

Some remarks on the nature of statistical inference

D.R. Cox

Nuffield College, Oxford, UK

and

Deborah Mayo

Virginia State University, Blacksburg, Virginia, USA

SUMMARY

After a brief discussion of the role and importance of objectivity in science and statistical analysis a short outline is given of a frequentist basis for statistical interpretation of data. The central role of a probability model of the data-generating process is stressed and the contribution of various kinds of conditioning explained. Some of the complications that can arise in applications are mentioned and a brief critique of alternative Bayesian approaches set out.

1 Preliminaries

Statistical methods are used to some extent in virtually all areas of science, technology, public affairs and private enterprise. The variety of applications makes any single unifying discussion difficult if not impossible. The present paper concentrates on the role of statistics in research in the natural and social sciences and the associated technologies, aiming to give a relatively nontechnical discussion of some of the conceptual issues involved.

The aim in some broad sense is to achieve enhanced understanding of the real world and in this some notion of objectivity is crucial. We therefore first discuss that notion briefly.

2 Objectivity

Objectivity in statistics, as in science more generally, is a matter of both aims and methods. Objective science, in our view, aims to find out what is the case as regards aspects of the world independently of our beliefs, biases, and interests; objective methods thus aim at critical control of inferences and hypotheses, constraining them by evidence and checks of error. The statistician is sometimes regarded as the gatekeeper of objectivity when the aim is to learn about those aspects of the world exhibiting haphazard variability, especially where methods take into account the uncertainties and errors by using probabilistic ideas in one way or another.

In one form, probability arises to quantify the relative frequencies of errors in a hypothetical long-run; in a second context probability purports to quantify the "rational" degree of belief, confirmation, or credibility in hypotheses. In the first "frequentist" camp, the aim of objective learning about the world is framed within a statistical model of the process postulated to have generated data. Frequentist methods achieve an objective connection to hypotheses about the data-generating process by being constrained and calibrated by the method's error probabilities in relation to these models: the probabilities derived from the modeled phenomena are equal to or are close to the actual or hypothetical relative frequencies of results in applying the method. In the second, degree-of-belief theory, by contrast, objectivity is bought by attempting to identify ideally rational degrees of belief controlled by inner coherency.

What are often called "objective" Bayesian methods fall under this second banner, and many although not all current Bayesian approaches appear to favour the use of special prior probabilities, representing in some sense an indifferent or neutral attitude (Berger, 2004). This is both because of the difficulty of eliciting subjective priors, and because of the reluctance among scientists to allow subjective beliefs to be conflated with the information provided by data. However, acknowledging that non informative priors do not exist, the "objective" (or default) priors are regarded largely as conventionally stipulated reference points to serve as weights in a Bayesian computation. We return to this issue later.

3 Formulation

An already idealized formulation is to recognize the following steps

- one or more research questions are formulated, very often on the basis of previous research
- data to address the question are either found or collected. A major Chapter in statistical theory addresses the design of experiments and observational studies, aiming to achieve unambiguous conclusions of as high a precision as is required. We shall not discuss this
- preliminary checks of data quality and simple graphical and tabular displays of the data are made. Sometimes, especially with very skilful design, little more analysis may be needed
- we consider, however, cases where more formal analysis is needed, both to extract as much information as possible from the data about the

research questions of concern and to assess the security of any interpretation reached.

4 Formal analysis: the start

The formal analysis proceeds broadly as follow.

First we divide the features to be analysed into two parts, and denote their full set of values collectively by y and by x , typically multi-dimensional. A probability model is formulated according to which y is the observed value of a vector random variable Y whose distribution depends on x , regarded as fixed. Note that especially in observational studies it may be that x could have been regarded as random but we choose not to do so.

Example 1. For a random sample of men we measure systolic blood pressure, weight, height and age. The research question may concern the relation between sbp and weight allowing for height and age and if so, specifically for that question, one would condition on the last three variables and represent by a model the conditional distribution of Y , sbp, given x , the other three variables. One would do this even if, say, one knew the distribution of age in the population.

We may call this *conditioning by model formulation*.

