Bayesian and frequentist approaches

G. Jogesh Babu*
Center for Astrostatistics, 326 Thomas Building,
The Pennsylvania State University, University Park, PA, 16802

Abstract

Though the Bayesian *vs.* frequentist debate is largely muted in the statistics community, it continues among those applying advanced statistical methods in the physical sciences. The debate often misses some salient features on both sides, relying on simplistic arguments based on some single aspect of the methodology such as p-values or priors. The discussion here argues that each approach has something to contribute. It is always good statistical practice to analyze the data by several methods and compare results.

1 Introduction

All models are wrong. But some are useful

– George E. P. Box[1]

The debate over Bayesian *vs.* frequentist statistical inference is largely over in the statistics community. Both Bayesian and frequentist ideas have a lot to offer practitioners. Each approach has a great deal to contribute to statistical analysis and each is essential for the full development of the other approach. Both methods often lead to the same solution when no external information (other than the data and the model itself) is introduced into the analysis. But these methods are not the same: they do different things. It

*This work is supported in part by NSF Grant AST-1047586

[1]Box is son-in-law of Sir Ronald Fisher, and is well known for his contributions to time series analysis, design of experiments and Bayesian inference

is very important to understand the assumptions behind the theories and to correctly interpret the mathematical conclusions produced. Using both approaches for an important problem is good practice. The union of frequentist and Bayesian procedures is discussed by Bayarri and Berger (2004), and this paper is based in part on their article.

In general, frequentist methods are computationally relatively simple. There is no need for numerical integration. Many of these methods, for sufficiently large data sets, are the locally most powerful tests possible. In many cases the frequentist and Bayesian interpretations are different: Bayesian methods are based on *decision theoretic principles*; actions are dictated by risk management by minimizing the expected loss under a chosen 'loss' function. Similar choices are needed in frequentist methodology to determine the optimal procedure (*e.g.* least squares or maximum likelihood estimation).

Frequentist doctrine is based on the idea that, with repeated use of a statistical procedure, the actual accuracy of the long-run average should not be less than the reported accuracy of the long-run average. This is really a joint frequentist-Bayesian principle. Consider the case of a 95% classical confidence interval for a parameter such as the 95% confidence interval for a Gaussian mean μ. The underlying presumption is that we repeatedly use this statistical model (a Gaussian distribution with the same mean) and the chosen statistical procedure. However, in practice the confidence procedure is used on a series of different problems with, possibly, different Gaussian means μ and different sets of data. Thus, in evaluating the procedure, we should be simultaneously averaging over the different means and data sets. This is in contrast to the common understanding of the frequentist principle which tends to focus on fixing the Gaussian model, using the confidence procedure repeatedly on data sets drawn repeatedly from a single model. This makes no sense in practice, as pointed out repeatedly by the distinguished mid-20th century statistician Jerzy Neyman. Conceptually, it is the combined frequentist-Bayesian average that is of practically importance.

2 Testing of Hypothesis

In the classical testing of hypothesis we have two competing hypotheses, the null hypothesis $\mathbf{H}_0 : \theta \in \Theta_0$ *vs.* a specified alternative hypothesis $\mathbf{H}_1 : \theta \in \Theta_1$. The classical testing of hypothesis tries to control two types of potential errors.

Type I error, also known as an error of the first kind, is the wrong inference made when a test rejects a true null hypothesis. In signal detection, a Type I error is a false positive. Type I errors are philosophically a focus of skepticism and Occam's razor. A Type I error is committed when we fail to believe a truth.

Type II error, also known as an error of the second kind, is the wrong decision that is made when a test fails to reject a false null hypothesis. A Type II error is often called a false negative where an actual 'hit' was disregarded by the test and seen as a 'miss' when checking for a single condition with a definitive result of true or false. The error is made when a test fails to reject a false null hypothesis, when we believe there is no signal when in actuality there is a signal.

