approximation over a moderate range. The data array can be presented as

\[
\begin{array}{cccc}
T_1 & x_{11} & \cdots & x_{1n_t} \\
& y_{11} & \cdots & y_{1n_t} \\
\vdots & \vdots & & \vdots \\
T_t & x_{t1} & \cdots & x_{tn_t} \\
& y_{t1} & \cdots & y_{tn_t} \\
\end{array}
\]

If the treatments are part of a factorial design, then \( n_1 = \cdots = n_t \). The data can appear as in Figure S.13.

The investigation typically will involve the following succession of models: an overall general level, \( y = \beta_1 + \sigma z \); dependence on the concomitant variables, which here is affine dependence on \( x \), \( y = \beta_1 1 + \beta_2 x + \sigma z \); dependence on the concomitant variables with allowance for different levels for different treatments, \( y = \Sigma \gamma_i 1_i + \delta x + \sigma z \), where \( 1_i \) is the 1-vector for treatment \( T_i \).

This succession of models is represented by the sequence of solid arrows in the following display presenting fitted models:
Sec. 12.5: Analysis of Covariance

FIGURE S.13
Three observations at each of three treatments.

\[ y = b_1 + \quad \quad \quad \rightarrow \quad y = b_1 + b_2 x + \]

\[ y = \sum c_i x_i + \quad \quad \rightarrow \quad y = \sum c_i x_i + dx + \]

We have noted in the introductory remarks that there are computational advantages to using the simplicity and orthogonality of factorial designs as long as possible and only terminally using the regression procedures. In the preceding array this corresponds to following the dashed arrows involving the various fits appropriate to a factorial design and then using regression to obtain the fitted model in the right-hand column. This is the computation procedure we follow in this section.

First we consider the cumulative table appropriate to the particular factorial design. Ordinarily, this records the squared length of the projections of the response in the succession of subspaces. In the present case we record the inner product matrix for the projections of both \( y \) and the concomitant vectors in the succession of subspaces. For example, the projections of \( x \) and \( y \) on the vector 1 are \( \bar{x} \) and \( \bar{y} \), where \( \bar{x} = T(x)/N \) and \( \bar{y} = T(y)/N \), and these capital \( T \)'s are used to denote totals as in earlier sections of this chapter; thus the inner product matrix is

\[
\begin{pmatrix}
  T_x^2(x)/N & T_x(x)T_y(y)/N \\
  T_x(y)T_y(x)/N & T_y^2(y)/N
\end{pmatrix}.
\]

In this way we obtain Table S.13.

The values in Table S.13 are now complemented with respect to the inner product matrix,

| TABLE S.13
Cumulative table |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>DF</td>
<td>SPxx</td>
<td>SPxy</td>
</tr>
<tr>
<td>(1) Mean</td>
<td>1</td>
<td>( T_x^2(x)/N )</td>
<td>( T_x(x)T_y(y)/N )</td>
</tr>
<tr>
<td>(2) Treatments</td>
<td>( t )</td>
<td>( \sum T_x^2(x)/n_i )</td>
<td>( \sum T_x(x)T_y(y)/n_i )</td>
</tr>
<tr>
<td>(3) Response</td>
<td>( N )</td>
<td>( \sum x_{ij}^2 )</td>
<td>( \sum x_{ij}y_{ij} )</td>
</tr>
</tbody>
</table>
for the initial vectors. By doing this we obtain the successive inner product matrices for residuals (Table S.14).

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SPxx</th>
<th>SPxy</th>
<th>SPyy</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0] Null</td>
<td>N</td>
<td>$\sum x_i^2$</td>
<td>$\sum x_i y_i$</td>
<td>$\sum y_i^2$</td>
</tr>
<tr>
<td>[1] Mean</td>
<td>$N - 1$</td>
<td>$\sum x_i^2 - \frac{T^2_x(x)}{N}$</td>
<td>$\sum x_i y_i - \frac{T_x(y) T_x(y)}{N}$</td>
<td>$\sum y_i^2 - \frac{T^2_y(y)}{N}$</td>
</tr>
<tr>
<td>[2] Treatments</td>
<td>$N - t$</td>
<td>$\sum x_i^2 - \sum \frac{T^2_x(x)}{n_i}$</td>
<td>$\sum x_i y_i - \sum \frac{T_x(y) T_i(y)}{n_i}$</td>
<td>$\sum y_i^2 - \sum \frac{T^2_y(y)}{n_i}$</td>
</tr>
</tbody>
</table>

Now recall the step-by-step procedure described in Section 11.2. For example, to introduce linear regression on $x$ after fitting the mean involves a single step using the residuals of $y$ and $x$ with respect to the mean (note the penultimate remarks in the third paragraph of Section 12.3E); thus in the present notation the regression coefficient is $SP_{11}^{xy}/SP_{11}^{xx}$ using the values in row [1], and the reduction in the sum of squares for the response is given by $(SP_{11}^{xy})^2/SP_{11}^{xx}$.

We now record in Table S.15 the sum of squares of residuals obtained following the solid arrows in the display (1) earlier in this section. Note that a prime indicates a shift from the left side of the display to the right side.

The sum of squares for the orthogonal components can be obtained by differencing Table S.15, giving the results shown in Table S.16. Thus with allowance made for linear dependence on $x$ we can test for treatment effects by calculating the treatment MS divided by residual MS in Table S.16.

**B EXAMPLE**

Consider a response $y$ whose level is known to adjust linearly with respect to a concomitant variable $x$. And suppose that three treatments $T_1$, $T_2$, and $T_3$ are applied treatments that could provide an additive adjustment to the response level. The data shown in Table S.17 were obtained. Some calculated results are given with the table:

$$35.64 = \sum T_i^2(x)/n_i, \quad 84.5521 = \sum T_i^2(y)/n_i,$$
$$52.5413 = \sum T_i(x) T_i(y)/n_i;$$
the total sums of squares and products are

$$\begin{align*}
35.64 & \quad 84.5521 \\
52.5413 & \quad 40.12 \\
& \quad 90.9700 \\
57.815 &
\end{align*}$$

The cumulative table recording inner products for projections into the successive linear spaces is shown in Table S.18.
### TABLE S.15
Cumulative table: residuals

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS Reduction</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0] Null</td>
<td>$N$</td>
<td>$S_{[0]}yy = \sum y_i^2$</td>
<td></td>
</tr>
<tr>
<td>[1] Mean</td>
<td>$N - 1$</td>
<td>$S_{[1]}yy = \sum y_i^2 - \frac{T^2(y)}{N}$</td>
<td></td>
</tr>
<tr>
<td>[1'] Linear (x)</td>
<td>$N - 2$</td>
<td>$S_{[1]}^{2}xy \over S_{[1]}^{2}xx$</td>
<td>$S_{[1]}^{2}yy \over S_{[1]}^{2}xx$</td>
</tr>
<tr>
<td>[2'] Treatments</td>
<td>$N - t - 1$</td>
<td>$S_{[2]}^{2}xy \over S_{[2]}^{2}xx$</td>
<td>$S_{[2]}^{2}yy \over S_{[2]}^{2}xx$</td>
</tr>
</tbody>
</table>

### TABLE S.16
Analysis-of-covariance table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Mean</td>
<td>1</td>
<td>$\frac{T^2(y)}{N}$</td>
</tr>
<tr>
<td>1' Linear (x)</td>
<td>1</td>
<td>$S_{[1]}^{2}xy \over S_{[1]}^{2}xx$</td>
</tr>
<tr>
<td>2' Treatments</td>
<td>$t - 1$</td>
<td>$S_{[1]}^{2}yy - S_{[2]}^{2}yy - \frac{S_{[1]}^{2}xy}{S_{[1]}^{2}xx}$</td>
</tr>
<tr>
<td>3 Residual</td>
<td>$N - t - 1$</td>
<td>$S_{[2]}^{2}yy \over S_{[2]}^{2}xx$</td>
</tr>
</tbody>
</table>

### TABLE S.17
Treatment data

<table>
<thead>
<tr>
<th>$T_1$</th>
<th>$T_2$</th>
<th>$T_3$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x$</td>
<td>$y$</td>
<td>$x$</td>
<td>$y$</td>
</tr>
<tr>
<td>1.1</td>
<td>0.95</td>
<td>1.4</td>
<td>2.43</td>
</tr>
<tr>
<td>2.2</td>
<td>2.42</td>
<td>2.0</td>
<td>2.88</td>
</tr>
<tr>
<td>3.1</td>
<td>3.04</td>
<td>3.0</td>
<td>4.44</td>
</tr>
<tr>
<td>6.4</td>
<td>6.41</td>
<td>6.4</td>
<td>9.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>35.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>52.5413</td>
</tr>
</tbody>
</table>

### TABLE S.18
Cumulative table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>$SP_{xx}$</th>
<th>$SP_{xy}$</th>
<th>$SP_{yy}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Mean</td>
<td>1</td>
<td>35.2044</td>
<td>53.4000</td>
<td>81.0000</td>
</tr>
<tr>
<td>(2) Treatments</td>
<td>3</td>
<td>35.6400</td>
<td>52.5413</td>
<td>84.5521</td>
</tr>
<tr>
<td>(3) Response</td>
<td>9</td>
<td>40.1200</td>
<td>57.8150</td>
<td>90.9700</td>
</tr>
</tbody>
</table>
TABLE S.19  
Cumulative table; residuals

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SPxx</th>
<th>SPxy</th>
<th>SPyy</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0] Null</td>
<td>9</td>
<td>40.1200</td>
<td>57.8150</td>
<td>90.9700</td>
</tr>
</tbody>
</table>

TABLE S.20  
Cumulative table; residuals

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS Reduction</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0] Null</td>
<td>9</td>
<td>90.9700</td>
<td></td>
</tr>
<tr>
<td>[1] Mean</td>
<td>8</td>
<td>9.9700</td>
<td></td>
</tr>
<tr>
<td>[1] Linear (x)</td>
<td>7</td>
<td>3.9654</td>
<td>6.0046</td>
</tr>
<tr>
<td>[2] Treatments</td>
<td>5</td>
<td>6.2080</td>
<td>0.2099</td>
</tr>
</tbody>
</table>

The inner products for residuals are obtained by complementing with respect to (3), Table S.19. The regression on x can now be removed at each stage after the mean, Table S.20.

The sums of squares for the orthogonal components relative to the succession of models (1) is now obtained by differencing; see Table S.21. The observed $F$ for testing treatments is 69 and the 5 and 1 percent points are 5.79 and 13.27. Under the assumed model this provides overwhelming evidence against the hypothesis of no treatment effects. It is easy to check that a test neglecting covariance analysis would give an $F$ value just larger than 1; and this would give no evidence against the hypothesis of no treatment effects.

TABLE S.21  
Analysis-of-covariance table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>81.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1’ Linear (x)</td>
<td>1</td>
<td>3.9654</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2’ Treatments</td>
<td>2</td>
<td>5.7947</td>
<td>2.897</td>
<td>$F = 69$</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.2099</td>
<td>0.042</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>90.9700</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

C  
EXERCISE

1  
Fourteen pigs were randomly assigned to 3 diets for a 2-month period; see Exercise 12.1.1. The following table records (initial weight, final weight) for each pig:

<table>
<thead>
<tr>
<th>Diet</th>
<th>Initial Weight</th>
<th>Final Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>T₁</td>
<td>(69, 82)</td>
<td>(73, 96)</td>
</tr>
<tr>
<td>T₂</td>
<td>(52, 74)</td>
<td>(59, 90)</td>
</tr>
<tr>
<td>T₃</td>
<td>(53, 86)</td>
<td>(58, 98)</td>
</tr>
</tbody>
</table>

Calculate the analysis-of-variance table and make appropriate tests for dependence on the diets.
### APPENDIX II

