

On the foundations: statistical models and inference*

D. BRENNER and D. A. S. FRASER

University of Toronto

Key words and phrases: Inference, deduction, inference base, distribution form, categorical information, additives.

AMS 1980 subject classification: Primary 62A99.

ABSTRACT

This paper presents an overview of some recent results concerning statistical models and inference, specifically: grounds for statistical models, types of models that simplify by standard probability analysis, the use of categorical information in the reduction of the model with data, and the role of additives in the inference process. The relevant technical material has been developed elsewhere.

1. PREFACE

Statistical inference is taken to be the deductive process in which the *model for an investigation* together with *data* from an investigation, collectively called the *inference base*, is analyzed to determine the implications concerning the unknowns of the investigation, *the inferences*. Inference, thus, is the process of determining “what the model and data say concerning the unknowns”. As such, inference is the central unifying part of statistics.

The statistical model that is used as a component of the inference base can, of course, affect the inferences from the inference base. Validity of the inferences, then clearly devolves from validity of the model. Accordingly, we distinguish the wide-sense use of a model as hypothetical, exploratory, tentative, or for-the-sake-of-argument from the restricted use as a valid model. More specifically, all our considerations of an inference base are in the context that the included statistical model is valid; this in itself imposes certain requirements on that model.

Also of importance in the customary inference arguments are the overt and covert principles, criteria, and methods that are not in the inference base, and thus are, in a clear technical sense, *additives*.

This paper presents an overview of some recent technical material bearing on the inference process: grounds for statistical models, types of models that simplify by standard probability analysis, the use of categorical information in the reduction of an inference base, and the role of additives in the inference process. The relevant technical material has been developed elsewhere, and the present discussion centers on the broader aspects.

2. THE INFERENCE BASE

We take inference to be the central deductive process of statistics. Clearly such a deductive process must start somewhere—we call the formal starting point *the*

* Specially invited paper.

inference base. More incisively we argue that if there is a formal deductive process being examined, then there is *the given*, which in the present context is the aforementioned inference base.

We take the formal starting point to be valid — as providing a valid presentation of information concerning an investigation. Such information involves observations as part of the investigation and background information concerning possibilities for the investigation. We refer to these as *the data* and *the model* respectively.

Accordingly, we take the formal starting point for the inference process to be the *inference base* $I = (M, D)$, consisting of the *data* D and the *model* M . The results of the inference process are called *the inferences*. The notion of the inference base was proposed and discussed in Fraser (1979). Note that the inferences can be viewed as a function of M and D .

As a pure case, we could envisage the context where things are simple and can be made mathematically explicit, thus giving rise to a clear specification of the model; more generally, the model would provide some reasonable approximation, or have some relative level of validity depending on the nature of the background information. These qualifications of the model in terms of degree of accuracy, or degree of validity, would, in turn, be qualifiers for the inferences coming from the deductive process.

3. THE MODEL

Traditionally, the substance of the statistical model is taken to be a collection of probability models, each available as the best description of an observable response variable. In certain formal contexts, the model, then, is a primitive object consisting of a space, an algebra, and a class of measures: (S, A, μ) with $\mu = \{P_\theta: \theta \in \Omega\}$. In other contexts, perhaps most, it is recorded in a more developed form as a class of probability density functions, $\mathcal{F} = \{f_\theta: \theta \in \Omega\}$, where a supporting space, algebra, and dominating measure are taken as background. In terms of the objects that are *given*, these are not the same. On the other hand, we do know that various classes of densities can give rise to the same class of probability measures. Are there other than arbitrary grounds for some particular choice? Or, in some sense, does it matter? Does a moderate amount of indeterminacy or slackness matter — matter from the point of view of accuracy, or (in a less tangible sense) in the effect on inferences?

These questions raise larger questions. Do the preceding primitive models adequately cover all scientific, engineering, and commercial contexts, or are there, in fact, additional elements from these contexts that should properly be included in the model? Certainly, mathematical elements are often added, for convenience or interest in particular analyses: for instance, continuity and differentiability are added to the space S or the densities \mathcal{F} . These added refinements, however, are rarely mentioned at the model presentation stage. To what extent, then, are these seeming “additives” in some sense truly implicit at that stage?

An examination of physical theories, of the description of cause-effect relationships, and of the presentation of circuit diagrams shows more than just a response space with measures or densities, even including the extra smoothness properties. On the other hand, in these richer contexts should some richest model be pursued?

Certainly a scientific investigation cannot just be dismissed as “ y_1, \dots, y_n independently and identically distributed $N(\mu, \sigma^2)$ ”, assuming just a space \mathbb{R}^n with, perhaps, a distinguished coordinate system. But how much more is really needed?

These questions have been addressed in Fraser (1979), under the headings *descriptive* and *exhaustive*: that elements in the model should describe elements in the physical situation, and that distinguished elements in the situation should have correspondents in the model. In part, these requirements amount to accuracy and coverage, and we recognize these as the two primary properties of a model with respect to its validity.

