STATISTICS, FOUNDATIONS

D.A.S. Fraser, University of Toronto

I. Background
II. Overview
III. Probability Model
IV. Statistical Model
V. Statistical Theory
VI. Foundations
VII. Principles
VIII. Likelihood Asymptotics
GLOSSARY

Ancillary:
A statistic with a fixed distribution, free of the parameter; for applications a physical interpretation is appropriate.

Conditioning:
A principle that recommends conditioning on an appropriate or relevant ancillary.

Likelihood:
A principle that recommends using only the observed likelihood function for statistical inference.

Probability model:
A mathematical construct for describing the long-run behavioural properties of a system that is actually or conceptually performable or repeatable under essentially constant conditions.

Statistical Model:
A probability model with free parameters such that certain particular values for the parameters provide a probability model that is a good approximation to the process being examined.

Statistical Reduction:
Sufficiency can reduce the dimension of the appropriate variable and lead to a marginal model. Conditioning can reduce the dimension of the free variable and lead to a conditional model.

Statistics seeks patterns and relationships within information collected concerning real world contexts. The contexts can range from sharply defined scientific issues, to social science matters, to wide ranging industrial, commercial, or financial issues, to the contents of massive data storage devices. The information can be the collected results, tabulations, or electronic records obtained, or to be obtained, concerning the real world contexts. And the patterns and relationships are the simplified descriptions including cause-effect relationships concerning the real world contexts being investigated.

Statistics is the methodology and theory for this process of extracting the important patterns and relationships from the collected information. The foundations of statistics
are the basic elements of theory and principle that structure the planning and collection of information and the extraction of patterns and relationships.
I. BACKGROUND

Some elements of statistical thought were present even in Biblical times. But it was money and gamblers in sixteenth century France that brought mathematical skills to the calculation of probabilities for games of chance, with the intent of increasing financial gains. Also, astronomers had many precise measurements but measurement error created deviations from theoretical patterns, and mathematical skills from Gauss, Laplace and many others were applied to this departure of data from theory. More recently the social sciences collect data where underlying patterns are heavily masked by variation, both variation in the subject material and variation due to measurement error. Also in agricultural experimentation the long delay from planning to data collection provides a strong incentive to carefully plan an investigation and to carefully analyze the results in order to maximize the quantity and quality of the conclusions obtained from the long process from initial planning to final presentation. But now statistics has entered into almost all areas of human endeavor.
II. OVERVIEW

Statistics involves the planning of investigations, the carrying out of the investigation, the collection of data, the processing of the data, and the analysis through to conclusions from the investigation and data. Some of these components however can exist almost in isolation from the others; for example, the analysis of very large data sets on consumer choices can in practice be almost totally separate from the original collection of the data.

Early activity focussed on variation in results from quite well defined processes such as the games of chance in sixteenth century France. The variation or randomness in the results was formalized as probability theory which is now a well established component of statistical modelling.

Early formulations of this theory gave the structure needed to describe the measurement error in traditional physics and astronomy. In these areas there would be an underlying law or theory that would prescribe deterministic results and then the actual data would depart from this due typically to small measurement errors. Probability theory provided a way to model the error and this combined with the deterministic pattern led to statistical models. These were probability models but with free parameters corresponding to characteristics of the underlying deterministic process.

This type of statistical model however has much broader application than the traditional measurement error context. It could describe variation in the underlying material being investigated and even variation in the quantities that were of direct interest. This provides applications to the social sciences and to industrial, commercial, financial and other areas.

Statistical methodology however extends far beyond this probabilistic modelling of error and variation. Small and large aggregations of data can be examined to get summary presentations, to elicit patterns for further investigation, to seek anomalies relative to previous patterns, and generally to obtain guidance for future analysis and development often with a distinct profit motive not unrelated to that of the sixteenth century origins.
of the discipline.

As indicated above a traditionally central part of statistics is concerned with investigations that lead to a statistical model, a probability model with free parameters corresponding to unknowns in the investigation. The primary statistical problem is then the analysis of data that is viewed as coming from this statistical model. Some areas of application may have special types of model form and be developed within related disciplines often with quite special notation. For example, actuarial science examines lifetime of people or objects obtaining probabilities of living, of dying, of accidents in time intervals. More recently this has extended into general insurance, financial, and related areas. Other specialized areas, include operations and management research, control theory, econometrics, and many biological, industrial, management, and financial research areas. Interpreted broadly statistics is the theory and methodology of knowledge acquisition, the use of this knowledge to manage and control, and the source of structure for many substantive areas of intellectual activity.
III. PROBABILITY MODEL

As indicated above the development of probability methods provided the major ingredient for the development of statistics and continues now as a significant core component.

