

FREQUENTIST AND BAYESIAN STATISTICS: A CRITIQUE (KEYNOTE ADDRESS)

D.R. COX

Nuffield College, Oxford OX1 1NF, UK

E-mail: david.cox@nuf.ox.ac.uk

The broad distinctions between the frequentist and Bayesian approaches to statistical inference are outlined and some brief historical background given. The advantages and disadvantages of the frequentist discussion are sketched and then two contrasting Bayesian views given. The difficulties with the notion of a flat or uninformative prior distribution are discussed.

1. Introduction

There are two broad approaches to formal statistical inference taken as concerned with the development of methods for analysing noisy empirical data and in particular as the attaching of measures of uncertainty to conclusions. The object of this paper is to summarize what is involved.

The issue is this. We have data represented collectively by y and taken to be the observed value of a vector random variable Y having a distribution determined by unknown parameters $\theta = (\psi, \lambda)$. Here ψ is a parameter of interest, often corresponding to a signal whereas λ represents such features as aspects of the data-capture procedure, background noise and so on. In this, probability is an (idealized) representation of the stability of long-run frequencies, whereas ψ aims to encapsulate important underlying physical parameters that are free from the accidents of the specific data under analysis.

How should we estimate ψ and how should we express our uncertainties about ψ ?

In the following discussion we assume that the probability model correctly represents the underlying physics. This means that issues of model criticism and possible model reformulation that arise in many other applications of statistical methods can be disregarded.

2. Two Broad Avenues

There are two broad routes to an answer, both with variants.

In the first, the so-called *frequentist* approach, we continue to use probability as representing a long-run frequency. Because ψ is typically an unknown *constant*, it is not in this setting meaningful to consider a probability distribution for ψ . Instead we mea-

sure uncertainty via procedures such as confidence limits and significance levels (p -values), whose behaviour is calibrated by their appealing properties under hypothetical repetition. In that the procedure is calibrated by what happens when it is used, it is no different from other measuring devices.

In the second approach, we do aim to attach a probability distribution to the unknown ψ . For this it is essential to extend or change the notion of probability so that it is concerned with uncertainty of knowledge rather than with variability of outcome. Such an approach involves what used to be termed one of inverse probability; it is now generally termed *Bayesian*.

Note that even in those situations where there is a collection of similar parameters that can be regarded as having a probability distribution in the frequency sense it is virtually always necessary to specify their distribution in terms of hyperparameters and a part of the problem of inference is transferred to that for the hyperparameters.

3. A Simple Preliminary

The essence of the Bayesian argument is as follows. Suppose that the possible sets of data that might arise are $\mathcal{D}_1, \mathcal{D}_2, \dots$ and that the possible explanations are $\mathcal{E}_1, \mathcal{E}_2, \dots$, and that all events listed have meaningful probabilities. Then

$$\begin{aligned} P(\mathcal{E}_k | \mathcal{D}_j) &= P(\mathcal{E}_k \cap \mathcal{D}_j) / P(\mathcal{D}_j) \\ &= P(\mathcal{D}_j | \mathcal{E}_k) P(\mathcal{E}_k) / P(\mathcal{D}_j) \\ &\propto P(\mathcal{D}_j | \mathcal{E}_k) P(\mathcal{E}_k). \end{aligned}$$

The proportionality is taken as the explanations vary for specified data and the relation has the normalizing constant $1/P(\mathcal{E}_j)$. The words used for the three terms in this last equation, which has the form of an

inversion equation, are respectively posterior probability, likelihood and prior probability.

Essentially the same relation holds for probability distributions and parameters in the form

$$f_{\Theta|Y}(\theta | y) \propto f_{Y|\Theta}(y | \theta) f_{\Theta}(\theta).$$

To obtain the posterior density of the parameter of interest we integrate out with respect to λ .

4. A Brief History

These issues have a long history. Laplace used what are now called Bayesian arguments with a flat prior, whereas Gauss, especially in his work on the optimality properties of the method of least squares, used frequentist concepts. The Irish algebraist Boole strongly criticized flat priors as representations of ignorance or indifference and similar points were made later in the 19th century by Venn. By the end of the 19th century inverse probability was widely regarded as unsatisfactory for inferential purposes.

Pioneering papers on the formulation of statistical inference by the geneticist and statistician R.A.Fisher, especially a major paper in 1922, laid the foundations for a frequentist theory. Some years later Neyman, then in Warsaw, and E.S.Pearson in London began a systematic development designed originally to clarify Fisher's ideas. Later, only partly because of personal friction between Fisher and Neyman, the differences became accentuated and two broad but rather ill-defined schools of frequentist inference can be discerned.

