Statistical Science
2003, Vol. 18, No. 1, 1–32

Could Fisher, Jeffreys and Neyman Have Agreed on Testing?

James O. Berger

Abstract. Ronald Fisher advocated testing using p-values, Harold Jeffreys proposed use of objective posterior probabilities of hypotheses and Jerzy Neyman recommended testing with fixed error probabilities. Each was quite critical of the other approaches. Most troubling for statistics and science is that the three approaches can lead to quite different practical conclusions.

This article focuses on discussion of the conditional frequentist approach to testing, which is argued to provide the basis for a methodological unification of the approaches of Fisher, Jeffreys and Neyman. The idea is to follow Fisher in using p-values to define the "strength of evidence" in data and to follow his approach of conditioning on strength of evidence; then follow Neyman by computing Type I and Type II error probabilities, but do so conditional on the strength of evidence in the data. The resulting conditional frequentist error probabilities equal the objective posterior probabilities of the hypotheses advocated by Jeffreys.

Key words and phrases: p-values, posterior probabilities of hypotheses, Type I and Type II error probabilities, conditional testing.

1. INTRODUCTION

1.1 Disagreements and *Disagreements*

Ronald Fisher, Harold Jeffreys and Jerzy Neyman **disagreed** as to the correct foundations for statistics, but often agreed on the actual statistical procedure to use. For instance, all three supported use of the same estimation and confidence procedures for the elementary normal linear model, **disagreeing** only on the interpretation to be given. As an example, Fisher, Jeffreys and Neyman **agreed** on $(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$ as the 95% confidence interval for a normal mean, but insisted on assigning it fiducial, objective Bayesian and frequentist interpretations, respectively. While the debate over interpretation can be strident, statistical practice is little affected as long as the reported numbers are the same.

James O. Berger is the Arts and Sciences Professor of Statistics in the Institute of Statistics and Decision Sciences at Duke University, Durham, North Carolina 27708-0251 (e-mail: berger@stat.duke.edu).

The situation in testing is quite different. For many types of testing, Fisher, Jeffreys and Neyman **disagreed** as to the basic numbers to be reported and could report considerably different conclusions in actual practice.

EXAMPLE 1. Suppose the data, $X_1, \ldots, X_n$, are i.i.d. from the $\mathcal{N}(\theta, \sigma^2)$ distribution, with σ^2 known, and $n = 10$, and that it is desired to test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. If $z = \sqrt{n}\bar{x}/\sigma = 2.3$ (or $z = 2.9$):

- Fisher would report the p-values $p = 0.021$ (or $p = 0.0037$).
- Jeffreys would report the posterior probabilities of H_0, $\Pr(H_0|x_1, \ldots, x_n) = 0.28$ [or $\Pr(H_0|x_1, \ldots, x_n) = 0.11$], based on assigning the hypotheses equal prior probabilities of $1/2$ and using a conventional Cauchy$(0, \sigma)$ prior on the alternative.
- Neyman, had he prespecified Type I error probability $\alpha = 0.05$, would report $\alpha = 0.05$ in either case (and a Type II error probability β or power function).

The discrepancy between the numbers reported by Fisher and Jeffreys are dramatic in both cases, while the discrepancy between the numbers reported by

Fisher and Neyman are dramatic primarily in the second case. Even if one goes past the raw numbers and considers the actual "scales of evidence" recommended by the three, significant differences remain (see, e.g., Efron and Gous, 2001).

The ***disagreement*** occurs primarily when testing a "precise" hypothesis as above. When testing a one-sided hypothesis, such as $H_0 : \theta \leq 0$, the numbers reported by Fisher and Jeffreys would often be similar (see Casella and Berger, 1987, for discussion—but see Berger and Mortera, 1999, for an alternative perspective). Here precise hypothesis does not necessarily mean a point null hypothesis; the discussion applies equally well to a small interval null hypothesis (see Berger and Delampady, 1987). Also, the null hypothesis can have nuisance parameters that are common to the alternative hypothesis.

We begin, in Section 2, by reviewing the approaches to testing espoused by Fisher, Jeffreys and Neyman and the criticisms each had of the other approaches. The negative impact upon science that has resulted from the ***disagreement*** is also discussed. In Section 3, we describe the conditional frequentist testing paradigm that is the basis of the unification of the three viewpoints. Section 4 discusses how this would have allowed Fisher, Jeffreys and Neyman to simply **disagree**—that is, to report the same numbers, though assigning them differing interpretations. Section 5 discusses various generalizations of the unified approach.

Before beginning, a few caveats are in order. The first is about the title of the article. Fisher, Jeffreys and Neyman all held very strong opinions as to the appropriateness of their particular views of statistics, and it is unlikely that they would have personally reached agreement on this issue. What we are really discussing, therefore, is the possibility of a unification being achieved in which the core principles of each of their three schools are accommodated.