It is not possible to minimize both the errors simultaneously. The Neyman-Pearson theory that underlies the testing of hypotheses methodology, controls Type I error at a certain *significance* level and minimize Type II errors to maximize the *power* of the test. The power function $\beta(\theta)$ at a given parameter value θ of the alternative hypothesis is the probability of correctly rejecting the null hypothesis, when the true parameter is θ.

In classical testing, the null and alternative hypotheses are not treated symmetrically while they are in Bayesian analysis. In the latter approach, the **Bayes factor** B in favor of Θ_0 is given by

$$B = \frac{\text{Posterior odds ratio}}{\text{Prior odds ratio}} = \frac{\alpha_0 \pi_1}{\alpha_1 \pi_0},$$

where $\alpha_i = P(\Theta_i|x)$, and π_i are posterior and prior probabilities respectively, $i = 0, 1$. The Bayesian analysis treats both hypotheses symmetrically, and chooses H_0 over H_1 if B is greater than 1. However, the procedure does not provide error probabilities. In a simple *vs.* simple hypothesis testing, the Bayes factor is simply the likelihood ratio, $f(x|\theta_0)/f(x|\theta_1)$, where f denotes the density.

There is a close resemblance between the classical rejection region and the rejection region of Bayesian test by decision theoretic approach under binary $(0-1)$ loss. It is given by

$$C = \left\{ x : P(\Theta_1|x) > \frac{K_1}{K_0 + K_1} \right\}$$

Here K_i are cost factors in the loss function.

Harold Jeffreys, a distinguished geophysicist and Bayesian proponent, proposed use of objective posterior probabilities of hypotheses. Jerzy Neyman recommended testing with fixed error probabilities. Ronald Fisher advocated testing using p-values, the probability of observing an X *more extreme* than the actual data x. Each was quite critical of the other approaches.

What is worrisome is that the three methods can lead to altogether different practical conclusions (Berger, 2003). The criticisms by Fisher and Neyman against the Bayesian approach were quite general. They felt that it is difficult to choose a prior distribution for Bayesian testing. It is framed in the jargon of objectivity *vs.* subjectivity, and sometimes formulated in terms of a preference for the frequency interpretation of probability. Both Fisher and Jeffreys criticized the errors in classical Neyman-Pearson testing of hypotheses for not reflecting the variation in evidence as the data range over the rejection regions. Fisher also criticized frequentist testing because of its need for a fully specified alternative hypothesis and for the associated difficulty of having to deal with a power function depending on unknown parameters.

2.1 p-value against H_0

One of the major criticism of frequentist hypothesis testing is the use of p-values. Recall that the p-value against H_0, when $\theta = \theta_0$, is the probability of observing an X more extreme than the actual data x. The p-value is thus strongly biased *against* the theory we are testing. The concept of p-value is not part of the Neyman-Pearson theory of testing of hypothesis and p-values do not have a frequentist justification. They are problematic in the view of many statisticians. It is not a posterior probability of a hypothesis. For a vast majority of cases, a p-value of 0.05 means that one can be pretty sure that H_0 is wrong. However, any reasonably fair Bayesian analysis will show that there is at best very weak evidence against H_0. For Fisher, a p-value of 10^{-20} seems to reject much more strongly than a p-value of 10^{-2}, leading him to push for the use of p-values. Contrary to popular perception, it was Karl Pearson, and not Fisher, who introduced the p-value (Inman, 1994).

In Neyman-Pearson theory, the p-value itself does not matter, it only matters that $p < \alpha$, the significance level. That is, Neyman-Pearson testing theory reports the same error probability regardless of the size of the test statistic. A p-value of 1% does not state that the probability of the hypothesis H_0 is

1%! It can be seen using simulations (see http://www.stat.duke.edu/∼berger) that a p-value of 5% can result from data where H_0 is true in 20-50% of cases!