#### TABLES

<table>
<thead>
<tr>
<th>$z_\alpha$</th>
<th>0.00</th>
<th>0.01</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.05</th>
<th>0.06</th>
<th>0.07</th>
<th>0.08</th>
<th>0.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.5000</td>
<td>0.4980</td>
<td>0.4920</td>
<td>0.4880</td>
<td>0.4840</td>
<td>0.4801</td>
<td>0.4761</td>
<td>0.4721</td>
<td>0.4681</td>
<td>0.4641</td>
</tr>
<tr>
<td>0.1</td>
<td>0.6020</td>
<td>0.5632</td>
<td>0.5252</td>
<td>0.4843</td>
<td>0.4443</td>
<td>0.4044</td>
<td>0.3644</td>
<td>0.3245</td>
<td>0.2846</td>
<td>0.2436</td>
</tr>
<tr>
<td>0.2</td>
<td>0.7027</td>
<td>0.6168</td>
<td>0.5299</td>
<td>0.4390</td>
<td>0.3490</td>
<td>0.2590</td>
<td>0.1690</td>
<td>0.0790</td>
<td>0.0390</td>
<td>0.0190</td>
</tr>
<tr>
<td>0.3</td>
<td>0.8012</td>
<td>0.7583</td>
<td>0.6754</td>
<td>0.5670</td>
<td>0.4670</td>
<td>0.3670</td>
<td>0.2670</td>
<td>0.1670</td>
<td>0.0670</td>
<td>0.0170</td>
</tr>
<tr>
<td>0.4</td>
<td>0.8996</td>
<td>0.8409</td>
<td>0.7372</td>
<td>0.6132</td>
<td>0.4932</td>
<td>0.3632</td>
<td>0.2432</td>
<td>0.1232</td>
<td>0.0232</td>
<td>0.0032</td>
</tr>
<tr>
<td>0.5</td>
<td>0.9805</td>
<td>0.9050</td>
<td>0.8205</td>
<td>0.7105</td>
<td>0.5805</td>
<td>0.4405</td>
<td>0.3005</td>
<td>0.1405</td>
<td>0.0305</td>
<td>0.0005</td>
</tr>
<tr>
<td>0.6</td>
<td>0.9524</td>
<td>0.8729</td>
<td>0.7567</td>
<td>0.6167</td>
<td>0.4867</td>
<td>0.3467</td>
<td>0.2067</td>
<td>0.0867</td>
<td>0.0267</td>
<td>0.0067</td>
</tr>
<tr>
<td>0.7</td>
<td>0.9189</td>
<td>0.8370</td>
<td>0.7109</td>
<td>0.5809</td>
<td>0.4409</td>
<td>0.3009</td>
<td>0.1609</td>
<td>0.0609</td>
<td>0.0209</td>
<td>0.0009</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8772</td>
<td>0.8071</td>
<td>0.6901</td>
<td>0.5501</td>
<td>0.4201</td>
<td>0.2901</td>
<td>0.1501</td>
<td>0.0501</td>
<td>0.0101</td>
<td>0.0001</td>
</tr>
<tr>
<td>0.9</td>
<td>0.8320</td>
<td>0.7732</td>
<td>0.6562</td>
<td>0.5262</td>
<td>0.3962</td>
<td>0.2762</td>
<td>0.1462</td>
<td>0.0462</td>
<td>0.0062</td>
<td>0.0000</td>
</tr>
<tr>
<td>1.0</td>
<td>0.7868</td>
<td>0.7162</td>
<td>0.5939</td>
<td>0.4739</td>
<td>0.3439</td>
<td>0.2239</td>
<td>0.1039</td>
<td>0.0039</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1.1</td>
<td>0.7357</td>
<td>0.6535</td>
<td>0.5135</td>
<td>0.3835</td>
<td>0.2635</td>
<td>0.1435</td>
<td>0.0235</td>
<td>0.0035</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1.2</td>
<td>0.6815</td>
<td>0.5911</td>
<td>0.4411</td>
<td>0.3011</td>
<td>0.1811</td>
<td>0.0611</td>
<td>0.0011</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1.3</td>
<td>0.6258</td>
<td>0.5363</td>
<td>0.3863</td>
<td>0.2463</td>
<td>0.1263</td>
<td>0.0063</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1.4</td>
<td>0.5681</td>
<td>0.4852</td>
<td>0.3352</td>
<td>0.1952</td>
<td>0.0952</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1.5</td>
<td>0.5094</td>
<td>0.4363</td>
<td>0.2863</td>
<td>0.1463</td>
<td>0.0463</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1.6</td>
<td>0.4497</td>
<td>0.3870</td>
<td>0.2370</td>
<td>0.0970</td>
<td>0.0070</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1.7</td>
<td>0.3887</td>
<td>0.3263</td>
<td>0.1763</td>
<td>0.0363</td>
<td>0.0063</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1.8</td>
<td>0.3275</td>
<td>0.2643</td>
<td>0.1143</td>
<td>0.0143</td>
<td>0.0043</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1.9</td>
<td>0.2660</td>
<td>0.2043</td>
<td>0.0543</td>
<td>0.0043</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>2.0</td>
<td>0.2042</td>
<td>0.1443</td>
<td>0.0243</td>
<td>0.0043</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>2.1</td>
<td>0.1441</td>
<td>0.0841</td>
<td>0.0041</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>2.2</td>
<td>0.0840</td>
<td>0.0240</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>2.3</td>
<td>0.0239</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>2.4</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

**TABLE 1**

Normal(0, 1) distribution: right-tail probabilities $\alpha = P(z \geq z_\alpha) = 1 - G(z_\alpha)$
### Table 2: Student(\( f \)) Distribution

Student(\( f \)) distribution: right-tail percentage points \( t_{\alpha} \); \( \alpha = P(t \geq t_{\alpha}) \)

<table>
<thead>
<tr>
<th>( f )</th>
<th>0.25</th>
<th>0.1</th>
<th>0.05</th>
<th>0.025</th>
<th>0.01</th>
<th>0.005</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.000</td>
<td>3.078</td>
<td>6.314</td>
<td>12.706</td>
<td>31.821</td>
<td>63.657</td>
</tr>
<tr>
<td>2</td>
<td>0.816</td>
<td>1.886</td>
<td>2.920</td>
<td>4.303</td>
<td>6.965</td>
<td>9.925</td>
</tr>
<tr>
<td>3</td>
<td>0.765</td>
<td>1.638</td>
<td>2.353</td>
<td>3.182</td>
<td>4.541</td>
<td>5.841</td>
</tr>
<tr>
<td>4</td>
<td>0.741</td>
<td>1.533</td>
<td>2.132</td>
<td>2.776</td>
<td>3.747</td>
<td>4.604</td>
</tr>
<tr>
<td>5</td>
<td>0.727</td>
<td>1.476</td>
<td>2.015</td>
<td>2.571</td>
<td>3.365</td>
<td>4.032</td>
</tr>
<tr>
<td>6</td>
<td>0.718</td>
<td>1.440</td>
<td>1.943</td>
<td>2.447</td>
<td>3.143</td>
<td>3.707</td>
</tr>
<tr>
<td>7</td>
<td>0.711</td>
<td>1.415</td>
<td>1.895</td>
<td>2.365</td>
<td>2.998</td>
<td>3.499</td>
</tr>
<tr>
<td>8</td>
<td>0.706</td>
<td>1.397</td>
<td>1.860</td>
<td>2.306</td>
<td>2.896</td>
<td>3.355</td>
</tr>
<tr>
<td>9</td>
<td>0.703</td>
<td>1.383</td>
<td>1.833</td>
<td>2.262</td>
<td>2.821</td>
<td>3.250</td>
</tr>
<tr>
<td>10</td>
<td>0.700</td>
<td>1.372</td>
<td>1.812</td>
<td>2.228</td>
<td>2.764</td>
<td>3.169</td>
</tr>
<tr>
<td>11</td>
<td>0.697</td>
<td>1.363</td>
<td>1.796</td>
<td>2.201</td>
<td>2.718</td>
<td>3.106</td>
</tr>
<tr>
<td>12</td>
<td>0.695</td>
<td>1.356</td>
<td>1.782</td>
<td>2.179</td>
<td>2.681</td>
<td>3.055</td>
</tr>
<tr>
<td>13</td>
<td>0.694</td>
<td>1.350</td>
<td>1.771</td>
<td>2.160</td>
<td>2.650</td>
<td>3.012</td>
</tr>
<tr>
<td>14</td>
<td>0.692</td>
<td>1.345</td>
<td>1.761</td>
<td>2.145</td>
<td>2.624</td>
<td>2.977</td>
</tr>
<tr>
<td>15</td>
<td>0.691</td>
<td>1.341</td>
<td>1.753</td>
<td>2.131</td>
<td>2.602</td>
<td>2.947</td>
</tr>
<tr>
<td>16</td>
<td>0.690</td>
<td>1.337</td>
<td>1.746</td>
<td>2.120</td>
<td>2.583</td>
<td>2.921</td>
</tr>
<tr>
<td>17</td>
<td>0.689</td>
<td>1.333</td>
<td>1.740</td>
<td>2.110</td>
<td>2.567</td>
<td>2.898</td>
</tr>
<tr>
<td>18</td>
<td>0.688</td>
<td>1.330</td>
<td>1.734</td>
<td>2.101</td>
<td>2.552</td>
<td>2.878</td>
</tr>
<tr>
<td>19</td>
<td>0.688</td>
<td>1.328</td>
<td>1.729</td>
<td>2.093</td>
<td>2.539</td>
<td>2.861</td>
</tr>
<tr>
<td>20</td>
<td>0.687</td>
<td>1.325</td>
<td>1.725</td>
<td>2.086</td>
<td>2.528</td>
<td>2.845</td>
</tr>
<tr>
<td>21</td>
<td>0.686</td>
<td>1.323</td>
<td>1.721</td>
<td>2.080</td>
<td>2.518</td>
<td>2.831</td>
</tr>
<tr>
<td>22</td>
<td>0.686</td>
<td>1.321</td>
<td>1.717</td>
<td>2.074</td>
<td>2.508</td>
<td>2.819</td>
</tr>
<tr>
<td>23</td>
<td>0.685</td>
<td>1.319</td>
<td>1.714</td>
<td>2.069</td>
<td>2.500</td>
<td>2.807</td>
</tr>
<tr>
<td>24</td>
<td>0.685</td>
<td>1.318</td>
<td>1.711</td>
<td>2.064</td>
<td>2.492</td>
<td>2.797</td>
</tr>
<tr>
<td>25</td>
<td>0.684</td>
<td>1.316</td>
<td>1.708</td>
<td>2.060</td>
<td>2.485</td>
<td>2.787</td>
</tr>
<tr>
<td>26</td>
<td>0.684</td>
<td>1.315</td>
<td>1.706</td>
<td>2.056</td>
<td>2.479</td>
<td>2.779</td>
</tr>
<tr>
<td>27</td>
<td>0.684</td>
<td>1.314</td>
<td>1.703</td>
<td>2.052</td>
<td>2.473</td>
<td>2.771</td>
</tr>
<tr>
<td>28</td>
<td>0.683</td>
<td>1.313</td>
<td>1.701</td>
<td>2.048</td>
<td>2.467</td>
<td>2.763</td>
</tr>
<tr>
<td>29</td>
<td>0.683</td>
<td>1.311</td>
<td>1.699</td>
<td>2.045</td>
<td>2.462</td>
<td>2.756</td>
</tr>
<tr>
<td>30</td>
<td>0.683</td>
<td>1.310</td>
<td>1.697</td>
<td>2.042</td>
<td>2.457</td>
<td>2.750</td>
</tr>
<tr>
<td>40</td>
<td>0.681</td>
<td>1.303</td>
<td>1.584</td>
<td>2.021</td>
<td>2.423</td>
<td>2.704</td>
</tr>
<tr>
<td>60</td>
<td>0.679</td>
<td>1.296</td>
<td>1.671</td>
<td>2.000</td>
<td>2.390</td>
<td>2.660</td>
</tr>
<tr>
<td>120</td>
<td>0.677</td>
<td>1.289</td>
<td>1.658</td>
<td>1.980</td>
<td>2.358</td>
<td>2.617</td>
</tr>
</tbody>
</table>

\( N(0, 1) \) \( \infty \) | 0.674  | 1.282 | 1.645 | 1.960  | 2.326  | 2.576  |
<table>
<thead>
<tr>
<th>$m$</th>
<th>0.995</th>
<th>0.975</th>
<th>0.950</th>
<th>0.900</th>
<th>0.800</th>
<th>0.700</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.043927</td>
<td>0.039821</td>
<td>3.84146</td>
<td>5.02389</td>
<td>6.63490</td>
<td>7.87944</td>
</tr>
<tr>
<td>2</td>
<td>0.010025</td>
<td>0.0050636</td>
<td>5.99147</td>
<td>7.37776</td>
<td>9.21034</td>
<td>10.5966</td>
</tr>
<tr>
<td>3</td>
<td>0.0071721</td>
<td>0.0215795</td>
<td>7.81473</td>
<td>9.34840</td>
<td>11.3449</td>
<td>12.8381</td>
</tr>
<tr>
<td>4</td>
<td>0.0206990</td>
<td>0.0484419</td>
<td>9.48773</td>
<td>11.1433</td>
<td>13.2767</td>
<td>14.8602</td>
</tr>
<tr>
<td>5</td>
<td>0.411740</td>
<td>0.831211</td>
<td>11.0705</td>
<td>12.8325</td>
<td>15.0863</td>
<td>16.7496</td>
</tr>
<tr>
<td>6</td>
<td>0.675727</td>
<td>1.237347</td>
<td>12.5916</td>
<td>14.4494</td>
<td>16.8119</td>
<td>18.5476</td>
</tr>
<tr>
<td>7</td>
<td>0.898265</td>
<td>1.68987</td>
<td>14.0671</td>
<td>16.0128</td>
<td>18.4753</td>
<td>20.2777</td>
</tr>
<tr>
<td>8</td>
<td>1.344419</td>
<td>2.17973</td>
<td>15.5073</td>
<td>17.5346</td>
<td>20.0902</td>
<td>21.9550</td>
</tr>
<tr>
<td>10</td>
<td>2.15585</td>
<td>3.24697</td>
<td>18.3070</td>
<td>20.4831</td>
<td>23.2993</td>
<td>25.1882</td>
</tr>
<tr>
<td>13</td>
<td>3.56503</td>
<td>5.00874</td>
<td>22.3621</td>
<td>24.7356</td>
<td>27.6883</td>
<td>29.8194</td>
</tr>
<tr>
<td>14</td>
<td>4.07468</td>
<td>5.62872</td>
<td>23.6848</td>
<td>26.1190</td>
<td>29.1413</td>
<td>31.3193</td>
</tr>
<tr>
<td>15</td>
<td>4.60094</td>
<td>6.26214</td>
<td>24.9958</td>
<td>27.4884</td>
<td>30.5779</td>
<td>32.8013</td>
</tr>
<tr>
<td>16</td>
<td>5.14224</td>
<td>6.90766</td>
<td>26.2962</td>
<td>28.8454</td>
<td>31.9999</td>
<td>34.2672</td>
</tr>
<tr>
<td>17</td>
<td>5.69724</td>
<td>7.56418</td>
<td>27.5871</td>
<td>30.1910</td>
<td>33.4087</td>
<td>35.7185</td>
</tr>
<tr>
<td>18</td>
<td>6.26481</td>
<td>8.23070</td>
<td>28.8693</td>
<td>31.5264</td>
<td>34.8053</td>
<td>37.1564</td>
</tr>
<tr>
<td>19</td>
<td>6.84398</td>
<td>8.90655</td>
<td>30.1435</td>
<td>32.8523</td>
<td>36.1908</td>
<td>38.5822</td>
</tr>
<tr>
<td>20</td>
<td>7.43386</td>
<td>9.59083</td>
<td>31.4104</td>
<td>34.1696</td>
<td>37.5662</td>
<td>39.9968</td>
</tr>
<tr>
<td>21</td>
<td>8.03366</td>
<td>10.28293</td>
<td>32.6705</td>
<td>35.4789</td>
<td>38.9321</td>
<td>41.4010</td>
</tr>
<tr>
<td>22</td>
<td>8.64272</td>
<td>10.9823</td>
<td>33.9244</td>
<td>36.7807</td>
<td>40.2894</td>
<td>42.7956</td>
</tr>
<tr>
<td>23</td>
<td>9.26042</td>
<td>11.6885</td>
<td>35.1725</td>
<td>38.0757</td>
<td>41.6384</td>
<td>44.1813</td>
</tr>
<tr>
<td>24</td>
<td>9.88623</td>
<td>12.4011</td>
<td>36.4151</td>
<td>39.3641</td>
<td>42.9798</td>
<td>45.5585</td>
</tr>
<tr>
<td>25</td>
<td>10.5197</td>
<td>13.1197</td>
<td>37.6525</td>
<td>40.6465</td>
<td>44.3141</td>
<td>46.9278</td>
</tr>
<tr>
<td>26</td>
<td>11.1603</td>
<td>13.8439</td>
<td>38.8852</td>
<td>41.9232</td>
<td>45.6417</td>
<td>48.2899</td>
</tr>
<tr>
<td>27</td>
<td>11.8076</td>
<td>14.5733</td>
<td>40.1133</td>
<td>43.1944</td>
<td>46.9630</td>
<td>49.6449</td>
</tr>
<tr>
<td>28</td>
<td>12.4613</td>
<td>15.3079</td>
<td>41.3372</td>
<td>44.4607</td>
<td>48.2782</td>
<td>50.9933</td>
</tr>
<tr>
<td>29</td>
<td>13.1211</td>
<td>16.0471</td>
<td>42.5569</td>
<td>45.7222</td>
<td>49.5879</td>
<td>52.3356</td>
</tr>
<tr>
<td>30</td>
<td>13.7867</td>
<td>16.7908</td>
<td>43.7729</td>
<td>46.9792</td>
<td>50.8922</td>
<td>53.6720</td>
</tr>
<tr>
<td>40</td>
<td>20.7065</td>
<td>24.4331</td>
<td>55.7585</td>
<td>59.3417</td>
<td>63.6907</td>
<td>66.7659</td>
</tr>
<tr>
<td>50</td>
<td>27.9907</td>
<td>32.3574</td>
<td>67.5048</td>
<td>71.4202</td>
<td>76.1539</td>
<td>79.4900</td>
</tr>
<tr>
<td>60</td>
<td>35.5346</td>
<td>40.4817</td>
<td>79.0819</td>
<td>82.3976</td>
<td>88.3794</td>
<td>91.9517</td>
</tr>
<tr>
<td>70</td>
<td>43.2752</td>
<td>48.7576</td>
<td>90.5312</td>
<td>95.0231</td>
<td>100.425</td>
<td>104.215</td>
</tr>
<tr>
<td>80</td>
<td>51.1720</td>
<td>57.1532</td>
<td>101.879</td>
<td>106.629</td>
<td>112.329</td>
<td>116.321</td>
</tr>
<tr>
<td>90</td>
<td>59.1963</td>
<td>65.6466</td>
<td>113.145</td>
<td>118.136</td>
<td>124.116</td>
<td>128.299</td>
</tr>
<tr>
<td>100</td>
<td>67.3276</td>
<td>74.2219</td>
<td>124.342</td>
<td>129.561</td>
<td>135.807</td>
<td>140.169</td>
</tr>
</tbody>
</table>