The preceding, however, does not focus on the seemingly pragmatic question of the effect on the inferences. This is sensitivity analysis, or in a more limited sense, robustness analysis. In a larger context, this presumes that the inferences are known, that the inference problem has been solved, or, in a small and expedient sense, that it has been replaced by some limited facet such as point estimation. Clearly, without some fundamental solution of the inference problem, the question of whether or not it matters cannot easily be addressed or examined.

So far in this section we have been discussing the *model for a system*. Such a model is a part of the organization of scientific information and may have many roles and facets. For inference, however, we are working from a particular investigation with particular data. The background information concerning possibilities for the investigation forms the *model for the investigation*. In brief, this is the model for the system as specialized and made relevant to that investigation. Some criteria can be mentioned, but they evolve directly from the fact that we are presenting formally the possibilities for the restricted context. For example, criteria such as "do not treat a known as an unknown" and "do not treat an unknown as a known" may sound unneeded but can have fundamental import. Of course, this is a way of saying — "do not distort or abuse the available information". For a discussion of a significant example, see Feuerverger and Fraser (1980).

The model for the investigation, as a description of possibilities for the investigation, has been discussed in Fraser (1979). In particular, the model for an investigation can be restricted by information from the data itself. For example, an initial assembly of model and data may show that certain things are known categorically concerning basic parameter values or concerning what was initially treated as a variable. Some aspects of this will be discussed in the next section.

A related fundamental question concerns just how large the statistical model should be. As an illustration of the direction we have in mind, consider a response known to be Student(7) in shape, with known scaling but unknown location μ . The conventional modelling procedure involves assembling all the relocated Student(7) distributions as a collection indexed by the real line. In the actual application, there is, however, just one Student(7) distribution. Accordingly, we can ask: is a model that is \mathcal{R} -fold more complicated than the reality being modelled actually necessary, appropriate, or even relevant? Ideally, the model would have just one distribution. In fact, direct analysis in Brenner and Fraser (1980, 1981) supports the use of a single distribution for the basic variable, the *distribution form* of the response variable. Given that one can observe distribution form *apart* from location, the basic modelling criteria require the use of this objective element. The unknown characteristic then concerns how the distribution form is presented on the range of the response variable; this involves the collection of relocation transformations, but is mathematically and conceptually more elemental than the \mathcal{R} -fold collection of distributions. Analysis in Brenner and Fraser (1980, 1981) and, more discursively, in Fraser (1979) examines this choice of a structural model (Fraser 1968) over the usual transformation-parameter model; it involves a collapsing and simplification.

4. NECESSARY REDUCTION

Consider the analysis of an inference base (M, D) —the model for the investigation and the data from the investigation, as initially assembled.

In the preceding sections, we referred to the model for an investigation as the description of possibilities for the investigation. Now suppose that some categorical information is available as part of the inference base; by categorical, we mean true-or-false, 0-1, information concerning elements of the investigation. Accordingly, then, the strict model does not include “possibilities” that are in fact not possibilities, that are excluded by the categorical information. The initially presented model and data contain elements that do not belong, and are accordingly removed as a *necessary reduction* of the initial inference base. The operating scientific view, or principle, is not to treat a known as an unknown, that is, not to misrepresent.

There are at least three clear directions in which the preceding reductions can occur. For detailed discussion see Fraser (1979).

For a first direction, let θ in Ω be a parameter representing characteristics of the investigated system. An observed response y , together with the model, may imply categorically the fact $\theta \in D(y)$, a proper subset of Ω . Then, as the model for the investigation delineates possibilities for the investigation, it follows that the range for θ is correspondingly limited: to use the full space Ω is to treat a known as an unknown is to *misrepresent*. In certain contexts a technical point can arise if the class $\{D(y)\}$ is not a partition of Ω , but the details are straightforward. For discussion, including a specific and striking example, see Feuerverger and Fraser (1979).

For a second direction, consider a model containing a component probability space. By the presentation of such a space, we mean there corresponds in the application an objective random component, a subsystem whose random characteristics are fully known, apart from any parameter or unknowns of the investigation. Also, suppose the observed response provides an observed value for this random component. The observed value from the “probability space” then gives conditional probabilities for the remaining variables in the model. The conditional probabilities then provide the *description of possibilities* for the investigation, and thus form the model for the investigation. This conditioning is *not* based on an ancillary argument but directly on the modelling process itself. The same scientific view is operating here: do not treat a known as an unknown; do not present a whole range of possibilities for a particular variable when, in fact, the range has been precisely restricted. For details, see Fraser (1979, Section 3-2).

The third direction relates to the special situation mentioned at the end of the preceding section—there is a basic distribution for form, and the response is a presentation of it. An observed value for the response then delineates, by means of the presentation, a range of values for the latent variable for distribution form. There is thus an observed value for a *function* of that variable, and nothing further concerning the original value. In a manner similar to that just described, we find that the strict model for the investigation involves the conditional distributions. More generally, we can have situations that involve an additional parameter for distribution form. The model covering the observable variable is then the marginal model for the function just mentioned, and the model for the unobserved portion is the previously mentioned conditional model. For further details, see Fraser (1976b; 1979, Sections 3.3, 2.3).