Criticism of p-values ignores the power function, though it has quite a bit of information. As mentioned above, Bayesian analysis can be questioned because of the choice of the prior. A Bayesian has no recourse but to attempt subjective specification of the feature. Too often p-values are misinterpreted as error probabilities, which results in considerable overstatement of the evidence against the null hypothesis. Neyman criticized p-values for violating the frequentist principle. Jeffreys (1961) articulates that "... a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred". Thus he felt that the logic of basing p-values on a tail area, as opposed to the actual data, was silly.

2.2 Conditional Testing

Berger, Brown and Wolpert (1994) proposed a solution for hypothesis testing by defining the Type I and Type II errors conditioned on the observed values of a statistic measuring the strength of evidence in the data. Consider the following simple hypothesis testing problem based on data X.

$$H_0 : f = f_0 \; vs. \; H_1 : f = f_1$$

The conditional Type I and Type II error probabilities based on a selected statistic, a function of the data X denoted $S = S(X)$, that measures the *strength of the evidence* in data X, for or against the hypotheses, are computed as

$$\begin{aligned} \alpha(s) &= P(\text{Type I error}|S = s) \\ &\equiv P_0(\text{reject } H_0|S(X) = s) \\ \gamma(s) &= P(\text{Type II error}|S = s) \\ &\equiv P_1(\text{accept } H_0|S(X) = s), \end{aligned}$$

where P_i refer to probability under H_i. Here S and the associated test utilize the p-values to measure the strength of the evidence in the data. Until this point there is no connection with Bayesianism. Conditioning is completely accepted and encouraged within the frequentist paradigm.

Suppose the hypotheses have equal prior probabilities of $1/2$. The Bayesian connection arises as it is known that

$$\alpha(s) = \frac{B(s)}{1 + B(s)} \quad \text{and} \quad \gamma(s) = \frac{1}{1 + B(s)}, \tag{1}$$

where $B(s) = \alpha(s)/\gamma(s)$ is the Bayes factor (or likelihood ratio in this simple setting).

The expressions (1) are precisely the Bayesian posterior probabilities of H_0 and H_1. A conditional frequentist can simply compute the objective Bayesian posterior probabilities of the hypotheses, and declare that they are the conditional frequentist error probabilities. There is no need to formally derive the conditioning statistic or perform the conditional frequentist computations. Thus the resulting conditional frequentist error probabilities equal the objective posterior probabilities of the hypotheses advocated by Jeffreys. This represents a profound alliance between the frequentist and Bayesian approaches.

3 Frequentist procedures in Bayesian methodology

Sometimes Bayesian methods depend on frequentist procedures. Few are discussed briefly below.

Bayesian Computation: *Gibbs sampling* and other *Markov chain Monte Carlo* (MCMC) methods have become relatively standard to deal with hierarchical, multilevel or mixed models. They are also used in estimating the posterior distribution and the mean. These methods are commonly known as Bayesian computation methods. They have become very popular recently, and Bayesian computation is now easier than computation via more classical routes for many complex problems. It is not necessarily due to their intrinsic virtues. However, frequentist reasoning is at the heart of any MCMC method. Diagnostics for MCMC convergence are almost universally based on frequentist tools.

Development of prior distribution: A subjective Bayesian does not bother about frequentist ideas if the chosen prior accurately reflects prior beliefs. However, it is very rare to have a mathematical prior distribution that accurately reflects all prior beliefs. One can utilize frequency methods to develop a prior distribution. Consider the simple case where the only unknown model parameter in the problem under consideration is a Gaussian mean μ. The complete specification of the prior distribution for μ involves an infinite number of conditions, such as assignment of values for the probability

that μ belongs to the interval $(-\infty, r]$ for any rational number r. Note that the probabilities of such half-open intervals uniquely determine the complete probability distribution on the line. In practice, only a few values are ever made, such as choosing prior as a Cauchy density with median and first quartile specified. Clearly, the effect of features of the prior that were not extracted is problematic.