A pure context involves a process or system that can be operated or performed repeatedly under essentially constant condition. For example, an ordinary six faced die can be shaken or tossed and the upward face observed, or the time to failure observed when an electronic component from a stable manufacturing process is tested, or many more. The elementary probability model involves a sample space $S$ of possible outcomes values, a class $\mathcal{A} = \{A\}$ of events or subsets $A$ of $S$ that are of interest, and a probability $P(A)$ for each event $A$. For the die $S = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{A}$ consists of all possible subsets of $S$ such as $\{2, 4, 6\}$, $\phi = \{\}$, or $S$ itself; and for a symmetrical die with implied symmetric probabilities $p(A) = \#A/6$ where $\#A$ is the number of points in $A$. Figure 1 records the amount of probability at each point in $S$ for this symmetric case. For the lifetime example, $S$ consists of the positive real line or perhaps for convenience the whole real line $R$, $\mathcal{A}$ consists of all intervals and things predicated by intervals under countable union and intersection, and $P$ could have many forms depending on the context. One simple possibility is the normal distribution with $P(A) = \int_A f(y)dy$ where $f(y) = (2\pi\sigma^2)^{-1/2}\exp\{-(y - \mu)^2/2\sigma^2\}$ which is located at some point $\mu$ and scaled by $\sigma$. For a location $\mu = 0$ and $\sigma = 1$ Figure 2 records the density function labelled $n = \infty$. The other curves provide examples from the Student family which provides longer tailed distributions than the normal. Of course these distributions are on the whole real line, in an absolute sense, whereas the lifetime variable is necessarily positive. Nonetheless they may provide good approximations in many applications which is the objective of probability modelling. A somewhat different possibility is provided by the exponential ($\theta$) model $f(y; \theta) = \theta^{-1}\exp\{-y/\theta\}$ on the positive line. This can describe a constant failure rate situation and $\theta$ is the mean lifetime. This model is on the positive real line. Of course other models can describe lifetime depending on the underlying process being investigated.
The probability $P(A)$ of an event $A$ is the proportion of occurrences of the event in a long sequence of repetitions; ideally it is the limit of the proportion as the number of repetitions goes to infinity. In applications it is an empirical phenomenon that the proportion $\hat{P}_n(A)$ of occurrence of $A$ in a long sequence of repetitions goes to a limit, or otherwise some change in conditions has occurred that has modified the process or system. This empirical phenomenon for a particular process may be difficult to verify beyond some reasonable approximation, and its assumption is typically based on the degree to which initial conditions are seen to be constant under repetitions.

Probability theory constructs and examines very general models and develops techniques for calculating probabilities for various types of events and also techniques for calculating many important characteristics of probability models.
IV. STATISTICAL MODEL

A basic statistical model is a probability model with free parameters to allow a range of possibilities, one of which leads to a model that provides a reasonable approximation to a process or system being investigated. For example in some context the normal \((\mu, \sigma^2)\) mentioned in II could be very appropriate. Also the die example in Section II might be generalized to have probabilities \(p_1, p_2, p_3, p_4, p_5, p_6\) at the six sample points 1, 2, 3, 4, 5, 6 where necessarily \(p_i \geq 0\) and \(\Sigma p_i = 1\), these being a consequence of probabilities viewed as proportions in a large aggregate.

Thus a general version of the basic model would have a space \(S\), a collection \(\mathcal{A}\) of subsets, and a probability measure \(P(A; \theta)\) for each subset \(A\) in \(\mathcal{A}\). The probability measure includes a free parameter \(\theta\) in a space \(\Omega\) that allows a range of possibilities for the measure; in an application it would be implicit that some one value of \(\theta\) provided probabilities that closely approximated the behaviour in the application.

The basic model as just described is concerned with frequencies or proportions on a space of possibilities. It does not distinguish component elements of structure or the practical significance of the various components. This basic model has been central to statistics throughout the twentieth century.