The preceding three directions all involve the direct use of the categorical infor-

mation contained in the inference base. An overall approach to categorical information would, in particular, embrace the preceding three cases, and, as such, would verify the mutual consistency of the three methods. The details for this are in preliminary form.

There is a further direction in which categorical information is used and that concerns the theoretical assessment of statistical models. For example, it is used in Brenner and Fraser (1980) to provide a classification of the transformation-parameter models in accord with the degree to which conditionality analysis has a categorical basis.

More recently, the assessment of the givens and arguments has resolved the dilemmas with Birnbaum's argument from sufficiency and conditionality to likelihood, cf. Evans and Fraser (1982).

5. SIMPLER MODELS

In Section 3, we discussed aspects of the formation of statistical models, and noted the fundamental requirement that there be a one-one correspondence between elements in the model and physical elements that have been identified in the context under investigation.

We mentioned there that under certain circumstances, it would be possible to model the actual distribution directly — a single distribution in the model corresponding to a single distribution in the physical situation. The fundamental importance of this is simply that it becomes possible to use the actual probabilities of the system for purposes of inference.

The more common modelling in such contexts involves an *infinity* of different distributions as an assembly to describe the investigation. The general requirement of a one-one correspondence would then choose the simpler single-distribution model over the infinitely more complex model as a description of possibilities for the investigation. In a less formalistic way, we could cite Occam's razor.

In this section, we discuss aspects of this special, but rather widely applicable, modelling situation. Specifically, we consider contexts where the general shape or form of the response distribution is known from background information, or known up to a parameter, say λ in Λ , where λ could be quite general.

What are the special circumstances that lead to information that the basic shape or form of the distribution is known? This has been examined in Brenner and Fraser (1980), with details limited to the case of a single distribution form rather than the parametrized case with λ in Λ . The more general case is currently in manuscript form.

Different parts of the response space and difference possibilities for the response distribution are examined by means of a class of *scans* — maps from an examination space to the response space. The identification of distribution form is then, in part, analogous to the situation of having a waveform hold on a CRT display. The paper just cited records assessments in terms of observability, samplability, and compatibility. All three equally support the basic condition on the class of scans: that it be a *platform*, technically defined, for observing a distribution form.

The discussion so far has centered on the modelling process. Certain aspects of the discussion can however be examined as a within-model question. This has fairly fundamental implications, and can be formulated fairly simply: Consider a probability space and some type of information processor that gives information concerning

a realized value on the probability space. What conditions validate the use of conditional probability?

The grounds for conditional probability are not widely examined in the statistical literature. The pattern in most statistical texts is to introduce the definition of the conditional probability $P(A : C)$ and explore a few straightforward examples. In fact, serious pitfalls exist, and seemingly open paradoxes can be found in the literature (Freund 1965). An approach with a detailed evaluation of conditions for, and risks with, conditional probability may be found in Fraser (1976b, Section 4.1).

Now consider a probability space and an information processor. The grounds for conditional probability determine conditions on the information processor; these have been examined in Brenner and Fraser (1979). Some risks that exist when the conditions are not fulfilled may be found in the discussion to McGilchrist (1973). The model mentioned at the beginning of this section is a special case in which the presentation functions form the information processor. The within-model discussion just given provides further support for the conditions that are relevant at the modelling stage itself.

An alternative approach to some of these questions is to take an ordinary statistical model involving a class of densities, and examine it for symmetries using transformations appropriate or relevant to the response space. Such an analysis is given in Brenner and Fraser (1981), and provides additional support for the direct modelling discussed at the beginning of this section.

In conclusion, we note that the results of this section all focus on the importance and relevance of the structural model to a wide class of problems. A survey of the general theory may be found in Fraser (1979).

6. ADDITIVES

An examination of the statistical literature can give a rather disturbing impression—that principles and criteria, such as sufficiency and ancillarity, are added iteratively in *ad hoc* ways to get answers, and that postscript judgements are made, such as “I can get a particular answer; all I have to do is use Macho’s prior.”

Clearly we should question the validity of conclusions achieved by appeal to *additives*, particularly principles and criteria that are invoked primarily to get to an answer. And we should ask what the grounds are for any particular prior, grounds other than to reach an otherwise preselected posterior distribution. If the only goal is to reach the particular posterior, the fastest and most unequivocal route is to start directly with the posterior as the particular *additive*.

Examples such as these may sound extreme, but exist widely, often persuasively presented. They all involve *additives*, say A , beyond the basic inference base (M, D) . In effect, (M, D) is augmented to (M, D, A) .

The role of additives has been discussed briefly in Brenner, Fraser, and Monette (1981). The view is taken there that the additives in any problem should be clearly recognized and formulated. The view is also expressed that the additives should be rephrased or recast in order to be seen as an explicit component of the inference base at the initial formulation stage—as opposed to being introduced during the inference argument.

The deductive inference process, in this more general context, is viewed, then, as finding the implications concerning unknowns that follow deductively from the *extended inference base* (M, D, A) . The reverse concern centers on how much the