We consider here parametric models in which the probability density of Y is taken in the form $f_Y(y; \theta)$, where θ is typically a vector of parameters $\theta = (\psi, \lambda)$, where the parameter of interest ψ addresses the research question of concern and the nuisance parameter λ is needed to complete the specification. Virtually all such models are to some extent provisional. Dependence on x is not shown explicitly in this notation.

Crucial conceptual issues concern the nature of the probability model and in particular the role of probability in it.

Probability models range from purely empirical representations of the pattern of haphazard variability observed to ones that include substantial elements of the underlying science base.

The former get their importance as providing a framework for statistical methods, for example some form of regression analysis, that find fruitful application across many fields of study. The latter provide a stronger link with interpretation. An intermediate class of models of increasing importance in observational studies, especially in the social sciences, represent a potential data generating process and hence may give a pointer towards a causal interpretation. Parameters, especially parameters of interest, are intended to encapsulate important aspects of the data generating process separated off from the accidents of the specific data under analysis. Probability is to be regarded as directly or indirectly based on the empirical stability of frequencies under real or hypothetical repetition.

Example 2. Cox and Brandwood(1959) put in order, possibly of time, the works of Plato taking *Laws* and *Republic* as reference points. The data were the stresses on the last 5 syllables of each sentence and these were assumed to have probability distributions over the 32 possibilities. What can probability mean in such a case?

Two challenges now arise. How do we use the data as effectively as possible to learn about ψ ? How can we check on the appropriateness of the model?

5 An initial solution

We aim to make a minimal choice of $s = s(y)$ such that the distribution in the model essentially factorizes as

$$f_S(s; \theta) f_{Y|S}(y; s),$$

that is, such that the distribution of Y given the value of S does not depend on the unknown θ . We now argue as follows. It would be equivalent to be given the data in two stages:

- we are told the value of s
- then we learn the value of the remaining parts of the data.

Now so long as the model is appropriate the second phase is equivalent to a random draw from a *totally known* distribution, and could just as well be the outcome of a random number generator. Therefore all the information about θ is, so long as the model is appropriate, locked in s . Secondly, in so far in some relevant respect the remaining parts of the data are not concordant with being from the known distribution, doubt is thrown on the model. It is crucial that any account of statistical inference provides a conceptual framework for this process of model criticism, even if in practice the criticism is often done relatively informally.

How then are we to extract answers to the research question out of $f_S(s; \theta)$; all that the reduction to s has done is to reduce the dimensionality of the data.

6 Formal analysis: the next step

There are now three, or perhaps four, approaches that might be taken and controversies about them have rumbled on for more than 200 years. The following discussion is merely an outline of a few of the issues involved.

The fourth approach is to appeal to what Hacking called the law of the likelihood. This amounts to looking at ratios $f_S(s; \theta_1)/f_S(s; \theta_0)$ of likelihoods for different parameter values, the ratios depending only on s , and to regard these as summaries of what the data convey about θ . There are various

reasons why this is inadequate except for very simple problems, just one being the difficulty of dealing with nuisance parameters.

The other three methods are

- the use of probability as representing personalistic degree of belief
- the use of probability as representing impersonal or rational degree of belief
- an indirect use of the frequentist view of probability to characterize methods of analysis

We concentrate here on the third. The personalistic approach, whatever merits it may have as a representation of personal belief and personal decision making, is in our view inappropriate for the public communication of information that is the core of scientific research, and other areas too. The objectivist Bayesian view addresses the same issues as the frequentist approach; some of the reasons for preferring the frequentist approach are sketched in Section .

The central point is that the focus of interest ψ is typically an unknown constant and if we were to aim at talking about a probability distribution for ψ an extended notion of probability would then be unavoidable.

With such a generalized notion of probability it is possible, formally at least, to assign a probability distribution to ψ given the data and to assess the probability that some hypothesis about ψ is in some sense true. This is done by what used to be called inverse probability and is nowadays referred to as a Bayesian argument.

There are two main formulations used in the frequentist approach to the summarization of evidence about the parameter of interest ψ . We shall for

simplicity suppose from now on that for each research question of interest ψ is one dimensional.