Source: "Tables of the Percentage Points of the $\chi^2$-Distribution," *Biometrika*. Vol. 32 (1941), pp. 188–189, by Catherine M. Thompson; reproduced by permission of Professor E. S. Pearson and the Biometrika Trustees.

### Table 4

The $F(m, n)$ distribution: right-tail percentage points $F_{1-n}; \alpha = P(F \geq F_{1-n})$ (continued)

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$m$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>12</th>
<th>15</th>
<th>20</th>
<th>30</th>
<th>60</th>
<th>120</th>
<th>$\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>20</td>
<td>2.97</td>
<td>2.59</td>
<td>2.38</td>
<td>2.25</td>
<td>2.18</td>
<td>2.09</td>
<td>2.04</td>
<td>2.00</td>
<td>1.96</td>
<td>1.94</td>
<td>1.89</td>
<td>1.84</td>
<td>1.79</td>
<td>1.74</td>
<td>1.68</td>
<td>1.64</td>
<td>1.61</td>
</tr>
<tr>
<td>0.05</td>
<td>30</td>
<td>2.86</td>
<td>2.49</td>
<td>2.28</td>
<td>2.14</td>
<td>2.05</td>
<td>1.98</td>
<td>1.93</td>
<td>1.88</td>
<td>1.85</td>
<td>1.82</td>
<td>1.77</td>
<td>1.72</td>
<td>1.67</td>
<td>1.61</td>
<td>1.54</td>
<td>1.50</td>
<td>1.46</td>
</tr>
<tr>
<td>0.025</td>
<td>40</td>
<td>2.79</td>
<td>2.39</td>
<td>2.18</td>
<td>2.04</td>
<td>1.95</td>
<td>1.87</td>
<td>1.82</td>
<td>1.77</td>
<td>1.74</td>
<td>1.71</td>
<td>1.66</td>
<td>1.60</td>
<td>1.54</td>
<td>1.48</td>
<td>1.40</td>
<td>1.35</td>
<td>1.29</td>
</tr>
<tr>
<td>0.01</td>
<td>50</td>
<td>2.72</td>
<td>2.35</td>
<td>2.13</td>
<td>1.99</td>
<td>1.90</td>
<td>1.82</td>
<td>1.77</td>
<td>1.72</td>
<td>1.68</td>
<td>1.65</td>
<td>1.60</td>
<td>1.54</td>
<td>1.48</td>
<td>1.41</td>
<td>1.32</td>
<td>1.26</td>
<td>1.19</td>
</tr>
<tr>
<td>0.005</td>
<td>60</td>
<td>2.66</td>
<td>2.28</td>
<td>2.06</td>
<td>1.92</td>
<td>1.83</td>
<td>1.75</td>
<td>1.70</td>
<td>1.65</td>
<td>1.61</td>
<td>1.56</td>
<td>1.50</td>
<td>1.44</td>
<td>1.38</td>
<td>1.31</td>
<td>1.23</td>
<td>1.17</td>
<td>1.11</td>
</tr>
</tbody>
</table>
CHAPTER 1: PROBABILITY

1.1 $\mathcal{B}' = \{\emptyset, H, T, E, H \vee T = E', H \vee E = T', T \vee E = H', \{U\}\};$ corresponding probabilities are $0, 0.53, 0.47, 0, 1, 0.53, 0.47, 1$. 1.2 $\emptyset$ (null event); $P, T, U, P \vee T, 0$ to 10 coin flip; $P \vee U, 0$ coin flip or more than 10, $T \vee U, at least one coin flip, \{U\}$. 1.5 In correspondence with Exercise 1.2: $\emptyset, P, S \cap P', U, S, P \vee U, P', \{U\}$. 1.8 Necessary and sufficient condition for the cardinality of $\mathcal{B}$ to be $2^k$ is that $k$ be the number of nonnull events among $A_1A_2'\ldots A_k\ldots A_k$. 2.1 $S = \{1, 2, 3, 4, 5, 6\}; \ A = \{2, 4, 6\}; \ B = \{3, 6\}; \ C = \{1, 2, 3\}; \ A \cap B = \{6\}; \ A \cap C = \{2\}; \ A' \cap B' = \{1, 2, 3, 4, 5\}; \ B = \{1, 6\}; \ {2, 5\}; (3, 4\}; \ (4, 3\}; \ (5, 2\}; \ (6, 1\}; \ \{1, 4\}; \ (2, 5\}; \ (3, 6\}; \ (4, 1\}; \ (5, 2\}; \ (6, 3\}; \ (1, 5\}; \ (2, 6\}; \ (5, 1\}; \ (6, 2\}; \ 2.4 S = \{1, 1\}; \ 1, 2\}; \ 1, 3\}; \ 1, 4\}; \ 1, 5\}; \ 1, 6\}; \ 2, 2\}; \ 2, 3\}; \ 2, 4\}; \ 2, 5\}; \ 2, 6\}; \ 3, 3\}; \ 3, 4\}; \ 3, 5\}; \ 3, 6\}; \ 4, 4\}; \ 4, 5\}; \ 4, 6\}; \ 5, 5\}; \ 5, 6\}; \ 6, 6\}; \ A = \{1, 5\}; \ 2, 4\}; \ (3, 3\}; \ (1, 4\}; \ (2, 5\}; \ (3, 6\}; \ (4, 4\}; \ (5, 5\}; \ (6, 6\}; \ \{\emptyset\}; 2.6 S = \{g, g\}; \ (g, b\}; \ (b, g\}; \ \{g, g\}; \ (g, b\}; \ (b, g\}; \ \{b, b\}; \ 2.8 S = \{(y_1, y_2); y_1 \neq y_2, y_1 = AS, KS, \ldots, 2C\}. 52 \times 51 = 26526, \ \{(y_1, y_2); y_1 \neq y_2, y_1 = AS, KS, \ldots, 2S\}. 13 \times 12 = 156. 2.10 S = \{(1, 2, 3, 4), (1, 2, 4, 3), \ldots, \text{giving all 24 rearrangements of (1, 2, 3, 4)}\}; A_1 = \{(1, 2, 3, 4), (1, 2, 4, 3), (1, 3, 2, 4), (1, 3, 4, 2), (1, 4, 2, 3), (1, 4, 3, 2)\}; E_2 = \{(1, 2, 3, 4), (1, 3, 2, 4), (2, 1, 3, 4)\}; E_3 = \emptyset, E_4 = \{(1, 2, 3, 4)\}; 2.13 \{(y_1, y_2); y_1 + y_2 \leq 10\}; \ \{(y_1, y_2); y_1 \leq 5, y_2 \leq 5\}; \ \{(y_1, y_2); \text{at least one } y_1 \leq 5\}; \ \{(y_1, y_2); y_1 \leq 5\}; \ R \times \{(y_2); y_2 \leq 5\}; \ \{(y_1, y_2); y_1 + y_2 \leq 5\}; \ \{(y_1, y_2); -5 \leq y_2 - y_1 \leq 5\}. 2.20 \text{Given } A_1, A_2, \ldots \in \mathcal{A}; \text{then } A_1, A_2, \ldots \in \mathcal{A}; \text{then } \bigcap A_i \in \mathcal{A}; \text{then } A_i \in \mathcal{A}; \text{using Problem 2.19}; \text{thus closed under } \sigma\text{-intersection.} 2.23 \mathcal{P}(\emptyset) \in \sigma\text{-algebra; it contains } \emptyset. 3.1 S_0 = \{0, 1\}^3; \ \overline{P}(A) = c(A \cap S_0)/8; 1/8, 3/8, 1/8. 3.3 \rho(0, 0) = q^2, \rho(1, 0) = p(0, 1) = pq, \rho(1, 1) = p^2, P(H \cup T) = p^2 + pq + p = 1. 3.4 \overline{P}(A) = c(A \cap S_0)/12; \overline{P}(E_2) = 6/12; \overline{P}(E_3) = 6/12; \overline{P}(E_4) = 0. 3.7 S = \{(y_1, y_2); y_1 
eq y_2, y_1 = 1, 2, 3, 4\}; \text{same as for Exercise 3.4}. 3.8 0.6, 0.2, 0.5. 3.10 0.2, 0.2; 0.04; \pi/100. 3.11 A = A_1 \cup A_2 \cup \ldots; \text{then } \mu(A) = \mu(A_1) + \Sigma_{\sigma} \mu(\emptyset); \text{then } \Sigma_{\sigma} \mu(\emptyset) = 0; \text{then } \mu(\emptyset) = 0; \text{note that } A \text{itself could be just } \emptyset. 3.12 \text{Let } A = A_1 \cup \ldots \cup A_k, \text{where the } A_i \text{ are disjoint}; \text{then } A' = A_1 \cup \ldots \cup A_k \cup \emptyset \cup \emptyset \ldots; \text{with disjoint components; then } \mu(A) = \mu(A_1) + \ldots + \mu(A_k) + \Sigma_{\sigma} \mu(\emptyset) = \Sigma_{\sigma} \mu(\emptyset), \text{using Problem 11}. 4.1 3/6 + 3/6 - 9/36 = 3/4. 4.2 1 - P(E_2) = 23/24. 4.4 E_0 = A_1A_2. E_1 = A_1A_2 \cup A_1A_3 \cup A_2A_3 \cup A_2. 48 \times 52/51 = 2, 48 \times 4/52 \times 51, 4 \times 3/52 \times 51. 4.6 1 - 1/521. 4.8 \Sigma_{i=1}^{n} 1/2^{i-1} = 1/7. 4.10 6/24; 1 - 6/24. 4.12 2/10, 2/10, 4/10 - 4/100. 4.13 P(A_1 \cup A_2 \cup A_3) = P(A_1 \cup A_2 \cup A_4) - P(A_1) - P(A_2) - P(A_3) and one further step. 4.17 U_{i=1}^{n} A_i \text{ is monotone decreasing to } \phi, \text{thus } \lim P(U_{i=1}^{n} A_i) = 0; \text{thus } P(U_{i=1}^{n} A_i) = P(A_i) + \ldots + P(A_k) + P(U_{k+1}^{n} A_i): \text{take the limit}; \text{P}(U_{i=1}^{n} A_i) = \Sigma_{i=1}^{n} P(A_i). 5.7 A_2 - A_1 = (A_1 \cup A_2) \Delta A_1; A_1 \Delta A_2 = (A_1 - A_2) \cup (A_2 - A_1). 5.9 \text{Let } A \text{ and } B \text{ be finite unions of disjoint left open–right closed intervals on } \mathbb{R}; \text{then } A \cup B \text{ and } A - B \text{ are also finite.
unions of left open-right closed intervals. \[ E(Y) = \sum_n \lambda e^{-\lambda n}/(y - 1)! = \lambda \sum_n \lambda^2 e^{-\lambda y}/y! = \lambda. \]

**CHAPTER 2: PROBABILITY ON THE LINE AND PLANE**

1.1 \( F(y) = 0. [y]/6, 1\) according as \( y < 0, 0 \leq y \leq 6, 6 < y; 5/6 - 3/6, 1 - 4/6. \]
1.3 \( 0.3679, 0.3679/1, 0.3679/2 = 0.1840, 0.1840/3 = 0.0613, 0.0613/4 = 0.0153, \ldots; 0.3679, 0.3679, 1 - 0.3679 - 0.3679. \]
1.4 \( e^{-2}\{1 + 2\}, 2e^{-2}. \]
1.5 \( 1 - 5e^{-2}. \]
1.8 \( F(y) = 0, [y]/N, 1\) according as \( y < 0, 0 \leq y \leq N, N < y; (N/2)/N, ([N/4] - [N/2])/N. \]
1.10 See Section 3.5A. 14/16, 174/246. 1.12 See Section 3.5C. 68/70, 55/70.
1.13 \( G_1(\pm \infty) = F_1(\pm \infty) = F_1(\pm \infty) = 1; \) other properties of a distribution function hold; \( G_1(y) = \sum_i \rho_i(t)/\sum \rho_i(t) \) and thus \( G_1 \) is a discrete distribution function.
1.16 \( y \in S_d \iff \rho(y) > 0 \iff y \in A_n \) for some \( n; \) thus \( S_d = \cup A_n \lim A_n; c(A_n) < n; \) thus \( S_d \)

is finite or countable.