Some more structured models have been proposed. For example in the measurement context one could model the measurement errors as \(z_1, \ldots, z_n\) from some error distribution \(f(z)\). The error distribution could be the standard normal in Figure 2 or it could be one of the Student distributions with parameter value equal to say 6, which provides more realistic, longer tails. The statistical model would then be \(y_i = \mu + \sigma z_i\) where \(\mu\) and \(\sigma\) represent the response location and scaling. This alternative more detailed model makes the process of statistical inference more straightforward and less arbitrary. For some details see Fraser (1979).

This raises the more general question of what further elements of structure from applications should reasonably be included in the statistical model. It also raises the question
as to what modifications might arise in the statistical analyses using the more detailed or more specific models.

One modelling question was raised by Cox (1958). The context involved two measuring instruments, $I_1$ producing a measurement $y$ that is normal $(\mu, \sigma_1^2)$, and $I_2$ producing a measurement $y$ that is normal $(\mu, \sigma_2^2)$. A coin is tossed and depending on heads or tails we have $i = 1$ or 2 with probability 1/2 each; the corresponding instrument $I_i$ is used to measure \( \theta \). The standard modelling view would use the model $f(i, y; \theta) = (1/2)g(y - \theta; \sigma_i^2)$ where $g(z, \sigma^2)$ designates a normal density with mean 0 and variance $\sigma^2$. In an application the data would be $(i, y)$ with model $f(i, y; \theta)$. But clearly when the measurement is made the instrument that is used in known. This suggests that the model should be $g(y - \theta, \sigma_1^2)$ if $i = 1$ and be $g(y - \theta, \sigma_2)$ if $i = 2$; this can viewed as the conditional model given the value of the indicator $i$.

Traditional statistical theory would give quite different results for the first model than for the second model. A global model may be appropriate for certain statistical calculations. But in an application when you know how the measurement of $\theta$ is obtained, it seems quite unrealistic to the purposes of inference to suggest that the instrument might have been different and that this alternative should be factored into the calculations.
V. STATISTICAL THEORY

Much of statistical theory through until the early decades of the twentieth century was concerned with the analysis of data say $\mathcal{D}$ together with a statistical model say $\mathcal{M}$ representing the source of the data. Thus, what can we conclude from $(\mathcal{M}, \mathcal{D})$ concerning the unknown true value of the parameter say $\theta$ in the statistical model? This represented a substantial portion of the discipline until mid twentieth century and with ups and down remains a major portion of the core discipline.

Some major early workers in the nineteenth century such as Laplace promoted the idea of attaching a probability distribution say $\pi(\theta)$ to the origins of the parameter $\theta$. This might be something approximately uniform representing diffuse information concerning $\theta$ or something more local on certain ranges for $\theta$ as opposed to other ranges of values. This had been promoted earlier by Bayes (1763) and intermittently has had periods of prominence through to the mid-twentieth century.

An alternative decision theoretic approach was promoted by work of Neyman & Pearson (1933) and later generalized by Von Neumann and Morgenstern (1947) and Wald (1950). This viewed statistics as needing a decision function for any application: the data combined with the decision function would produce a decision concerning the unknown. This approach dominated most theoretical thought through until 1955. There is of course the question of whether things can properly or should properly be mechanized in this way. There is also the issue of whether this methodological approach can produce sensible answers to the broad range of practical problems. By the mid 1950’s there were substantial criticisms of the decision theory approach, in particular there had been a major failure of the theory to produce reasonable statistical procedures for a broad range of problems.

In the mid 1950’s two publications Fisher (1956) and Savage (1954) substantially altered the directions of statistics and opened wide areas for development. Fisher proposed insightful methods based on the earlier view of examining the model data combination $(\mathcal{D}, \mathcal{M})$. Savage favored the Bayesian approach emphasizing the use of personal priors to
represent the latent views of the investigator concerning possible values for the parameters.

Both these directions opened new opportunities to a discipline that had become partially paralyzed by the decision theoretic approach and by its inability to produce answers for wide ranging problems.
VI. THE FOUNDATIONS

The foundations of statistics have changed and evolved with time. The early use of probability for statistical analysis was closely tied to the development of the least squares method, a widely used technique dating from Laplace and earlier. The Bayesian approach also comes from this same earlier period. Neither could be viewed at that time as an all embracing foundation for statistics.

The decision theory approach however did present itself as an all embracing theory: start with a model and a utility function and derive the optimum decision procedure for producing a decision from data. Its claims were all encompassing but as mentioned above it failed to deliver for the broad needs of statistics.