The first is the provision of sets or intervals within which ψ is in some sense likely to lie (confidence intervals) and the other is the assessment of concordancy with a specified value ψ_0 . It turns out to be a good idea to concentrate on the latter.

That is, we address the question of assessing the concordancy of the data with ψ_0 , conveniently called the null hypothesis and denoted by H_0 .

7 The significance test

How would we test such a hypothesis if it was deterministic? We would find one (or more) relevant observable features exactly predicted by the hypothesis and then check whether it (or they) are as predicted.

So far as feasible we do the same in the stochastic situation. That is, we look for a feature t of the data, in the light of the previous discussion a function of s , such that

- the probability distribution of the random variable T is exactly known when the null hypothesis is true, so that in particular the distribution does not depend on nuisance parameters
- the larger the value of t the greater the discrepancy with the null hypothesis

In the second phase of the interpretation we compare the observed value t with its predicted value under the null hypothesis by finding for any observed value t

$$p = P(T \geq t; \psi = \psi_0).$$

We look to see how extreme t is in its probability distribution under H_0 . An extremely small value of p corresponds to an extreme discordancy with the null hypothesis. In applications it is usually enough to report p approximately. A hypothetical interpretation of a given p is as follows: if the data under analysis were regarded as just decisive against the null hypothesis, then of a long run of cases in which the null hypothesis is true it would be wrongly rejected in a proportion p .

Note that this is a theoretical calibration of a significance test regarded as a measuring instrument for concordancy. It is not an instruction on how to use the procedure for accepting or rejecting the null hypothesis. Rather, in the first place at least, we regard p as a measure of concordancy with the null hypothesis, calibrated as are other measuring instruments by performance on (hypothetical) use. In addition, however, this kind of hypothetical reasoning is relevant to the case at hand not solely or primarily in view of the long run rates, but rather in what those rates reveal about the data-generating source of phenomenon. The error-based calculations reassure us that incorrect interpretations are being avoided in the particular case.

8 A typology of null hypotheses

The formulation just sketched is sufficiently rich to cover a number of distinct cases confusion amongst which is one reason why discussion of the whole subject of significance tests contains so many misunderstandings.

Some distinct possibilities are as follows:

- there may be a full family of probability models embracing both the null hypothesis and a set of other possibilities for comparison
- the null hypothesis may be clearly defined but alternatives specified

only somewhat qualitatively, for example by supposing that departures from the null hypothesis of interest are shown by certain features of the data being extreme

In the latter case, but not the former, choice of the test statistic t comes from outside the model, that is from *qualitative* consideration of the alternatives of interest.

Some but not all of the following possibilities apply only to the full parametric formulation, that is to the former of the above two cases.

- the null hypothesis $\psi = \psi_0$ may plausibly be exactly (or very nearly) true and subject-matter interest focuses on whether there is evidence against or, sometimes, in favour of it
- we may have a dividing null hypothesis in that the value ψ_0 divides the parameter space into two qualitative different parts. Concordancy with the null hypothesis means that the data are indecisive as to which part holds. There may be no special reason for thinking the null hypothesis itself to be intrinsically plausible
- a powerful method of summarizing information about ψ in general is to test concordancy with every possible value of ψ , taken in turn, and to assemble all those values of ψ concordant with the data at various levels of p , forming what are typically systems of confidence intervals

An interpretation of extreme discordancy with H_0 is that H_0 is untenable in the light of the data, or that the data are seriously defective. An interpretation of a moderately large value of p , for example $p > 0.2$, say is initially that the data are concordant with H_0 in the respect tested. To go further and claim that such a set of data were positive support for H_0 would

require showing that the data are discordant with the kind of departure from H_0 of subject-matter concern. The difficulty with treating a modest value of p as evidence in favour of H_0 is that apparent agreement between the data and H_0 can occur even if rivals to H_0 seriously different from it are true. This issue is particularly acute when data are limited. However sometimes we may be able to use tests that would with high probability have reported such a discrepancy were it present and then absence of a discrepancy would be evidence for H_0 .

9 Conditioning and the definition of p

The definition of p refers to behaviour in a set of hypothetical repetitions.