2.1 Monotone with \( G(\infty) = 0, G(\infty) = 1; \) \( \rho(y) = 2y \) on \( (0, 1) \) and zero otherwise; \( 1/\sqrt{2}. \]
2.2 \( 0.866, 0.5. \quad 0.159; \quad 0.01e^{-1/2}/\sqrt{2}. \]
2.4 \( 0.325, 0.0116. \]
2.6 \( g(t) \geq 0, \int g(t) dt = 1; \) \( G(\infty) = 1, \) \( G^{-1} = \sqrt{2}. \]
2.8 \( g(t) \geq 0, \int g(t) dt = 1; \) \( G(\infty) = 1, \) \( G^{-1} = \sqrt{2}. \]
2.10 \( g(t) \geq 0, \int g(t) dt = 1; \) \( G(\infty) = 1, \) \( G^{-1} = \sqrt{2}. \]
2.14 \( G(x) = \exp \{-e\sqrt{x} \}; \quad -\ln 2. \]
2.16 \( \Gamma(m/2) \int g(x) dx = \int_0^\infty e^{-x/2}(x/2)^{m/2-1} dx/2 = \Gamma(m/2) \quad \) for \( m/2 > 0; \)
\( \Gamma(2/2) \leq 0 = \int_0^\infty e^{-x/2}(x/2)^{2/2-1} dx/2 = \int_0^\infty h(u) du \) with \( p = m/2 \) and \( t > 0. \]
2.21 \( \Gamma(1) = \int_0^\infty e^{-t} dt = 1; \quad \Gamma(p + 1) = \int_0^\infty e^{-t} p(t) dt = \Gamma(p). \]
2.22 \( \int_0^\infty |x|\rho(t, u) t = t. \quad \Gamma(p) \)(p) \]
\( (1/\sqrt{2}). \]
3.1 \( \{a, b, c\} \quad F(a,b,c) = \{x : (y) = 0, \quad F(y) = F(y_1) = F(y_2) = F(y_3). \]
3.4 \( c^2 \leq c < 1 \) and \( 1 \)
3.6 \( f(t) = 0, \int f(t) dt = 0. \]
3.10 \( f(t) = 0, \int f(t) dt = 0. \]
4.1 \( F_y(w) = w \) on \( (0, 1). \]
4.4 \( \rho_y(w) = 1/6 \) on \( 1, 4, 9, 16, 25, 36 \) and \( 0 \) otherwise.
4.5 \( \exp \{-\Lambda\} \) for \( \Lambda > 0 \) and \( 0 \) otherwise.
4.7 \( e^\alpha \quad \) for \( \alpha < 0 \) and \( 1 \) otherwise.
4.9 \( F_y(y) = 1 - (1 + y)^{-1} / y; \quad \) \( F_y(y) = 1 - e^{-y} \) for \( y > 0. \]
4.11 \( F_y(y) = 1 - G(\infty) \leq 1 \quad e^{-y} \) for \( y > 0. \]
4.24 \( a_1 = a_1, \quad a_2 = a_2; \quad a_2 = a_2; \quad a_1 = a_1 / a_1^2 (1 - a_1^2), \quad a_1 = a_1 / a_1^2 (1 - a_1^2). \]
5.1 \( \cos^{-1}(1/2), \quad \cos^{-1}(\sqrt{5}/\sqrt{6}), \quad \cos^{-1}(15/\sqrt{354}), \quad \cos^{-1}(100/\sqrt{30}/\sqrt{354}). \]
5.2 \( 41, \)
5.4 \( -1/2, -1/2, 1/2, 3/2. \]
5.6 \( x'Mx = n'^{-1}x'^{-1} \quad (\Sigma x)_y/n = \alpha^2 x. \]

**CHAPTER 3: MARGINAL PROBABILITY**

1.1 \( \rho(s) = 1/8 \) for \( s \in \{i_1, i_2, i_3\} \) with \( i_j = 0, 1. \]
1.2 \( P_y(y) = 1/8, 3/8, 3/8, 1/8 \)
for \( y = 0, 1, 2, 3 \). e.g., note that \( P_y(s) = P_y(y) = \{y = 0, 1, 0, 0, 0, 1\}; \)
1.3 \( F_y(y) = y^2/2, 1 - (y - 1)/2 \)
for \( 0 < y < 1, 1 < y < 2; \quad f_y(y) = y, 2 - y \) for \( 0 < y < 1, 1 < y < 2 \) and \( 0 \) otherwise.
1.4 \( \text{The line } y_1 + y_2 + c = 0; \quad \text{the circle } y_2^2 + y_3^2 = d; \quad \text{preimage of } (c, d) \)
\( \{y_2 \leq 2d - c^2/2, c/2 \leq y_2 \leq 2d - c^2/2\}. \]
1.6 \( y = c \) is a plane, \( y_1 + y_2 + y_3 = 3c \) perpendicular to the vector \( 1 = (1, 1, 1); \) \( s_2^2 = (c, d) \) is a cylinder of radius \( \sqrt{n} \) with axis the line \( L(1). \)
which gives a circle of radius $\sqrt{n - 1} \sqrt{d}$ centered at $(c, c, c)$ on the plane through $(c, c, c)$ perpendicular to 1. 1.8 $s \in Y^{-1}(U_{B_j}) \iff Y(s) \in U_{B_j} \iff Y(s) \in B_j$ for some $j \iff s \in Y^{-1}(B_j)$. 1.10 Consider a rotation of axes through angle $\pi/4$; $F_{w}(w) = G(w/\sqrt{2})$, where $G$ is the $M(0, 1)$ distribution function; $M(0, \sqrt{2})$.

2.1 $H_1, H_2$. 2.2 $0.10, 0.45, 0.80, 1.00; 0.20, 0.50, 0.80, 1.00$. 2.3 $F_1(y_1) = 1 - (1 + y_1)^{-1}, F_2(y_2) = 1 - (1 + y_2)^{-1}$ on positive axes; $f_1(y_1) = (1 + y_1)^{-2}, f_2(y_2) = (1 + y_2)^{-2}$ on positive axes. 2.6 $F_{Y_1}(y_1) = (G(y_1) + H(y_1))/2, F_{Y_2}(y_2) = 0, 1/2, 1$ on $y_2 \leq 0, 0 \leq y_2 < 1, 1 \leq y_2 \leq \infty$. 2.9 $G(w) = 0, w, 1$ for $w < 0, 0 \leq w < 1, 1 \leq w$. $H(r) = r^2, 1$ for $r < 0, 0 \leq r < 1, 1 \leq r$. 2.11 $F_{y_1}(y_1) = 1 - (1 + y_1)^{-a}, F_{y_2}(y_2) = 1 - (1 + y_2)^{-a}$ on positive axes. 2.12 $\rho(x_1, x_2) = 1/12, \rho(y_1) = 1/6$ otherwise. 2.14 $f_{y_1 < y_2} f_{y_1} f_{y_2} = f_{y_1 < y_2} f_{y_2} f_{y_1}$ and hence each is $1/2$. 2.17 $f_{y_1}(y_1) = 2e^{-2y_1}, f_{y_2}(y_2) = 2(e^{-y_2} - e^{-2y_2})$ on positive axes. 2.18 $f_{y_1}(y_1) = 2(1 - y_1^2)/\pi$ on $(-1, +1)$ and zero otherwise, $f_{y_1} = f_{y_2}$.

3.1 Independence implies factorization (Proposition 6). Factorization gives $F_X(x) = g(x)h(+\infty), F_Y(y) = g(+\infty)h(+\infty)$ so that $F_X(x)F_Y(y) = g(x)h(y)$, which is $F_{XY}(x, y)$. 3.2 Independence implies factorization (Proposition 7). Factorization gives $p_{XY}(x, y) = g(x)h(y), p_X(x) = g(x), p_Y(y) = h(y)$, so that $p_{XY}(x, y) = g(x)h(y)$, which is $p_{X,Y}(x, y)$. 3.5 Add chi-squares one at a time and correspondingly add degrees of freedom.

3.7 $F_{\rho(w)}(w) = 1, 3, 6, 10, 15, 21, 25, 27, 25, 21, 15, 10, 6, 3, 1$ at $w = 3, 4, \ldots$. 15.12 $|\partial(u_1, u_2)/\partial(x_1, x_2)| = \exp\{-(x_1^2 + x_2^2)/2\}/2\pi; f_{x_1, x_2} = g(u_1, u_2)|\partial(u_1, u_2)/\partial(x_1, x_2)|$.

4.4 $(1 + 1)^n = 2^n; (1 - 1)^n = 0$. 4.12 $1/6^5, 1/2^5, 5^5/6^5, 5^4/2^6, 5^3/2^6$. 4.14 $1/2^5, 10/2^6$.

4.16 $\binom{13}{5} 4^5 /\binom{52}{5}$. 4.20 $\binom{n}{y_1 \ldots y_k} k^n$. 4.5.1 $1/8, 3/8, 3/8, 1/8$ at 0, 1, 2, 3. 5.3 As for Exercise 5.1. 5.5 $1/27$ at $(3, 0, 0), (0, 3, 0), (0, 0, 3), 3/27$ at $(1, 2, 0), (0, 1, 0), (2, 0, 1), (1, 0, 2), (0, 2, 1), (0, 1, 2)$.

6/27 at $(1, 1, 1)$. 5.7 As for Exercise 5.5. 5.11 $\binom{5}{y_1} \binom{5}{y_2} \rho p_1(1 - p_1)^{y_2} \rho p_2(1 - p_2)^{y_1}$ for $y_1 = 0, 1, 2, 3, 4, 5$. 5.13 $20\rho(y) = 1, 9, 9, 1$ at 0, 1, 2, 3. 5.15 As for Exercise 5.13.

5.16 $\binom{4}{y_1} \binom{48}{13 - y_1} /\binom{52}{13}$ at $y = 0, 1, 2, 3, 4$. 5.18 $3/10, 3/10, 3/10, 1/10$ at $(1, 1, 1), (1, 0, 2), (0, 1, 2), (0, 0, 3)$. 5.20 $\binom{13}{x} \binom{13}{y} /\binom{26}{52}$. 5.22 $\binom{10}{y} /\binom{5}{5}$. 5.26 Note that $\binom{n}{y} /\binom{n}{y_1}$ is $(n - y + 1)/y$.

5.29 $\rho_{\ell}(t) = \rho((1 - \rho)^{t - y_1 - y_2 - 1} \Sigma_{y_1} / (t - y_1) /\binom{y_1}{t - y_1} /\binom{52}{26}$. 5.34 $\rho_{\ell}(y_1) = e^{-\ell - \ell_1 /\ell(t)}$.

5.37 $\rho_1(y_1) = (n - y_1)^{(n + 1)} /\binom{52}{13}$. 5.46 $\binom{10}{y_0 y_1 \ldots y_4} p^{y_0} (1 - p)^{y_1 - y_0} /\binom{52}{13}$. 5.49 $\Sigma_{y_1} 1 \Sigma_{y_1} = 10$, where $p_1 = \binom{48}{13 - 1} /\binom{13}{13}$

6.10 $|\partial(u_{a1}, \ldots, u_{an})/\partial(c_1, \ldots, c_n)| = 1$. 6.12 $|\partial(c_1, \ldots, c_n)/\partial(U_1, \ldots, U_k)| = 1; C_i = U_i - U_i - 1$.

CHAPTER 4: CONDITIONAL PROBABILITY

1.1 $1/3$. 1.2 $0$. 1.3 $1/2$. 1.4 $\binom{13}{x} \binom{39}{13 - x} /\binom{26 + x}{26}$. 1.7 Disjoint $A_i$, then $P(B \cup A_i) = P(B) P(A_i/C) = P(C)$. 5.34 $\rho_1(y_1) = e^{-\ell - \ell_1 /\ell(t)}$.

5.37 $\rho_1(y_1) = (n - y_1)^{(n + 1)} /\binom{52}{13}$. 5.46 $\binom{10}{y_0 y_1 \ldots y_4} p^{y_0} (1 - p)^{y_1 - y_0} /\binom{52}{13}$. 5.49 $\Sigma_{y_1} 1 \Sigma_{y_1} = 10$, where $p_1 = \binom{48}{13 - 1} /\binom{13}{13}$

6.10 $|\partial(u_{a1}, \ldots, u_{an})/\partial(c_1, \ldots, c_n)| = 1$. 6.12 $|\partial(c_1, \ldots, c_n)/\partial(U_1, \ldots, U_k)| = 1; C_i = U_i - U_i - 1$.

CHAPTER 4: CONDITIONAL PROBABILITY
\[ \sum P_\sigma(A_i); \quad P_\sigma(S) = P(S|C)/P(C) = 1. \]

1.9 \( 365^{th}/365^n; \) (1 \( \Sigma \delta^{-1} i/365 \) = \( 730 + n - n^2)/730 \) or, better, \( \exp(-\Sigma \delta^{-1} i/365) = \exp[-n(n-1)/730]; \) 23 by direct calculation.

2.1 (405, 1080, 1080, 480, 80)/3125 at (0, 1, 2, 3, 4), respectively; (243, 810, 1080, 720, 240, 32)/3125 at (0, 1, 2, 3, 4, 5), respectively. 2.3 \( 3/10, 6/10, 1/10 \) at 0, 1, 2, 1/10, 6/10, 3/10 at 0, 1, 2.

2.5 Number of sequences of three defects is \( 3^{th} \); of two good articles is \( 7^{th} \); number of ways of interfacing so that the defective is last is \( \binom{3}{2} \).