The view that probability is rational degree of belief following on from Laplace was studied in detail in a thesis by the economist John Maynard Keynes. The main work on this theme was done over the late 1920's and 1930's by the geophysicist H. Jeffreys and set out, in particular in a highly influential book *The theory of probability* in 1939. Discussion of how prior distributions might be determined in the absence of evidence have continued, the most notable work being that of J.M.Bernardo. A contrasting view of probability as a degree of belief emphasizes its personalistic character, in particular its link with personal decision making. An early influential contribution was by F.P. Ramsey. Independent major systematic developments were by de Finetti and L.J. Savage.

5. Outline of Frequentist Approach

A summary of the frequentist approach is as follows. In most situations a directly frequency-based concept of probability cannot be applied directly to the unknown of interest, ψ . Instead we introduce measures of security, p -values and confidence limits, whose interpretation is calibrated, as are other measuring instruments, by their properties when used. In this particular context, use is assessed by considering hypothetically how they perform when used repeatedly under the same conditions. The performance may be studied analytically or by computer simulation.

In particular a confidence set specifies all those values of ψ reasonably consistent with the data up to a specified level. In extreme cases, such sets may be the whole space or null, as when the data appear inconsistent with any possible value of ψ .

6. Critique of Frequentist Approach

Major advantages of the approach are that it provides a systematic approach to a wide range of statistical methods and one not requiring additional specification beyond that of the probabilistic representation of the data-generating process. It also gives a route to assessing methods that may have been suggested on relatively informal grounds.

A key problem in principle in frequentist formulations is that of ensuring that the long-run used in calibration is relevant to the analysis of the specific data being analysed. A more immediate issue in applying the ideas is that technically exact solutions are available only for a relatively limited class of situations. Usually, approximations have to be used based on asymptotic analysis and often implemented by computer simulation.

As an example of the last point, suppose that Y has a Poisson distribution with mean $(\gamma + \lambda)t_S$ and that independently Y_B has a Poisson distribution with mean λt_B , correspondingly to observation of first signal plus background and secondly to background alone. Then if interest lies in $\psi = \gamma/\lambda$ exact efficient estimation is possible based on the binomial distribution of Y given $Y + Y_B$ which is a binomial distribution with parameter

$$\frac{t_S(1 + \psi)}{t_S(1 + \psi) + t_B}.$$

But if interest lies in $\psi^* = \gamma$ itself no formally exact solution is available and we have to use an approxi-

mation, typically based on an asymptotic expansion. As with asymptotic expansions in other areas, some care is needed to ensure that the expansions yield good answers in the specific instance.

For example, if the amount of information on background is relatively large, that is the corresponding errors in estimating λ relatively small, the following approximate argument can be used.

For given y , let $p(y, \psi_0; \lambda)$ be the p -value for testing $\psi = \psi_0$, assuming λ is known. Let $\tilde{\lambda}$ be an unbiased estimate for λ with small variance $v(\lambda)$, all conditionally on y . Then a close approximation to the significance level adjusted for errors of estimation of λ is $p(y, \psi_0; \lambda^*)$, where

$$\lambda^* = \tilde{\lambda} - \frac{v(\tilde{\lambda}_0)\partial^2 p/\partial\lambda^2}{2\partial p/\partial\lambda}.$$

The final term has a direct generalization if λ is a vector and may be evaluated at $\lambda = \tilde{\lambda}$.

In particular for the above application with $y = 0$, the p -value for testing $\psi = \psi_0$ leading to an upper confidence limit for ψ is

$$\exp\{-\psi_0 - y_B/t_B + y_B/(2t_B^2)\}.$$

7. Critique of Bayesian Methods

To use Bayesian methods we have to extend the notion of probability so that we can specify a prior distribution for the unknown constant θ . That is we regard probability as measuring a degree of belief in an uncertain event or proposition. There are two radically different ways of doing this.

The first approach is personalistic in which $P(\mathcal{E}|\mathcal{I})$ denotes the degree of belief in \mathcal{E} held by a specific individual, conventionally denoted by You, given information \mathcal{I} . There is no suggestion that two different people with the same background information will have the same probability. The emphasis is on trying to achieve self-consistency, so-called coherency, in Your probability assessments. The second approach involves a notion of rational degree of belief and, commonly although not necessarily, an attempt, following Laplace, to address the question of assessing the evidence in a specific set of data by using a prior expressing a notion of indifference or ignorance in order to focus attention on the data.

These are to be regarded as two very different approaches and the following comments address them separately.