To establish a significance test we need to choose the statistic t and find a distribution for assessing its concordancy with H_0 . The approach to this depends somewhat on the type of null hypothesis involved. We deal here with an important situation where an essentially unique answer is possible.

Suppose first that there is a full model covering both null and alternative possibilities and that ψ is one-dimensional. We reduce by sufficiency to s .

If s itself is one-dimensional the test statistic must be a function of s and we can almost always arrange that s itself can be taken as the test statistic and its distribution thus found.

If s is multidimensional this argument will not work, We seek a factorization $s = (t, a)$, where t is one-dimensional and we can write

$$f_S(s; \psi) = f_A(a)f_{S|A}(s; a, \psi),$$

where the first factor is independent of θ .

Now we argue that it is equivalent to obtain the data in two steps

- we observe that $A = a$. This tells us nothing directly about ψ , although it may and indeed in general will say something about the amount of information actually achieved
- then we observe, conditionally on the first step, that $T = t$, an observation from the conditional distribution $f_{T|A}$

. The second step defines a unique p . We may call this *technical conditioning to ensure relevance*.

Example 3. Suppose n independent and identically distributed observations are each such as to be equally likely to be $\psi - 1$ or $\psi + 1$. The data take one of two possible configurations; either all the values are the same, y' , say, or there are two different values $y'' - 1, y'' + 1$, say both represented. The minimal sufficient statistic is the relevant y and an indicator as to which configuration obtains; the latter has a fixed distribution and hence one conditions on its value. The conclusion is thus that in the second case ψ is exactly known to be y'' whereas in the former two values $y' \pm 1$ are equally concordant with the data.

We may call this *technical conditioning to induce relevance*. The point essentially is that the marginal distribution of an estimate averaged over the possible configurations would be somewhat irrelevant for a particular set of data; in some other cases the base for conditioning may not be so obvious so that there is a need for a systematic formulation.

More commonly there is a nuisance parameter λ in addition to the parameter ψ of interest. In this case we aim to achieve a factorization $s = (t, c)$, where t is one-dimensional and C has a distribution depending only on λ and the distribution of T given $C = c$ depends only on ψ . A variant of the argument used twice above now shows that values of p for a hypothesis about ψ are to be calculated from this last conditional distribution. This

may be called *technical conditioning for separation from nuisance parameters*. In the Neyman-Pearson theory this is called the formation of regions of Neyman structure.

Example 4. Suppose that Y_1 and Y_2 have independent Poisson distributions of means respectively λ and $\psi\lambda$; that is, the ratio of the means is the parameter of interest. For any given value of ψ it can be shown that there is a sufficiency reduction to $c = y_1 + y_2$. That is, for any given value of ψ , the observed value c contains all the information about λ and there is a factorization into information about λ and the complementary term of the distribution of, say, Y_2 given $C = c$ which depends only on ψ and thus contains all the information about ψ so long as λ is regarded as completely unknown.

10 Some limitations

The constructions sketched above successfully underpin a substantial part of what may be called elementary statistical method, including key problems about binomial, Poisson and normal distributions, including the method of least squares for so-called linear models. When we go to more complicated situations the factorizations that underlie the arguments no longer hold.

In some generality, however, we may show that they hold approximately and we may use that to obtain p -values whose interpretation is, for example, only very mildly dependent on the values of nuisance parameters.

The corresponding Bayesian treatment does not involve mathematical approximations but does depend relatively critically on a precise formulation of the prior distribution.

11 Adaptation of analysis to the data

In the idealized model of the analysis of data summarized above, we start with a research question, find or collect relevant data, formulate an appropriate model of the data-generating process, and then, typically at or before the initial stages of data collection, determine the appropriate methods of analysis and finally proceed to the detailed analysis and interpretation. This is essentially the process considered in the formal theory of statistical analysis.

Now it would be foolish in the extreme to collect data without considering how to analyze it. Moreover it would be typical to consider not only what patterns of outcome are in some sense likely to arise but also other possibilities to ensure, if possible, that interpretation will be achievable in all reasonably predictable circumstances. The simple sequence from question, to data, to analysis, to interpretation is implicit in the description of p -values sketched above.