2.12 \( 1/2 \sqrt{(1-y_j^2)} \) for \( y_j \) in \( (\pm \sqrt{(1-y_j^2)}) \) and zero otherwise. 2.15 \( y_j/(1+y_j) \) is Pareto(\( \alpha + 1 \)). 2.17 Uniform on parallelogram \((0, 0), (1, 0), (2, 1), (1, 1)\); uniform \( (0, t) \) for \( 0 < t < 1 \) and uniform \((t - 1, 1) \) for \( 1 < t < 2 \). 2.21 \( \Gamma(r_{p+1} + \ldots + r_{k+p})/\Gamma(r_{p+1}) \ldots \Gamma(r_{k+p})/(Y_{p+1}/Y)^{r_{p+1} - 1} \ldots (Y_k/Y)^{r_{k+p} - 1} \) on \( 0 < Y_j, \Sigma_{k<p} Y_j < Y \).

2.25 Number of sequence of \( r \) defects \( D^n \), of \( n - r \) good is \((N - D)^{n-r} \); number of interfacings is \( \binom{n}{r} \).

2.29 \( 'r - 1 \) defects in first \( n - 1 \) and \( ' \)defective on nth.'

3.1 \( \rho(n) = q^{n-1}p \) with \( p = 1 - q = 2/9 \); see Example 1.4.4, 14/32, 98/386; \( r \) (win on \( nth \)) = \( q^n-1(1/18), \) \( r \) (lose on \( nth \)) = \( q^n-1(3/18) \).

3.3 \( \prod_i \frac{(1-e^{\lambda_i \theta \sigma})(1-e^{-\lambda_i \sigma})}{(1-e^{1/2})} = (\lambda_1 \sigma + \ldots + \lambda_i \sigma) \sum \frac{\lambda_1 \sigma + \ldots + \lambda_i \sigma}{\lambda_1 \sigma + \ldots + \lambda_i \sigma} \).

3.4 \( \prod_{i=1}^{n} \frac{(1-e^{-\lambda_i \theta \sigma})(1-e^{-\lambda_i \sigma})}{(1-e^{1/2})} = (\lambda_1 \sigma + \ldots + \lambda_i \sigma) \sum \frac{\lambda_1 \sigma + \ldots + \lambda_i \sigma}{\lambda_1 \sigma + \ldots + \lambda_i \sigma} \).

3.6 \( p_2(1, j) = 0.64, 0.21, 0.15, p_2(2, j) = 0.67, 0.33, p_3(3, j) = 0, 0.66, 0.34; p_4(4, j) = 0.32, 0.43, 0.25, p_5(5, j) = 0.256, 0.483, 0.261.

3.10 \( e^{-2}; \) \( \sigma^{2}; \) \( F_{\infty}(w) = e^{2w} \) for \( w < 0 \) and 1 otherwise.

CHAPTER 5: THE MEAN VALUE FOR REAL AND VECTOR DISTRIBUTIONS

1.1 1/4, 0.3 3/4, 3/5. 1.5 \( \infty \) is a constant.

2.1 2. 2.3 2. 2.5 2. 2.7 2. 2.9 2.1 2.11 \( E(y) = \sum_i \rho_\sigma(i) = \sum_i \rho_\sigma(i) + \ldots \).

2.13 \( \int f(x, y) dP(y) \geq \int f(x, k) dP(y) \).

2.15 \( \mathbb{E} \sum x_i \) is continuous in \( t; \mathbb{E} (\sum x_i - t \delta) = \mathbb{E} (\sum x_i - t \delta) \) provided there is no \( a_i \) in interval \((\pm \delta)\).

3.1 1. 3.3 4/7. 3.6 1/25. 3.7 1/20(5, -2, -3, 2, -8, -6, -3, 6, 9, 1/20(-2, -3).

3.29 \( h(t) = E((x - t)^2), h'(t) = 2t - 2 \mu, h''(t) = 2 \). 3.6 50,000.

3.32 \( \mathbb{E} \sum t \) is a variance, \( \geq 0 \); hence \( \mathbb{E} \) positive semidefinite.

3.36 \( E(x_k, x_l) = 0 \) for \( i \neq j; E(x^2) = E(x^2)/k = k/k.

4.1 \( 1/n\) is a constant. \( \Sigma (x_i - \bar{x}) = \Sigma x_i - \overline{x}^2 \).

4.3 16.8, 2.6752. 16.3485. 4.3 9.3, 3.25. 4.7 0.4947. 4.7 10.2, 0.4946. 4.7 \( \Sigma y_i = na + C \Sigma x_i \).

4.10 \( 1/n(n + 1), n(n + 1) \).

4.12 \( 1, 2, \ldots, n(n + 1) \).

4.13 \( u(n) = (n - r - s + 1)/(n + 1)^2 \).

5.1 \( 2 \) \((1 - y)^2; 1 - y)^2/18, 2.5 1/4 = E(1 - x)/2 \).

5.5 \( u(n) = (1 - u_0)(n - r + 1) \).

5.7 \( I = \Sigma \Sigma^{21} \Sigma_{12} = \Sigma \Sigma^{21} \Sigma_{12} = \Sigma \Sigma^{21} \Sigma_{12} \).

6.1 2/(r + 2).

6.3 \( \mu_r = \mu_r = p \rho + (-p)^r \).

6.6 \( (N + 1)^{(n+1)/N} \).

6.7 \( \mu_r \) is the sample average for \( \rho = x^2 \).

7.1 \( e^{1/4} + e^{2/4} + e^{3/4}/3, \) \((1 + 2 + 3)/3, \) \((1 + 4 + 9)/3. \) 7.2 \( 1 + 4e^t + 6e^2 + 4e^3 + e^4)/16 = (1 + e^t)^4/16. \) 7.3 \( f(x) e^{xt} dx = (e^t - 1)/t \).

7.7 \( \int e^{x^2} dx = \int e^{x^2} dx = \int (1 - 2t)^{-1/2} \).

7.33 \( 1 + 2t)^{-1/2} = (1 - 2t)^{-1/2}. \)

7.35 For \( \Sigma x_i, \) \( m(t) = \Pi \exp \{\mu t + n\alpha^2/2\}; \) for \( \bar{x}, \) \( m(t) = \exp \{\mu t + n\alpha^2/2\}. \)

7.36 For \( \Sigma x_i, \) \( m(t) = \Pi \exp \{\mu t + n\alpha^2/2\}; \) for \( \bar{x}, \) \( m(t) = \exp \{\mu t + n\alpha^2/2\}. \)
\[ m(t) = \prod \exp \{ \mu' t + t' \Sigma t / 2 \} ; \text{for } i, m(t) = \exp \{ n \mu' t / n + n' \Sigma t / 2 n' \} \]. 7.38 Moment-generating function for \( x \) is \( m_2(a) = m_{x^2}(1) \), which is determined by the distribution of \( a'x \). 7.41 \[ \mu_1 t_1 + \mu_2 t_2 + \{ \sigma_1^2 t_1^2 + 2 \sigma_1 \sigma_2 t_1 t_2 + \sigma_2^2 t_2^2 / 2 \} \).

CHAPTER 6: LIMITING DISTRIBUTIONS AND LIMITING FUNCTIONS

1.10 0.706. 1.2 \( \ln m(t) = -\sqrt{m/2} t - (m/2) \ln (1 - t/\sqrt{m/2}) = -\sqrt{m/2} t - (m/2)[1 - t / \sqrt{m/2} - (1/2)t^2 / (m/2) \ldots] \rightarrow t^2 / 2 \).

1.4 \( P(|y_n - c| \leq c) \rightarrow 1 \) implies that \( P(y_n < c - \epsilon) \rightarrow 0 \). \( P(y_n \leq c + \epsilon) \rightarrow 1 \). 1.7 Density for \( n \) is \( 1 - u_n^{n-1} \) for \( n u_n = y \) is \( (1 - y/n)^{n-1} \rightarrow e^{-y} \) for \( y \) on \((0, \infty)\). 1.9 The uniform distribution for \( u \) spreads the discrete probability \( \lambda e^{-\lambda y} / y! \) at \( y \) uniformly over \((y \pm 1/2)\) with density as given; \( (w - \lambda) / \lambda^{1/2} = (y - \lambda)/\lambda^{1/2} + u / \lambda^{1/2} \text{ final term } \rightarrow 0 \), and thus same limiting distribution by Theorem 4.

2.1 9604. 2.3 \( m_n \) is sample average for \( y^* \). 2.15 \( y_i - \mu \) has mean 0. Weak Law for \( y_i - \mu \) says that \( y_n \rightarrow \mu \).

CHAPTER 7: STATISTICAL INFERENCE

1.10 0.1392. \( (0.1392, 0.1392, 0.1392, 0.1392, 0.1392) \), \( (-0.0012, 0.0008, -0.0032, 0.0028, 0.0008) \), \( 0.000052 \). 1.2 21, \( (21, 21, 21, 21, 21, 21) \). \( (-2.7, -1.7, 0.2, 1.4, -3.9, 6.7) \), 14.456. \( (1.3, 90.5857, 90.5857, 1.1, 1.1, 1.1, 1.1) \). \( (-2.8857, -3.3857, 1.7143, 3.6143, 3.0143, -4.9857, 2.9143) \), 13.0381.

2.1 \( z = 5.3/0.775 = 6.839 \), a very extreme value for the standard normal; subject to the assumptions, this is strong evidence against the hypothesis. \( (285.3 \pm 1.52) \). \( (285.3 \pm 2.00) \).

2.2 \( z = 4.696 \), an extreme value for the standard normal; subject to the assumptions, this is strong evidence against the hypothesis. \( (23 \pm 2.00) \). \( (23 \pm 2.00) \).

2.4 \( t = 0.8571/1.3648 = 0.429 \), a reasonable value for the standard normal, no evidence against the hypothesis? \( (90.59 \pm 2.73) \).

2.5 \( t = -4.077, (3.921 \pm 0.333), (3.921 \pm 0.500) \). 2.6 \( z = 2.920 \), a value beyond the 1% point of the standard normal; subject to the assumptions, this is moderately strong evidence against the hypothesis. \( (2.0 \pm 2.15) \).

2.7 \( t = 0.2573/0.2627 = 0.979 \), a reasonable value for the standard normal, no evidence against the hypothesis. \( (0.257 \pm 0.79) \).

2.8 As sample sizes are equal, test is same as in preceding answer. \( (0.257 \pm 0.53) \).

2.9 \( (\bar{y}_2 - \bar{y}_1 \pm 3(s_1^2/m + s_2^2/n)^{1/2}) \). 2.10 \( (\bar{y}_2 - \bar{y}_1 \pm 2(1/m + 1/n)^{1/2}) (m - 1)^{1/2} / (m + n - 2)^{1/2} \).

2.11 Without loss of generality, take \( \mu = 0 \); use \( (n - 1)^{1/2} \rho = (\Sigma y_2^2 - \Sigma y_1^2)^{1/2} \) and take the mean value.

3.1 \( (1+ 52 \pm 94.67) \). \( (3.3 871.5, (871.5 \pm 220.5) \). \( (3.5 1325 \pm 306) \).

4.1 \( K = 5.6.25 \%, (0.136, 0.142) \). 4.2 \( K = 1, 21.9 \%, (17.1, 27.7) \). 4.3 \( T = 0.021 \, 6.25 \%, (0.136, 0.142) \). 4.4 \( T = -36.6.25 \%, (17.1, 27.7) \). 4.7 \( k = 3.5, \rho^2 + (1 - \rho)^2 \).

4.8 \( y_2 = \text{(number of } y_2 < y_1 + 1 = \Sigma y_i - y_i - 1 \). 4.10 \( (4.6; 2048) \text{(probability)} = 1, 1, 1, 2, 2, 3, 4 \) at 0. 1. 2. 3, 4, 5, 6; observed level of significance is 28/2048 = 1.36%; some evidence against hypothesis. \( \{6 - 33\}/11.25 = -2.4 \); observed level of significance is 1.64%. 4.13 \( (3.385, 4.61) \).

4.16 \( S = 4 \). \( (4 - 5.5)/1.20 = -1.25 \); no evidence against hypothesis. 4.19 \( U = 75 \) or \( 76, z = 0.95 \); no evidence against hypothesis. 4.23 Observed \( \Sigma y_2 = 86.0 \); 20 possible values \( \Sigma y_2 \) are 86.1, 86.0, . . . , 80.8, 80.7; observed level of significance is 20%; no evidence against hypothesis.

5.1 \( \{463.5 \pm 18.83\} \). Sample number 19 above UCL, sample number 14 below LCL, process not in control. Nine samples outside control limits based on specifications; does not conform to specifications.

CHAPTER 8: THE LIKELIHOOD FUNCTION IN STATISTICAL INFERENCE

1.2 \( c \exp (-6.3/\theta) / \theta^3 \). \(-6.3 \theta^{-1} - 3 \ln \theta + a \). \( 6.3 \theta^{-2} - 3 \theta^{-1} \). \( \tilde{\theta} = 2.1 \). 1.4 \( c \exp (-9.4 + 3\theta), \exp (-5.2 + \theta), \exp (1.6 - \theta), \exp (9.4 - 3\theta) \) on \( (-\infty, 2.1) \).

3)/4. 1.5 \min (\Sigma y_i - \lambda)^2 \text{ gives } \bar{y}; \lambda = \bar{y} \text{ gives } \bar{y}; \Sigma y_i/\lambda - n = 0 \text{ gives } \bar{y}. 1.9 (\bar{\beta}x)’(\bar{\beta}x) = 
abla' \bar{\beta} x \cdot \bar{\beta} = \Sigma x_i \hat{\beta}. 1.10 x \text{ is orthogonal to the residual } y - \bar{\beta} x; (y - \bar{\beta} x)'(y - \bar{\beta} x) = y' (y - 2.1 \begin{pmatrix} 4 \quad 5 \quad 6 \quad 7 \quad 8 \end{pmatrix}, 190; (-1, 2, 0, -2, 1), 10; 2, (4, 5, 6, 7, 8). 190 / 3, (-1, 2, 0, -2, 1), 10 / 5, (3, 7, 6, 5, 9), 200. 2.2 (1, 3, 5, 7, 9), 165; (1, 0, -2, 0, 1), 6; the same table as in Section C. 2.4

\[
P_1 = \frac{1}{10} \begin{pmatrix} 1 & 6 & 4 & 2 & 0 & -2 & 0 & 1 \ 4 & 3 & 2 & 1 & 0 \ 2 & 2 & 2 & 2 \ 0 & 1 & 2 & 3 & 4 \ -2 & 0 & 2 & 4 & 6 \ \end{pmatrix} \quad P_2 = \frac{1}{10} \begin{pmatrix} 4 & -4 & -2 & 0 & 2 \ -4 & 7 & -2 & -1 & 0 \ -2 & -2 & 8 & -2 & -2 \ 0 & -1 & -2 & 7 & -4 \ 2 & 0 & -2 & 4 & -4 \ \end{pmatrix} .
\]

2.7 (\bar{y}_1, 2, \bar{y}_p, \ldots \bar{y}_Q, \bar{y}_r, \ldots \bar{y}_r, \ldots \bar{y}_p, n \bar{y}_i^2 + n \bar{y}_i^2 + n \bar{y}_i^2 + n \bar{y}_i^2 + \bar{y}_1, \ldots \bar{y}_m - \bar{y}_1, \bar{y}_2 - \bar{y}_2, \ldots \bar{y}_n - \bar{y}_2, \bar{y}_3 - \bar{y}_3, \ldots \bar{y}_p - \bar{y}_p, \Sigma (y_i - \bar{y}_i)^2. 2.9 (Xb_2 - Xb_1)' (Xb_2 - Xb_1) = (b_2 - b_1)' X' X (b_2 - b_1). 2.10 (X'X) = (n, 0 / 0, \Sigma (x_i - \bar{x})^2). \bar{x}' = (\Sigma y_i, \Sigma x_i (x_i - \bar{x}))'; b = (x' x)' x' y = expression as given; y'y - y'xb = \Sigma y_f - \Sigma y_i \bar{y} = (x_i - \bar{x}) \Sigma y_i (x_i - \bar{x}) / \Sigma (x_i - \bar{x})^2.