8. Personalistic Theory

This approach has the ambitious aim, in particular, of introducing into the quantitative discussion uncertain information of a more general kind than is represented by statistical data in the narrow sense. In theoretical discussion it is usually set out as part of a theory of personal decision making. Suppose, in order to simplify the discussion, that there is available a source able to produce events with any specified probability p . Then Your probability of \mathcal{E} is a value of p such that you are indifferent as between

- a valuable prize if \mathcal{E} is true and zero if \mathcal{E} is false
- the same prize if an event with agreed probability p occurs and zero otherwise

A certain kind of consistency of behaviour can be shown to imply that the laws of probability theory hold. Note though that this is not a theory of empirical behaviour based on what people actually do but rather a specification of how they would have to behave to be self-consistent.

A very major difficulty with this as a basis for the public discussion of scientific evidence is that it treats personal intuition as on the same basis as evidence from hard data. More explicitly it treats all probabilities of, say, 0.5 as on an equal footing, whether they are based on careful stable measurements of frequency or on the most transitory of personal judgements. In some situations prior distributions based on a careful summary of expert judgement may be used quantitatively, but then scrutiny of their evidence-base is crucial.

This is not to deny the relevance of personal judgement for the individual decision-maker.

9. Probability as Rational Degree of Belief

While the notion of rational degree of belief can certainly be taken more broadly, for the most part it is associated with the use of priors that are in some sense flat, which aim to represent little or no prior information and which therefore induce posterior distributions having the same objective as frequentist methods, i.e. of summarizing what it is reasonable to learn from data plus assumptions about the structure of the data-generating process.

It is generally accepted from various philosophical standpoints that the notion of representing ignorance as such by a flat prior is treacherous, although in some fields the use of relatively flat priors as non-committal is quite widespread. The following points arise

- if θ has a flat, i.e. effectively uniform, prior then e^θ has an exponential distribution, so that choice of functional form of parameters would be important
- for a one dimensional parameter the Jeffreys prior, essentially uniform in a parameterization for which the Fisher information is constant, leads to a posterior distribution having very good frequentist properties
- flat priors for parameters with a large number of dimensions may give clearly unacceptable answers.

J.M. Bernardo has developed a systematic theory of reference priors. This is based on the notion of finding a prior weighting function that maximizes the expected discrepancy between prior knowledge and perfect knowledge obtained by a specified type of replication of the system. When the parameter space is finite it produces the maximum entropy prior of E.T. Jaynes and for a one-dimensional parameter the Jeffreys prior. Some difficulties are that when there are nuisance parameters

- finding the prior weight is often complicated
- the nuisance parameters have to be arranged in sequence of importance, even though none of them is of intrinsic interest
- if the parameter of interest changes the whole prior structure may change
- if the sampling rule or design changes the prior will in general change
- it is emphasized that the prior weights are not to be thought of as prior probabilities, raising a question-mark over the interpretation of the posterior
- many of the formal simplifications arising from all calculations being probabilistic are lost.

In general reference priors have some good frequentist properties but except in one-dimensional problems it is unclear that they have any special merit in that regard.

10. Concluding Remarks

In conclusion, the following points arise:

- formal inferential aspects are often a relatively small part of statistical analysis
- carefully used, the frequentist approach yields broadly applicable if sometimes clumsy answers
- in simple problems specially chosen prior distributions yield essentially the same answer
- in multiparameter problems flat priors can yield very bad answers
- injection of further information quantitatively through an informative prior may be helpful but scrutiny of the evidence base is essential.

These issues have a very extensive literature. Traditional accounts of the two frequentist viewpoints are by Fisher¹ and Neyman and Pearson² and of the two Bayesian approaches by Jeffreys³ and Savage⁴. An introductory comparative account is by Barnett⁵ and a systematic discussion by Cox and Hinkley⁶ and Cox⁷. The notion of reference priors is developed in detail by Bernardo⁸.

References

1. Fisher, R.A., *The logic of scientific inference*. Edinburgh: Oliver and Boyd (1956).
2. Neyman, J. and Pearson, E.S., *Joint statistical papers of J.Neyman and E.S.Pearson*. Cambridge University Press on behalf of Biometrika Trustees (1967).
3. Jeffreys, H., (1939 and subsequent editions). *The theory of probability*. Oxford University Press (1939).
4. Savage, L.J., *Foundations of statistics*. New York: Wiley (1964).
5. Barnett, V.D., *Comparative statistical inference*. 3rd edition. Chichester: Wiley (1999).
6. Cox, D.R. and Hinkley, D.V., *Theoretical statistics*. London: Chapman and Hall (1974).
7. Cox, D.R., *Principles of statistical inference*. To appear (2006).
8. Bernardo, J.M., Reference priors. To appear (2006).