Nevertheless harsh reality may show that the analysis initially planned is not appropriate, perhaps in minor respects of model formulation not affecting the primary research question, perhaps in changing the precise formulation of that question or, rather less commonly, in altering the whole focus of the investigation. The question then arises of how the interpretation of p -values and associated confidence intervals is affected by such changes in the implicit protocol for analysis.

In some cases the effect of these changes on interpretation is minimal. In others an allowance, for example for multiple testing is possible, in the light of a careful specification of the procedure of analysis actually employed. The most challenging possibility is where the whole focus of the investigation shifts to study an effect seemingly totally unanticipated. In situations where investigations can be repeated fairly quickly, the conventional view is,

of course, that in such situations independent replication is obligatory. In other contexts, standard analysis, frequentist or Bayesian can be made but has to be interpreted very cautiously. In a Bayesian analysis particular warning is needed against explanations that beforehand were ignored, and hence given an implicitly negligible prior probability, but which with the benefit of hindsight seem to have deserved high prior probability.

12 A review

The essence of the previous discussion is that once a potential probability model has been chosen, a crucial step, a series of factorizations allows us

- to separate a term containing all the information in the data about the model from another term allowing model criticism
- to assess concordancy with a specified value of a parameter of interest, the assessment being calibrate by repeat performance in hypothetical repetitions
- to establish the most relevant ensemble of repetitions for the unique data under analysis
- to deal with situations formulated to various levels of detail

The broad approach is also a powerful one for assessing the merits of alternative schemes of study design and analysis, an aspect that is not addressed here.

There are two further possible approaches to these issues. One involves a notion of probability as a personalistic degree of belief. It allows the incorporation of evidence other than such as can be modelled via a frequency

concept of probability but, by its very nature, is not focused on the extraction and presentation of evidence of a public and objective kind. The other approach based on a notion of rational degree of belief has in some respects similar objectives to the frequentist view sketched above, and often leads to formally very similar or even numerically identical answers. The most elaborately developed versions of this are the reference priors of Bernardo (2005). There are, however, substantial difficulties over the meaning to be attached both to the probabilities used to specify an initial state of knowledge and hence also to the final, or posterior, probabilities.

These are in outline as follows.

- the most fully developed such notion of probability as objective degree of belief is based on that of a reference prior chosen to maximize the contribution of the data to the resulting inference; for a thorough account, see Bernardo (2005). It is emphasized that the priors are not probabilities but merely a basis for comparison. What then is the interpretation of a posterior? Why is that a probability in any meaningful sense?
- not only may the calculation of a reference prior be relatively complicated but the prior for a particular parameter may depend on whether it is a parameter of interest or a nuisance parameter and even on the order in which nuisance parameters are specified
- many of the formal advantages claimed for the Bayesian approach, such as independence from stopping rules, no longer hold
- the underlying supposition that uncertainty about an event or proposition can always be encapsulated in a single real number may be useful for some purposes but in general is suspect. For example, it conflates

two situations, one based on a large amount of high quality data, the other on the merging of meagre data with a formalized prior, both leading to the same numerical probability.

There is not the apace to discuss here the personalistic Bayesian approach to these issues which we regard as dealing with distinct issues concerned with the beliefs of specific individuals. Its main appeal in statistical work is some mixture of internal formal concordancy with the apparent ability to incorporate information that is of a broader kind than that represented by a probabilistic model based on frequencies. The essential focus is too far from our concern with objectivity for this to be a satisfactory basis for statistical analysis in science.

For a general account of these issues, very much from a statistical perspective, see Cox (2006).

REFERENCES

- Berger, J. (2004). The case for objective Bayesian analysis. *Bayesian analysis* **1**, 1-17.
- Bernardo, J.M. (2005). Reference analysis. *Handbook of statistics* **35**. Amsterdam: Elsevier.
- Cox, D.R. (2006). *Principles of statistical inference*. Cambridge University Press.
- Cox, D.R. and Brandwood, L. (1959). On a discriminatory problem connected with the works of Plato. *J.R. Statist. Soc. B* **21**, 195-200