3.1 257.4359. 3.3 (t + w)/2; \text{ } \quad \text{At } + Bu \quad \text{with } A = (0.354318, 0.306698 / 0.306698, 0.354318), \quad B = (0.645682, -0.306698 / -0.306698, 0.645682). 3.4 b = (x' x)' x' y, (x' x)' x' y. 3.12 Theorem 7 with the following notations:

\[
E(t_1) = \frac{1}{\beta} \quad \text{M} = \begin{pmatrix} \Sigma y_i (\theta_0) & 0 \\ 0 & \Sigma y_i (\theta_0) \end{pmatrix}.
\]

4.1 \(\tau = \rho_0 + (n/\rho_0 q_0)^{-1} (\Sigma x_i/\rho_0 - (n - \Sigma x_i)/q_0) = \bar{x}; \quad \text{yes.} \quad 4.3 \alpha_0^2 + (n/2\alpha_0^2)^{-1} \quad -n/2\alpha_0^2 + \Sigma (y_i - \mu_0)^2/2\alpha_0^2 = \Sigma (y_i - \mu_0)^2/n; \quad \text{yes.} \quad 4.7 S(y|\theta) = \phi'(\theta) + \psi'(\theta) y; \quad E(S) = 0 \text{ gives } E(t) = -\psi'(\theta) + \psi'(\theta) y; \quad \text{an affine function of } S(y|\theta) \text{ that is independent of } \theta \text{ must be an affine function of } t; \quad \text{thus the corresponding parameter is an affine function of } \phi'(\theta) + \psi'(\theta).
\]

4.9 (1/\theta + 1/\theta + 1). 4.11 \(J_1(\theta_0) + J_2(\theta_0)^{-1}(J_1(\theta_0) (\theta_0 + J_2^{-1}(\theta_0) S(\Sigma | \theta_0) + J_2(\theta_0) (\theta_0 + J_2^{-1}(\theta_0) S(\Sigma | \theta_0)) = \theta_0 + (J_1(\theta_0) + J_2(\theta_0)^{-1}(S(\Sigma | \theta_0)) + S(\Sigma | \theta_0)).
\]

4.13 \(t = \left(\begin{array}{c}
\mu_0 \\
\alpha_0^2/n
\end{array}\right) / \left(\begin{array}{c}
\sigma^2/\mu_0 \\
2a_0^2/n
\end{array}\right) \quad \bar{y} = \left(\begin{array}{c}
\Sigma y_i - \mu_0/a_0^2 \\
\Sigma y_i - \mu_0/a_0^2 + \Sigma (y_i - \mu_0)^2/2\alpha_0^2
\end{array}\right).\]

4.14 thus no UMV unbiased estimate at information lower bound; however, later we will determine a UMV unbiased estimate.

5.1 Distribution of \(s = \Sigma y_i \text{ has density } \Gamma^{-1}(n) \theta^{-n} \exp \{-s/\theta\}; \text{ has exponential form; thus } s \text{ is complete. } E(s/n) = E(\bar{y}) = \theta; \text{ thus } \bar{y}, \text{ a function of } s \text{ is UMV unbiased. 5.4 Distribution of } y \text{ is binomial}(n, p) \text{ with } p \in [0, 1]. E(r(y)) = r(0) \begin{pmatrix} n \end{pmatrix}_0 q^n + r(1) \begin{pmatrix} n \end{pmatrix}_1 p q^{-n+1} + \cdots = 0 \text{ for all } p; \text{ with } p = 0, \text{ obtain } r(0) = 0; \text{ then divide out } p \text{ and consider the limit as } p \rightarrow 0, \text{ obtain } r(1) = 0; \text{ continue. 5.5 } y_{2 \hat{a}} - y_{1 \hat{a}} = -1/3. 5.9 E(r(y)) = \Sigma y_i \begin{pmatrix} D \end{pmatrix} (N - D) / (N - n) \text{ with } y \in (\max(0, n - N + D), \min(n, D)); \text{ take } D = 0 \text{ and obtain } r(0) = 0; \text{ then } D = 1 \text{ and obtain } r(1) = 0; \text{ continue. 5.13 Density for } (s_1, s_2) = (u_{1\theta}, u_{2\theta}) \text{ is } n(n - 1)(s_2 - s_1)^{-2} / \theta_1 (s_2 - s_1) \text{ on } (\theta_1, \theta_2) \text{ and zero otherwise; } E(r(s_1, s_2)) = 0 \text{ implies that } \int_0^1 \int_0^1 r(s_1, s_2) (s_2 - s_1)^{-2} ds_1 \text{ ds_2 = 0; } \theta_1 < \theta_2; \text{ differentiate re } \theta_2, \text{ then re } \theta_1, \text{ and obtain } r(\theta_1, \theta_2)(\theta_2 - \theta_1)^{-1} = 0 \text{ for all } \theta_1 < \theta_2. 5.17 \text{ Follow Example 4: } \int r^4(s) \exp(s' \theta) h(s) \text{ ds = } \int r^4(s) \exp(s' \theta) h(s) \text{ ds; then continue as Example 4, using uniqueness for multivariate moment-generating functions. 5.21 From Problem}
Answers

9.2 we have \( b = \sum (i - n/2)y_i/\Sigma (i - n/2)^2 \). 5.23 Let \( (s_1, \ldots, s_n) = (\Sigma y_i, \ldots, \Sigma y_n^2) \); then by Problem 17 \( (s_1, \ldots, s_n) \) is complete. But the order statistic is equivalent to \( (s_1, \ldots, s_n) \).

5.25 \( \text{var}(\sigma_1 + (1 - a)t_2) = a^2\sigma_2^2(\theta) + 2a(1 - a)\text{cov}(t_1, t_2) + (1 - a)^2\sigma_2^2(\theta) \geq a^2(\theta) \), which implies that \( \text{cov}(t_1, t_2) = a^2(\theta) \geq 0 \), which by covariance inequality implies that \( t_1 \) and \( t_2 \) are affinely related and hence equal.

6.1 \( N(108, 4/\sqrt{5}) \); 108. 6.3 All probability for Bb. 6.5 \( N(\mu y/\sigma_\theta^2 + \mu_\theta/\sigma^2 \omega^2, \omega) \), where \( 1/\omega^2 = n/\sigma_\theta^2 + 1/\omega^2 \). 6.7 \( \text{beta}(\alpha_0 + \Sigma x_i, \beta_0 + n - \Sigma x_i) \). 6.9 \( \mathcal{N}(\Omega (\Sigma_0^{-1} y + M_0^\top \tau_0; \Omega), \Omega) \), where \( \Omega^{-1} = \Sigma_0^{-1} + M_0^\top M_0 \).

CHAPTER 10: TESTING STATISTICAL HYPOTHESES

1.1 Critical values for \( \tilde{\gamma} \geq 0.1372 \); reject at 5% level.” 1.3 Critical values for \( \tilde{\gamma} \leq 19.73 \); accept at 5% level.” 1.5 Critical region is \( \{y : \tilde{\gamma} \geq 75 + 2.326 \times 0.3795 = 75.8833\} \): \( \mathcal{G}(\theta - 75/0.3795 - 2.326) \). 1.8 Critical region is \( \{y : \Sigma (y_i - \mu_\theta^2) \geq a^2(\theta)\} \). Test independent of \( \sigma_\theta \geq 0 \). \( \theta(\alpha) = 1 - H(\sigma_\theta^2(\alpha)/\sigma^2) \), where \( H \) is the distribution function for chi-square \( n \).

1.10 Reject for large \( t = \Sigma y_i\), that is, \( \phi(\gamma) = 1, a, 0 \) according as \( t = \Sigma y_i \geq k \), where \( P(t > k) = \alpha \), \( \theta(\alpha) = \alpha - P(t > k) \), \( \{P(t > k) - P(t > k)\} \) based on the Poisson \( n\theta_0 \). 1.14 Reject if \( \gamma = \max y_i > \theta_0 \) and for \( t \) values in \( (0, \theta_0) \) such that \( \int_0^{\theta_0} \phi(t) t^{n-1} - \theta_{\tilde{\gamma}}^n dt = \alpha \). Test does not depend on \( \theta(\tilde{\gamma} > \theta_0) \). 1.16 Let \( k \) be the \( \alpha \) point on the right tail of the \( G \) distribution: \( G(k) = 1 - \alpha = G(k - 0) \); choose \( a = G(k) - (1 - a) \).

2.1 Critical region is \( \{y : \tilde{\gamma} \leq \theta - \gamma_{\alpha}^2(\sqrt{n})\} \). 2.2 Critical region is \( \{y : 2 \Sigma y_i > \theta_0 y(\sqrt{n})\} \). 2.5 (a) As in Problem 10.1.10; (b) Test function is \( \tilde{\gamma}/\sqrt{n} \).

3.1 \( L(y) = \exp \{n(\tilde{\gamma} - \gamma)/\gamma^2\} \); same distribution. 3.3 \( L(y_1, y_2) = (\Sigma (y_i - \tilde{\gamma})^2 + \Sigma (y_i - \tilde{\gamma})^2 - \tilde{\gamma}_i^2)/((m + n)/m) \).

3.6 \( S \tilde{\gamma} = \tilde{\gamma}(\theta) \gamma(\theta) + \gamma(\theta) \tilde{\gamma}(\theta) \).

3.7 \( z = \ln \{a_0/\gamma - (1/2)(1/\gamma - 1/\gamma)\} \).

4.1 \( \chi^2 = 4(15.94804) = 63.7922 \) and \( \chi^2 = 16.75 \); observed value far out on the right tail; very strong evidence against the hypothesis. 4.5 \( \chi^2 = 4(3.6616) = 14.646 \) and \( \chi^2 = 14.45 \); moderate evidence against the hypothesis. 4.9 \( \chi^2 = 4(3.3032) = 13.213 \) and \( \chi^2 = 12.84 \); strong evidence against the hypothesis.

5.1 \( (13 \pm 1.96 \sqrt{0.05}/\sqrt{5}) = (13 \pm 0.215) \). 5.2 \( (13 \pm 2.78 \sqrt{0.05}/\sqrt{5}) = \)
(13 ± 0.278). 5.6 $P(y_{(1)} < \theta < y_{(n)}) = 1 - 1/2^{n-1}$. 5.7 The interval $(y_{(1)}, y_{(n)})$ does cover $\theta$. $P(y_{(1)} < \theta < y_{(n)}; y_{(n)} - y_{(1)} = r) = r/(1 - r)$ if $r < 1/2$ and 1 if $r \geq 1/2$.

**CHAPTER 11: LINEAR MODELS**

1.4 $d' = (-0.0333, 0.2666, -0.2333)/0.355903 = (-0.09366, 0.74927, -0.65561)$; density for \( u = (T - 0.65561)/1.40488 \) \( 2(1 + u)^{-3} \); central 95% interval for \( u \) is (0.01274, 5.3246) and for \( T = (0.67351, 8.13603) \); (2.5333 - 8.13603 × 0.355903, 2.5333 - 0.67351 × 0.355903) = (-0.362, 2.294). 1.6 Density is $n[(n-1)^{1/2}(n-1)]^{-1/2} (1 + x^2)^{1/2}$ on $x > 0$ and zero elsewhere; density is $n(n-1)(n-1)^{1/2}(n-1)^{-1}$ on $0 < y < \infty$ and zero elsewhere. 1.8 The likelihood function is greater than 3 between $\lambda = 1$ and $\lambda = 5$ and is greater than 2 between $\lambda = 0.5$ and $\lambda = 9$. Ninety percent confidence intervals for $\mu$ are: $\lambda = 1$, (1.029, 1.630); $\lambda = 3$, (0.891, 1.813); $\lambda = 6$, (0.857, 1.985); $\lambda = 9$, (0.851, 2.073); and $\lambda = \infty$, (0.817, 2.229).

<table>
<thead>
<tr>
<th>$\hat{L}(x_i)$</th>
<th>1</th>
<th>0.8$w_1$</th>
<th>3.840</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{L}(x_i \perp x_j)$</td>
<td>1</td>
<td>1.314$w_2$</td>
<td>0.394</td>
</tr>
<tr>
<td>$\hat{L}(x_1, x_2)$</td>
<td>4</td>
<td>$y - 0.8w_1 - 1.314w_2$</td>
<td>0.126</td>
</tr>
<tr>
<td>$\hat{L}(x_1, x_2, x_3)$</td>
<td>6</td>
<td>$y$</td>
<td>4.360</td>
</tr>
</tbody>
</table>

where $w_1 = (1, 1, 1, 1, 1, 1)$, $w_2 = x_2 - 1.91667x_1$.