The Bayesian approach has close ties to the decision theoretic but has more flexibility in allowing the personal prior or even in allowing many personal priors if many investigations are involved.

The frequentist approach as reemphasized by Fisher (1956) reinvestigated the earlier approach of just analyzing a model together with data.

The Bayesian and frequentist approaches brought rather flexible and innovative directions to statistical theory, after the rigidly of the decision theory approach. This was largely assisted by the increase in available computer power. Both have since existed side by side with intermittent tensions and conflicts. The two approaches represent a major portion of statistics, and have increasing overlap, and are growing closer in overall operation and objective.
VII. PRINCIPLES

As indicated above the combination of a statistical model and data represents the core concern of statistics. Various methods and principles have evolved for the analysis and simplification of the model data combination.

A. Sufficiency Principle

Fisher (1922) defined a sufficient statistic \( s(y) \) as a statistic with the property that the distribution of the basic variable \( y \) given \( s(y) \) is independent of the parameter, say \( \theta \). The idea being that one would use the observed value of the sufficient statistic together with the marginal model say \( g(y; \theta) \) for that statistic, and otherwise ignore the original value of \( y \). In a sense the actual value of \( y \) given \( s(y) \) can be viewed as an observation of pure error with no influence from the parameter \( \theta \), and as such can be viewed as irrelevant to any statistical assessment of \( \theta \). The sufficiency principle (S) is to use only the marginal model for the sufficient statistic in the inference process or assessment of data. As an example consider a sample \( y_1, \ldots, y_n \) from the normal \( (\mu, \sigma^2) \) distribution: \( \bar{y} \) is a sufficient statistic in the sense that the distribution of \( (y_1, \ldots, y_n) \) given \( \bar{y} \) does not depend on \( \mu \). Unfortunately, there are very few problems that admit a simple sufficient statistic that provides a dimension reduction such as the \( n \) to 1 found in this normal example. The concept and principle for sufficiency was widely accepted and adopted but turned out not to be available for most problems other than the familiar textbook cases. The detailed mathematical investigation of sufficiency thus unfortunately diverted statistical attention from seeking more fruitful directions for theory.

B. Likelihood Principle

Fisher (1922) introduced the concept of the likelihood function. The likelihood function \( L(\theta; y^0) \) from data \( y^0 \) with model \( f(y; \theta) \) is the density function \( cf(y^0; \theta) \) examined as a function of \( \theta \) for fixed data value. It would seem from a mathematical viewpoint to be a very obvious expression of what a function \( f(y; \theta) \) would say about \( \theta \) with \( y = y^0 \). It is
usually examined relatively, that is, one \( \theta \) value relative to another \( \theta \) value and this is expressed above by having an arbitrary constant \( c \) in the definition. The likelihood function \( L(\theta) = L(\theta; y^0) \) or its logarithmic version \( \ell(\theta) = \ell(\theta; y^0) = \log L(\theta; y^0) \) can be treated as a very informative numerical evaluation of one \( \theta \) value relative to another. The *likelihood principle* (L) is to use only the observed value \( \ell(\theta) \) of the likelihood function and not any other information from the original model-data combination. One could picture the plot of \( \ell(\theta) \) on a monitor; the principle would say that this plot was the full allowable information concerning \( \theta \) from the model data combination. Thus for the normal example in A above one would use only \( \ell(\theta) = a - n(\theta - \bar{y})/\sigma^2 \) with \( a \) arbitrary and no other information from the original sample \( (y_1^0, \ldots, y_n^0) \) or from the original model \( \Pi_i^n g(y_i - \theta; \sigma^2) \).

C. Weak Likelihood Principle

Sufficient statistics and the likelihood function as noted above date from Fisher's major 1922 paper. In Fisher (1925) he noted a very close connection between them: that the likelihood function as obtained from data was the best, later called minimal, sufficient statistic. Despite this asserted close equivalence the two concepts developed largely independently until the 1960's when informal discussion at meetings finally acknowledged that they were largely equivalent. Even now at the millenium there is little general recognition of the equivalence and only two or several introductory texts on mathematical statistics draw attention to the connection. The weak likelihood principle is to use only the observed likelihood function and the statistical model for the possible likelihood functions. Because of the link between the production of the likelihood function and the concept of minimal sufficiency it follows that the weak likelihood principle and the sufficiency principle are essentially equivalent. Note that the strong likelihood principle is to use the observed likelihood function but not any information concerning possible likelihood function as would be given by the corresponding model.