Consistency: Consistency is a statistical concept that the procedure under consideration approximates the truth in the presence of large data. This is one of the simplest frequentist estimation tool that a Bayesian can usefully employ. Bayes estimates are virtually always consistent if the parameter space is finite-dimensional. This need not be true if the parameter space is not finite-dimensional or in irregular cases.

In many applications in fields such as bioinformatics, the number of parameters increases with the amount of data. Consider the following classical *Neyman-Scott Problem* (Neyman and Scott, 1948). Let X_{ij} be independent data drawn from a Gaussian distribution with mean μ_i and common variance σ^2, $i = 1, \ldots, n;\ j = 1, 2$. Thus there are $2n$ data points and $n + 1$ parameters. The Jeffreys-rule prior is the most commonly used objective prior in such cases. It can be shown that the resulting Bayesian estimate of σ^2 approximates to half the value, leading to an inconsistent estimate. Thus the Jeffreys-rule prior is often inappropriate in high dimensional settings, yet it can be difficult or impossible to assess this problem within the Bayesian paradigm itself. Here μ_i are nuisance parameters that are of no interest to the central problem. The only parameter of interest is σ^2. The differences $X_{i1} - X_{i2}$ are distributed as independent Gaussian with mean zero and variance $2\sigma^2$. Now one can use the classical laws of large numbers and show that $\frac{1}{2n}\sum_{i=1}^{n}(X_{i1} - X_{i2})^2$ is a consistent estimator of σ^2.

Estimation: In standard parametric estimation problems, objective Bayesian and frequentist methods often provide similar or even identical answers. For the standard Gaussian linear model, frequentist estimates and confidence intervals coincide exactly with the standard objective Bayesian estimates and credible intervals. This occurs more generally in the presence of the so-called "invariance structure". The 'pivotal' structure lies at the heart of classical confidence interval construction using inversion. That is, once the estimator is 'standardized' by hypothetical subtraction of its theoretical mean and division by the square-root of the variance, the distribution is free

from any unknown parameters. It is also possible to achieve near-agreement between frequentist and Bayesian estimation procedures in more complicated problems.

4 Conclusions

Both Bayesian and frequentist methodology are here to stay. Neither will disappear in the foreseeable future. One should not conclude that all Bayesian or all frequentist methodology is fine. There are many areas of frequentist methodology that should be replaced by Bayesian methodology that provides superior answers. There are frequentist situations that cannot be handled by Bayesian procedures effectively. There are some Bayesian methodologies that have been exposed as having potentially serious frequentist problems. Bayesian and frequentist methods cater to different purposes. Philosophical unification of these two is unlikely, as each highlights a different aspect of statistical analysis. It is recommend that astronomers facing important modeling problems analyze their data using both approaches and use scientific judgment. Understanding the assumptions behind both theories and proper interpretation of their mathematical conclusions are essential.

References

M. J. Bayarri, and J. O. Berger, "The Interplay of Bayesian and Frequentist Analysis", *Statistical Science*, vol. **19**, n. 1, 58-80 (2004).

J. O. Berger, "Could Fisher, Jeffreys and Neyman have agreed on testing", *Statistical Science*, vol. **18**, n. 1, 1-32 (2003).

J. O. Berger, L. D. Brown, and R. Wolpert, "A unified conditional frequentist Bayesian test for fixed and sequential simple hypothesis testing", *Annals of Statistics*, vol. **22**, 1787-1807 (1994).

H.F. Inman, "Karl Pearson and R. A. Fisher on Statistical Tests: A 1935 Exchange from Nature", *The American Statistician*, **48**, 211 (1994).

H. Jeffreys, *Theory of Probability*, 3rd . ed, Oxford University Press. (1961).

J. Neyman, and E. L. Scott, "Consistent estimates based on partially consistent observations", *Econometrica*, **16**, 1-32 (1948).