<table>
<thead>
<tr>
<th>$\hat{L}(x_i)$</th>
<th>1</th>
<th>12.760$w_1$</th>
<th>814.0880</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{L}(x_i \perp x_j)$</td>
<td>1</td>
<td>2.730$w_2$</td>
<td>1.1028</td>
</tr>
<tr>
<td>$\hat{L}(x_i \perp x_j, x_k)$</td>
<td>1</td>
<td>$-1.002w_3$</td>
<td>0.0037</td>
</tr>
<tr>
<td>$\hat{L}(x_1, x_2, x_3)$</td>
<td>2</td>
<td>$y - 12.760w_1 - 2.730w_2 + 1.002w_3$</td>
<td>0.0055</td>
</tr>
<tr>
<td>$\hat{L}(x_i \perp x_j, x_k, x_l)$</td>
<td>5</td>
<td>$y$</td>
<td>815.2000</td>
</tr>
</tbody>
</table>

4.1 $20.5718v_1 + 6.0083v_2 = 0.80178v_2$. 20.5718$v_1 + 6.0083$v_2. 4.3 $F = 0.394/(0.126/4) = 12.51$ cf $F(1, 4)$; 5% and 1% points are 7.71, 21.20; significance at the 5% level (observed significance is 2.4%) $(1.314 \pm 2.776 \times 1.775 \sqrt{4.38} = (1.314 \pm 1.031)$. 4.4 $F = 0.0037/(0.0055/2) = 1.34$ cf $F(1, 2)$; $F_{0.05}(1, 2) = 18.5$: no significance at the 5% level. 4.1 $1.1028/(0.0055/2) = 401$ cf $F(1, 2)$; $F_{0.05}(1, 2) = 199$: significance at the 1/2% level (observed significance is 0.25%) $(2.730 \pm 4.303(0.00275)1/2(6.757)1/2) = (2.730 \pm 0.587)$. 4.2 $F_{0.05}(2, 11) = 3.98$. $F_{0.01}(2, 11) = 7.21$: significance at 5% level (observed significance is 2.6%). Evidence of dependence.

**CHAPTER 12: THE DESIGN OF EXPERIMENTS**

<table>
<thead>
<tr>
<th>1.1 Mean</th>
<th>1</th>
<th>13.454</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{L}$</td>
<td>2</td>
<td>458.85</td>
</tr>
<tr>
<td>Residual</td>
<td>11</td>
<td>487.15</td>
</tr>
</tbody>
</table>

$F_{0.05}(2, 11) = 3.98$. $F_{0.01}(2, 11) = 7.21$: significance at 5% level (observed significance is 2.6%). Evidence of dependence.
1.3 \[\begin{array}{|c|c|c|c|}
\hline
\text{Mean} & 1 & 1628.176 & \\
\hline
\text{A} & 4 & 2.224 & 0.556 & F = 23.2 \\
\hline
\text{Residual} & 5 & 0.120 & 0.024 & \\
\hline
\end{array}\]

\(F_{0.01}(4, 5) = 11.4\); significance at 1% level (observed significance is 0.2%). 1.4 Regression entries can be transferred from Problems 11.2.2 and 11.4.4 (multiply by 2 because MS’s estimate variance for individuals, whereas former MS’s estimate variance for an average of two individuals).

\[\begin{array}{|c|c|c|c|}
\hline
\text{Mean} & 1 & 1628.176 & \\
\hline
\text{A} & \text{linear} & 1 & 2.2056 & 2.2056 \\
\text{quadratic} & 1 & 0.0074 & 0.0074 \\
\text{other} & 2 & 0.0110 & 0.0055 \\
\text{Residual} & 5 & 0.1200 & 0.024 & \\
\hline
\end{array}\]

Test for other \(A\) dependence \(F = 0.0055/0.024 = 0.23\); no significance. Test for quadratic \(A\) dependence \(F = 0.0074/0.024 = 0.31\), no significance. Test for linear \(A\) dependence \(F = 2.2056/0.024 = 91.8, F_{0.01}(1, 5) = 16.3\); significance at 1% level (observed significance is 0.02%). \((2.73 \pm 2.57(0.024)^{1/2}(6.757/2)^{1/2}) = (2.73 \pm 0.732)\).

2.1 \[\begin{array}{|c|c|c|c|}
\hline
\text{Mean} & 1 & 4446.7500 & \\
\hline
\text{T} & 2 & 72.0600 & 36.0300 & F = 1080 \\
\text{P} & 1 & 39.6033 & 39.6033 & F = 1190 \\
\text{T} \times \text{P} & 2 & 0.1867 & 0.0934 & F = 2.8 \\
\text{Residual} & 6 & 0.2000 & 0.0333 & \\
\hline
\end{array}\]

\(F_{0.05}(2, 6) = 5.14\); no significance for \(T \times P\) interaction (observed significance is 14%); \(F_{0.01}(2, 6) = 10.9, F_{0.01}(1, 6) = 13.7\). Significance for \(T\) and for \(P\) at 1% level (observed significances are 0.02% and 0.04%).

\[\begin{array}{|c|c|c|}
\hline
\text{T linear} & 1 & 72.0000 & 72.0000 & F = 2160 \\
\text{other} & 1 & 0.0600 & 0.0600 & F = 1.8 \\
\hline
\end{array}\]

No significance for \(T\) other. Significance for \(I\) linear at 1% level (observed significance is 0.087%).
INDEX

absolutely continuous component, 52
absolutely continuous distribution, 53
  on $R^n$, 74
acceptance region, 413
acceptance sampling, 158, 175, 176
additive, 21
  finite, 28, 29
affine function, of normal variables, 234,
  240
algebra
  of events, 9
  smallest, 19
  of subsets, 14, 515
almost surely, 34
alternative hypothesis, 412
analysis of covariance, 594
analysis of variance, 483
  analysis of covariance, 596
  one-factor design, 494
  randomized blocks, 591
  three-factor design, 588
  two-factor design, 504
angle, 522
angular transformation, 446
antecedent event, probability for, 157, 160
average, 18, 207
  distribution for normal, 323
  finite population
    mean, 289
    variance, 289
  limiting value, 563
  mean, 208
  variance, 208
average on $R^n$, 211
  mean, 211
  transformations, 216
  variance, 211

basis, 520
Bayes analysis, other uses, 576
Bayes formula, 161, 573
Bayes solution, 576
Bernoulli($p$), 44
  probability-generating function, 237
  variance, 197
Bernoulli($p_1, \ldots, p_k$), 138
beta($\alpha, \beta$), 66
  generating functions, 239
  marginal from chi squares, 119
  mean, 205
  moments, 239
  variance, 205
beta function, 69
binary expansion, 517, 524
binary number, 33
  for events, 8
binomial($n, p$), 25, 50
  characteristic function, 232
  completeness, 406
  confidence interval, 354
  convolution, 147, 240
  density function, 554
  derivation, 137
  distribution function, 28, 147
  factorial moments, 238
  generating functions, 238
  mean, 192
  mode, 147
  moment-generating function, 232
  normal limit, 253
  Poisson limit, 147, 245
  recursion relation, 147
  tests, 354
  variance, 202
binomial theorem, 132
birth-and-death process, 538
blocks, randomized, 590ff.
Boolean algebra, 9, 10
Borel-Cantelli lemma, 517, 561
Borel sets, 15, 16
Bose-Einstein, 135
Brownian process, 541

615
cardinality, 22
cardinality measure, 22
Cauchy
  characteristic function, 240
  convolution, 240
  IPIT, 95
  Cauchy(μ, α), 61
cause-effect, 274
central limit theorem, 269
generalizations, 262, 263
proof, 261
central moments, 226
change of variable
density function, 83
  on R^n, 91
  on R^2, 87
  discrete, 81
distribution function, 80
  probability differential, 84, 91
characteristic function, 231
characteristic function on R^n, 235
Chebyshev inequality, 189, 198, 223
chi square(1), 107
chi square(2), IPIT, 95
chi square(m), 66
  convolution, 118, 127, 240, 527
  generating functions, 239
  mean, 205
  moments, 239
  proportion, 127
  ratio, 127
  sum, 127
  tables, 601
  variance, 205
chi-square test, 443
  of normal, 59
  of Poisson, 48
combinatorial symbol, 130, 132
  generalized, 131, 133
comparing models, 124
complement of event, 3
complement of unions, 18
complete statistic, 402
  example, 402
completeness, 402, 437
component, 522
composite hypothesis, 415
concomitant variable, 594
conditional distribution, for inference, 459
conditional mean, 217
conditional mean on R^n, 221
conditional probability, 155
  cautions, 161
  conditions for, 162
  independence, 159
  by probability density, 168
  general, 172
  by probability functions, 163
conditional sign symmetry, 296
conditional test, 437
conditional variance, 217
conditional variance on R^n, 221
confidence interval, 579
  acceptance region, 580
  binomial, 581
  component parameters, 583
  general formulation, 579
  large-sample, 353
  for location, 461
  for mean, 279, 461
  normal theory, 327, 330
  variance unknown, 283
for mean response with regression model, 482
for median, 295, 300, 461
  by pivotal quantity, 581
  power, 582
  for regression coefficient, 485
  for scale, 462
  for variance, 462
  normal theory, 328
for variance ratio, normal theory, 584
confidence regions (see confidence interval)
continuity, point of, 247
continuity theorem, 248
proof, 559
continuous component, 51
control, 491
control limits, 564
convergence
  almost surely, 556
  criteria, 557, 558
  in distribution, 246, 555
  of functions, 517
  in mean(), 556
  in probability, 555
convergence to a constant, 251
convolution, of normal, 234
convolution formula
  absolutely continuous, 115
  difference, 118
  discrete, 110
  linear combination, 119
  proportion, 119
  ratio, 119
correlation, 203
  sample, 213
correlation inequality, sample, 214
count data, 136ff., 442ff.
counting measure, 554
counting sequences, 128, 129
counting sequences of sets, 131
counting sets, 130
covariance, 198
  linear combinations, 200
  sample, 212
  convergence, 264
covariance inequality, 203
  sample, 214
Cramér–Rao inequality, 389
critical region, 413
cumulants, 241

data reduction, 98
decision function, 413
decision space, 413
decision theory, 413ff., 577ff.
decisions, 284
degrees of freedom, 483
De Morgan's laws, 18
density function, of a measure, 549
dependence
  additive, 501
  interactive, 501
  linear, 520
descending factorial, 129, 132
design factors, 491
development vector, 364
difference of means, 285, 331, 332
  nonparametric, 303ff.
  normal
    confidence interval, 332
    test, 331
diffuse prior, 577
Dirichlet($r_1, \ldots, r_{k+1}$), 79
  conditional mean, 223
  conditional variance, 223
  conditionals, 175
  marginals, 118
  mean, 206
  variance, 206
 discrete component, 43
discrete distribution, 24, 44
  on $R^k$, 73
discrete function, 519
disjoint events, 3
distribution, 43
distribution-free methods, 292ff.
distribution function, 40
  properties, 43
  on $R^k$, 70
  properties, 73
distribution generating function, 240
distributions
  absolutely continuous, 52
  on $R^k$, 74
  beta($p, q$), 66
  binomial($n, p$), 50
  Cauchy($\mu, \sigma$), 61
  chi-square, 66
  Dirichlet($r_1, \ldots, r_{k+1}$), 79
discrete, 44
  on $R^k$, 73
double exponential, 62
  exponential($\theta$), 60
  extreme value, 64
  $F(m, n)$, 66, 67
  gamma($\rho$), 65
  geometric($\rho$), 49
  hypergeometric($N, n, D/M$), 51
  Laplace, 62
  logistic, 63
  lognormal, 63
  negative binomial($r, p$), 50
  normal($\mu_1, \sigma_1, \sigma_2, \rho$), 95
  normal($\mu, \sigma$), 55
  normal($\mu; \Sigma$), 93
  normal on $R^k$, 76
  Pareto($\alpha$), 62
  on $R^k$, 79
  Student($\alpha$), 67
  on $R^k$, 79
  triangular, 60
  uniform($a, b$), 53
  uniform($1, \ldots, N$), 49
  Weibull($\beta$), 63
  dominated convergence theorem, 551
dominated derivative theorem, 551
double exponential, 62
efficiency, of unbiased estimators, 381
error distribution (see variation)
errors, 414
estimable parameter, 401
estimation, 362ff., 456ff.
  Bayes, 576
  conditional, 456ff.
  Gauss–Markov, 385, 389
  information inequality, 392
  large-sample, 394
  least-squares, 363, 370
  maximum-likelihood, 367
  for mean, 276
  method of moments, 366
  methods of combining, 379ff.
sufficiency and completeness, 398ff.
  for variance, 281
event, 2
  for a function, 100, 104
  as a subset, 12
events, algebra of, 9
exact size $\alpha$, 415, 418
expectation (see mean value)
experimental design, 273
experimental units, 271, 491
exponential form, 342
  completeness, 406
  likelihood statistic, 342
  location model, 345
  location scale model, 345
  reduced, 342
  sampling, 343
exponential(1)
  generating functions, 239
  mean, 205
  moments, 239
  sample property, 92
  variance, 205
expontntal(\theta), 60
completeness, 405
IPIT, 95
extension theorem, 515
extreme value, 64
IPIT, 95
transformation of, 94

\(F(m, n)\)
canonical, 66
generating functions, 239
marginal from chi squares, 119
mean, 205
moments, 239
from samples, 325
standard, 67
tables, 602, 603, 604
variance, 205

\(F\)-statistic distribution
normal theory, 528
noncentral, 529
factorial design, 490
factorial function, 25
factorial moments, 228
factorization theorem, 337
factors, 273
Fatou lemma, 552
Fermi–Dirac, 135
finite populations, 286
Fisher information, 351, 356
Fisher test, 298
fitted response, 364
fitting affine function, 472
fitting quadratic function, 473
frequency counts (see also binomial; hypergeometric; multinomial)
general, 143
double counts, 148
Fubini theorem, 552
fundamental lemma, 416, 418, 420
monotone likelihood, 425, 426
unbiased tests, 435, 437

gamma function, 69
gamma(1), 85
gamma(\rho), 65
convolution, 240
generating functions, 239
mean, 205
moments, 239
variance, 205
Gauss–Markov, 384, 385, 389
Gaussian process, 541
generating function cumulants, 241
geometric(\rho), 49
factorial moments, 238
generating functions, 238
mean, 204
variance, 204

hypergeometric\((N, n, p)\), 51
derivation, 140
factorial moments, 238
generating functions, 238
mean, 204
variance, 204
hypergeometric\((N, n, p_1, \ldots, p_k)\)
completeness, 406
conditionals, 174
derivation, 142
factorial moments, 239
marginals, 147
mean, 205
variance, 205