D. Conditioning Principle
Fisher (1925) defined an ancillary statistic as one with a fixed distribution, free of the parameter of the model. In more recent language we would view an ancillary statistic as one that presents natural variation or error present in the model. The name suggests that it is supportive for a process of inference. Fisher recommended that one condition on the value of the ancillary and thus use the conditional model given the observed value. The *ancillary principle* is to use only the conditional model and the observed value of the conditioned variable. The ancillary principle is often called the *conditioning principle* (C). As an example suppose \( y_1, y_2 \) come from the location model \( f(y - \theta) \). Then it is easily seen that \( a(y_1, y_2) = y_2 - y_1 \) has a fixed distribution free of \( \theta \); Fisher called it a configuration statistic. The conditional model for say \( \bar{y} \) given \( a(y_1, y_2) = a(y^0_1, y^0_2) = a^0 \) is then \( cf(\bar{y} - \theta - a^0/2)f(\bar{y} - \theta + a^0/2) \) and is a simple location model with parameter \( \theta \) and with observed value \( \bar{y} = \bar{y}^0 = (y^0_1 + y^0_2)/2 \). The corresponding \( p \)-value is

\[
p(\theta) = \frac{\int_{-\infty}^{\bar{y}^0} f(\bar{y} - \theta - a^0/2 , \bar{y} - \theta + a^0/2)d\bar{y}}{\int_{-\infty}^{\infty} f(\bar{y} - \theta - a^0/2 , \bar{y} - \theta + a^0/2)d\bar{y}}
\]

and it records the probability position of the observed \( \bar{y}^0 \) in its distribution conditional on the known error value \( a^0 \).

E. Connections among Principles

Birnbaum (1962) discussed the Sufficiency (S), Likelihood (L) and Conditioning (C) principles from an equivalence or set theoretic viewpoint. He showed that S plus C implied L. This caused some major stress in the statistical community as S and C seemed well founded while L seemed far too strong to most statisticians. Evans, Fraser & Monette (1986) then showed that C alone implied the likelihood L principle. There seems to be no widely held resolution among members of the statistical community concerning these interlinks. In fact many are unaware of key aspects of the conditioning principle or the arguments that support that principle. While Fisher (1925) certainly provided the key foundational approach and later (Fisher, 1934) analyzed the location and location
scale models, it was the Cox (1958) example (see Section IV) concerning the measuring instruments that triggered a reexamination of the ancillary approach to inference.

An example with nonuniqueness of the ancillary statistic may be found in Fisher (1956, p.47); also see Basu (1964) and Buehler (1982). This uncertainty kept the conditioning approach from general acceptance, indeed, it left a noticeable taint to the concept. More recent considerations however have emphasized a need for there to be some objective meaning for the ancillary to be used.

Objective ancillaries are central to the analysis of location-scale and transformation models when presented in terms of an error variable. Some general discussion may be found in Fraser (1968, 1979).
VIII. LIKELIHOOD ASYMPTOTICS

Recent likelihood asymptotic analysis developed from approximation theory, in particular, from the saddlepoint work of Daniels (1954) and Lugannani & Rice (1979). This led to the analysis of large sample likelihood functions Barndorff-Nielsen, (1986), Fraser & Reid (1995), Fraser, Reid, Wu (2000). With an increasing number \( n \) of coordinates and a fixed number of parameters it is found that with high accuracy, both theoretically \( O(n^{-3/2}) \) and empirically, there is an ancillary with an approximately fixed distribution and with a conditional distribution with the same dimension as the parameter. Also it is found that for a scalar component parameter there is a subsequent marginal distribution that provides a \( p \)-value for assessing the scalar parameter. Under mild regularity the \( p \)-value is unique.

In the early development of statistical inference, sufficiency was viewed as a prime focal concept, but its availability was extremely limited. The likelihood developments growing out of approximation theory have shown what the possible forms of a large sample model are and thus lead to the appropriate statistical analysis: first reduce by conditioning using the generally overlooked ancillary approach; then marginalize to obtain a \( p \)-value free of the nuisance parameters. For some recent details, see Fraser, Reid, Wu (2000) and Fraser, Wong, Wu (1999).
BIBLIOGRAPHY