IID, 125
independence
conditional density function, 569
by conditional probability, 155
conditional probability function, 166
by distribution function, 123, 127, 526
normal(\mu_1, \rho), 491, \sigma_1, \sigma_2, \rho), 531
normal(\mu; \Sigma), 531
by probability density, 124, 127
by probability function, 123, 127
independent algebras, 121
independent events, 120
independent functions, 122
independent random variables, 122
indicator function, 4, 8, 16, 511
positive axis, 302
unit interval, 334
inference, 269ff. (see also confidence interval; estimation; tests)
location-scale, 480
mean, 276ff.
finite population, 287
median, 294ff.
normal, 326
regression coefficients, 480
information, 351, 390
combining, 397
information inequality, 392, 396
vector, 396
inner product, 520
inner products of residuals, 469
insufficient reason, 577
integral
absolute value, 551
linearity, 550
nonnegative function, 547
real-valued function, 549
interaction, 503
contingency tables, 449
intersection of events, 3
inverse probability integral transformation, 87
inversion formula, 546
IPIT, 87
Jacobian, 88, 91
Jensen inequality, 286

k statistics, 241
Kolmogorov inequality, 223
Kolmogorov lemma, 561
Kolmogorov theorem, 562

Laplace, 52
  generating functions, 239
  mean, 205
  moments, 239
  variance, 205
large samples
  confidence intervals, 280, 284, 353, 357
  tests, 277, 282, 353, 357, 434
law of large numbers
  strong, 560
  weak, 256
least favorable distribution, 423
least favorable value, 421
least squares, 363
  fit of affine function, 378, 472
  fit of quadratic function, 473
  weighted, 378
least-squares equations, 373
least-squares estimate, 365
least-squares solution, 378
Lehmann–Scheffé, 402
likelihood
  independence, 336
  large samples, 348
likelihood function, 309, 312, 358
  assessment, 310
  Bernoulli, 315
  large-sample distribution of, 352, 356
  normal, 314, 316
  distribution of, 349
  for shape parameter, 461, 480
  from sufficient statistic, 338
likelihood principle, 311
likelihood ratio, 313, 416, 418
  generalized, 432
  large samples, 433
likelihood statistic, 332, 334, 398
  Bernoulli, 334
  fixed dimension, 344
  hypothesis testing, 428
  normal, 334
  uniform, 334
liminf, 517, 537
limit, of functions, 517
limiting distribution, 248
  by characteristic function, 248
  by density function, 247
  by moment-generating function, 248
limiting function, 555
limsup, 517, 537
linear function, of normal variables, 234, 240
linear model (see also regression model)
  for location, 372
  maximum likelihood, 388
  normal
    completeness, 407
    likelihood statistic, 347
linear regression, 224
linear subspace, 520
linearly dependent, 520
linearly independent, 520
locally most powerful, 427
locally unbiased, 391, 396
locally unbiased test, 435
location
  confidence interval for, 461
  inference for, 461
  test of, 461
location-scale model, 456
log-likelihood function, 312
logistic, 63, 81
  generating functions, 239
  mean, 205
  moments, 239
  variance, 205
lognormal, 63
  mean, 205
  moments, 239
  variance, 205
loss function, 414

Mann–Whitney confidence interval, 304
Mann–Whitney test, 303
marginal distribution, 99
  by characteristic function, 236
  by conditional probability, 159
  by distribution function, 105, 106
  for frequencies, 136
  for general frequencies, 143
  by measure, 101
  by moment-generating function, 236
  by moment sequence, 236
  by probability density, 111, 112, 115
  by probability function, 109
marginal model, 99, 102
marginal probability, 101 (see also marginal distribution)
Markov process, 539
Martingale, 421
matching problem, 144, 514
maximum likelihood
  least absolute durations, 370
  least squares, 368
maximum likelihood estimate, 315, 367
maximum likelihood method, 367
Maxwell–Boltzmann, 135
mean, 186
mean (continued)
  confidence interval for, 279, 283, 289, 
  295, 300 (see also large samples; 
  location: mean of normal)
  derivative of, 390, 396
  on R^k, 190
  symmetry, 187
  test of, 277, 282, 289, 294, 298 (see 
  also large samples; location: mean of 
  normal)
  mean of normal
    confidence interval, 327, 330
    test of, 326, 329
mean value
  additivity, 513
  bounded function, 182
  discrete function, 519
  general function, 183, 184
  properties, 185
  simple function, 181, 512
  in terms of distribution function, 554
measurable function. 101, 104, 534
  criterion for, 536
measurable space, 14
measure, 21
  discrete, 22, 24
  finite, 23
  probability, 23
  σ finite, 23
  symmetric, 22
measurement
  direct, 362
  indirect, 362
  wavelength, 407
measuring, 19
measuring area, 516
measuring length, 516
median
  efficiency, 381
  limiting distribution, 254
median test, 303
  confidence interval, 303
minimal sufficient statistic, 338
minimum variance, 382, 384, 385, 389, 
  393, 396, 402, 404
MLE, 315, 367
moment-generating function, 231
  properties of, 233
  on R^k, 235
moments, 224
  method of, 366
moments of sample, convergence, 264
monotone convergence theorem, 548
monotone likelihood ratio, 424
Mood-Westenberg test, 303
multinomial(n, p_1, . . ., p_k), 138
  completeness, 406
  conditional mean, 220
  conditional variance, 220
  conditionals, 166, 174
derivation, 138
factorial moments, 239
generating functions, 239
limiting distribution, 264
marginals, 147
mean, 205
  Poisson limit, 147
  variance, 205
multinomial theorem, 133
  multiplication for probabilities, 158
negative binomial(k, p), 50
  derivation, 147
  factorial moments, 238
  generating functions, 238
  mean, 204
  variance, 204
Newton-Raphson, 394
Neyman-Pearson lemma (see fundamental 
  lemma)
noncentral chi square(m), 529
noncentral F(m, n), 529
noncentral normal, 529
noncentral Student(n), 529
nonnormal variation, 457
nonparametric inference, 292ff.
normal
  conditionals, 172
  distribution theory, 319
  orthogonal transformation, 321
  tables, 599
normal(μ, σ), 55
  central moments, 227
  characteristic function, 234
  completeness, 406
  convolution, 127, 234, 240
  mean, 188
  moment-generating function, 234
  variance, 197
normal(μ, σ^2), completeness, 403
normal(μ; 2), 93
  completeness, 407
  conditional, 177
  conditional mean, 222
  conditional variance, 222
  convolution, 240
  likelihood statistic, 347
  marginals, 177
  mean, 192
  moment-generating function, 235
  transformation, 240
  variance matrix, 202
normal(μ_0, σ), completeness, 403
normal(μ_1, σ_2, σ_1, σ_2, p), 95
  conditional mean, 220
  conditionals, 177
  marginals, 177
  mean, 206
  variance, 206
normal(0, 1), 55
square of, 107, 116, 240
normal data, 57
normal on $R^n$, 77
normal process, 541
null hypothesis, 412

one-factor design, 491
operating characteristic, 414
order statistic, 18
order-statistic distributions, 532, 533
orthogonal, 522
orthogonal components, 322, 365, 375, 376
orthogonal transformation, 320
orthonormal basis, 321, 474

parameter change, information inequality, 397
Pareto(a), 62
generating functions, 239
IPIT, 95
mean, 205
moments, 239
on $R^n$, 79
conditionals, 175
marginals, 118
transformation of, 94
variance, 205
Parseval relation, 545
partition, 20
of events, 3
of a function, 103
percentile, limiting distribution, 254
personal prior, 577
PIT, 86
Pitman test, 304
pivotal quantity, 581
Poisson(λ), 2, 5, 45
completeness, 403
conditionals, 174
convolution, 164, 240
density function, 554
distribution function, 28, 147
factorial moments, 229, 238
generating functions, 238
limiting distribution, 453
mean, 188, 204, 519
mode, 147
normal limit, 249, 250, 253
recursion relation, 147
square root, 250
variance, 197, 204
Poisson data, 46
Poisson process, 538, 540
Polya urn, 542
positive definite, 522
greater than, 523
posterior distribution, 573

power function, 413
power set, 14
precision, 574
preimage, 100
prior distribution, 572
probability
a posteriori, 156
a priori, 156
at least one event, 36, 143, 514
conditional, 155
empirical example, 4, 8, 47, 58
marginal, 101
monotone property, 31
r events, 143, 514
r or more events, 515
probability density function, 52
properties, 54
on $R^n$, 74
properties, 75
probability differential, 56
probability function, 43
properties, 45
on $R^n$, 73
properties, 74
probability-generating function, 236
probability integral transformation, 86
probability measure
alternative definition, 33
definition, 23, 33
symmetric, 26
probability model, 26
probability of a limit, 32
probability of complement, 29
probability of union, 30
probability space, 27
product algebra, 124
product measure, 124
product space, 124, 534
projection, 468, 522
by differencing, 469
one-factor design, 493
into orthogonal complement, 469
two-factor design, 585
three-factor design, 503
projection matrix, 374
proportions, limiting value, 563

quadratic form, 522
quality control, 564
quantitative factor
one-factor design, 495
two-factor design, 507

random bits, 49
random digits, 49
random variable, 96, 102
randomization, 158, 274
randomized blocks, 590
randomized test function (see test function)
rank function, 302
rank statistic, 117
Rao–Blackwell, 399, 401
ratio of variances, normal
  confidence interval, 332
tests of, 332
reciprocal variance, as weights, 574
reduction, 98, 99, 100
regression coefficients, 374, 470
distribution theory, 480
regression model, 476
  conditional distribution, 479
distribution theory, 479, 481
  inference, 482
regression of y on x, 218
replication, 274
residual vector, 374
ring, 515
risk, 414
run test, 567
distribution, 569
Rutherford and Geiger, 46

σ algebra (see algebra)
sample: Bernoulli(p), 125
sample: normal(μ, σ), 126
sample: Poisson(λ), 126
sample from distribution, 125
sample point, 11
sample space, 11
scale
  confidence interval for, 462
  inference for, 462
test of, 462
Schwartz inequality, 554
score equation, 367
score function, 312
  combining, 397
  local tests, 428
  normal, distribution of, 350
  properties, 351, 390, 396
selection, 153
sequential analysis, 438
sequential test, 439
  critical values, 440
  optimality, 440
serial system, 31
sign confidence interval, 295
sign function, 293
sign test, 294
signed deviation confidence interval, 299
signed deviation test, 298
signed rank confidence interval, 302
signed rank test, 302
similar test, 432, 437
simple function, 511
simple hypothesis, 415
simple random sampling, 287
singular component, 523, 524
size α, 415
SLLN, 560
smallest ring, 515
squared length
  by differencing, 469
  of projection, 374
  of residual, 374
standard deviation (see also variance)
estimator, 406, 566
standard deviation of sample (see variance of sample)
state space, 537
statistical model, 271
  parametric, 308
statistically independent (see independence)
Stirling’s formula, 133, 136
stochastic process, 537
stratified sample, 291
strong law of large numbers, 560
Student(ν)
canonical, 67
  limiting distribution, 247
  marginal from (n, χ), 119
  mean, 205
  moments, 240
  on Rp, 79
  conditionals, 175
  marginals, 118
  standard, 67
tables, 600
  variance, 205
sufficient statistic, 337
  fixed dimension, 344
  hypothesis testing, 428
  sums of squares, 483
symmetric distribution, 26 (see also distributions, uniform)

t-statistic, 18
t-statistic distribution
  general theory, 461
  normal theory, 466, 527
  noncentral, 529
t-test, 438
ternary expansion, 524
test function, 418
tests, 412ff.
  of fit, 442
  of independence
    contingency tables, 449
  large sample, 353
  of location, 461
  of mean, 277
  finite population, 289
  variance unknown, 282
  of median, 294, 298
  of probability model, 445
of regression coefficients, 485
of scale, 462
of statistical model, fitted parameters, 448
of variance, 328, 484
of variance ratio, 332
three-factor design, 584
$2^3$ design, 588
transition matrix, 540
treatments, 491
triangular distribution, 60
two-factor design, 501
two-sample problems, 285
Type I error, 414
Type II error, 414

UMP, 417, 423, 426
UMP similar, 432
UMP unbiased, 431
UMV unbiased, 402
  by completeness, 402, 405
  examples, 404
  by information, 393, 396
  method, 395, 396
unbiased estimator, 276, 379, 380
  better than, 382
  combining, 382, 384, 388
  comparison, 380, 382
  efficiency, 381
exponential models, 395, 396
Gauss–Markov, 384, 385
general Gauss–Markov, 389
large sample, 394
likelihood statistic, 399, 401
local, 391, 393, 394, 396
minimum variance, 401
  uniqueness, 402
sufficient statistic, 399, 401
variance reduction, 399, 401
unbiased test, 431, 435
uniform$(a, b)$, 53
  mean, 187
  variance, 196
uniform$(1, \ldots, N)$, 49
  factorial moments, 238
  generating functions, 238
  mean, 204
  variance, 204
uniform$(0, \theta)$, completeness, 403
uniformity trials, 590
uniformly most powerful (see UMP)
union of complements, 18
union of events, 3
uniqueness theorem, 232, 544, 545, 547
  characteristic function, 232, 546
  moment-generating function, 232, 546, 547
  moment sequence, 546, 547
unknowns of a random system, 270

variable
  input, 1
  response, 1
variance, 194
  linear combination, 200
  transformations, 195
  zero, 198
variance about regression, 219
variance matrix, 199
  formulas, 199, 200
variance matrix of sample, 213
  transformations, 216
variance of normal
  confidence interval, 328
  test of, 328
variance of sample, 18, 58, 210, 273, 281
  convergence, 264
  distribution for normal, 323
  finite population, mean, 289
  mean, 210
  unbiasedness, 380
  variance, 210
variance ratio, 325
variance ratio test, 430
  power function, 430
variation, 1, 456, 476
  Student distribution for, 463
vector algebra, 520

Wald–Wolfowitz, 440
weak law of large numbers, 256
Weibull$(\beta)$, 63
  IPIT, 95
  mean, 206
  moments, 240
  transformation of, 94
  variance, 206
Wilcoxon test, 302, 304
WLLN, 256
  alternative proof, 265
  proof, 263