Another caveat is that this is not written as a historical work and quotations justifying the stated positions of Fisher, Jeffreys and Neyman are not included. Key books and publications of the three that outline their positions and give their criticisms of the other approaches include Fisher (1925, 1935, 1955, 1973), Neyman and Pearson (1933), Neyman (1961, 1977) and Jeffreys (1961). Other references and much useful historical discussion can be found, for instance, in Morrison and Henkel (1970), Spielman (1974, 1978), Carlson (1976), Savage (1976), Hall and Selinger (1986), Zabell (1992), Lehmann (1993), Johnstone (1997), Barnett (1999) and Hubbard (2000). Furthermore, Fisher, Jeffreys and Neyman were statisticians of great depth and complexity, and their actual viewpoints toward statistics were considerably more subtle than described herein. Indeed, the names Fisher, Jeffreys and Neyman will often be used more as a label for the schools they founded than as specific references to the individuals. It is also for this reason that we discuss Neyman testing rather than the more historically appropriate Neyman–Pearson testing; Egon Pearson seemed to have a somewhat eclectic view of statistics (see, e.g., Pearson, 1955, 1962) and is therefore less appropriate as a label for the "pure" frequentist philosophy of testing.

A final caveat is that we mostly avoid discussion of the very significant philosophical differences between the various schools (cf. Braithwaite, 1953; Hacking, 1965; Kyburg, 1974; Seidenfeld, 1979). We focus less on "what is correct philosophically?" than on "what is correct methodologically?" In part, this is motivated by the view that professional agreement on statistical philosophy is not on the immediate horizon, but this should not stop us from agreeing on methodology, when possible, and, in part, this is motivated by the belief that optimal general statistical methodology must be simultaneously interpretable from the differing viewpoints of the major statistical paradigms.

2. THE THREE APPROACHES AND CORRESPONDING CRITICISMS

2.1 The Approaches of Fisher, Jeffreys and Neyman

In part to set notation, we briefly review the three approaches to testing in the basic scenario of testing simple hypotheses.

Fisher's significance testing. Suppose one observes data $X \sim f(x|\theta)$ and is interested in testing $H_0 : \theta = \theta_0$. Fisher would proceed by:

- Choosing a test statistic $T = t(X)$, large values of T reflecting evidence against H_0.
- Computing the p-value $p = P_0(t(X) \geq t(x))$, rejecting H_0 if p is small. (Here, and throughout the paper, we let X denote the data considered as a random variable, with x denoting the actual observed data.)

A typical justification that Fisher would give for this procedure is that the p-value can be viewed as an index of the "strength of evidence" against H_0, with small p indicating an unlikely event and, hence, an unlikely hypothesis.

Neyman–Pearson hypothesis testing. Neyman felt that one could only test a null hypothesis, $H_0 : \theta = \theta_0$, versus some alternative hypothesis, for instance, $H_1 : \theta = \theta_1$. He would then proceed by:

- Rejecting H_0 if $T \geq c$ and accepting otherwise, where c is a *pre-chosen* critical value.
- Computing Type I and Type II error probabilities, $\alpha = P_0(\text{rejecting } H_0)$ and $\beta = P_1(\text{accepting } H_0)$.

Neyman's justification for this procedure was the frequentist principle, which we state here in the form that is actually of clear practical value. (See Neyman, 1977. Berger, 1985a and b contain discussions relating this practical version to more common textbook definitions of frequentism.)

FREQUENTIST PRINCIPLE. In repeated practical use of a statistical procedure, the long-run average actual error should not be greater than (and ideally should equal) the long-run average reported error.

The Jeffreys approach to testing. Jeffreys agreed with Neyman that one needed an alternative hypothesis to engage in testing and proceeded by:

- Defining the Bayes factor (or likelihood ratio) $B(x) = f(x|\theta_0)/f(x|\theta_1)$.
- Rejecting H_0 (accepting H_0) as $B(x) \leq 1$ [$B(x) > 1$].
- Reporting the objective posterior error probabilities (i.e., the posterior probabilities of the hypotheses)

$$\Pr(H_0|x) = \frac{B(x)}{1 + B(x)} \quad \left(\text{or } \Pr(H_1|x) = \frac{1}{1 + B(x)}\right) \tag{1}$$

based on assigning equal prior probabilities of $1/2$ to the two hypotheses and applying the Bayes theorem.

Note that we are using "objective" here as a label to distinguish the Jeffreys approach to Bayesian analysis from the subjective approach. Whether any approach to statistics can really claim to be *objective* is an issue we avoid here; see Berger and Berry (1988) for discussion.

2.2 Criticisms of the Three Approaches

The discussion here will be very limited: Fisher, Jeffreys and Neyman each had a lot to say about the other approaches, but space precludes more than a rather superficial discussion of their more popularized criticisms.

Criticisms of the Bayesian approach. Fisher and Neyman felt that it is difficult and/or inappropriate to choose a prior distribution for Bayesian testing. Sometimes criticism would be couched in the language of objectivity versus subjectivity; sometimes phrased in terms of the inadequacy of the older inverse probability version of Bayesianism that had been central to statistical inference since Laplace (1812); and sometimes phrased in terms of a preference for the frequency meaning of probability.

The comments by Fisher and Neyman against the Bayesian approach were typically quite general, as opposed to focusing on the specifics of the developments of Jeffreys. For instance, the fact that the methodology proposed by Jeffreys can lead to Bayesian confidence intervals that are also asymptotically optimal frequentist confidence intervals (Welch and Peers, 1963) did not seem to enter the debate. What could be viewed as an analogue of this result for testing will be central to our argument.

Criticisms of Neyman–Pearson testing. Both Fisher and Jeffreys criticized (unconditional) Type I and Type II errors for not reflecting the variation in evidence as the data range over the rejection or acceptance regions. Thus, reporting a prespecified $\alpha = 0.05$ in Example 1, regardless of whether $z = 2$ or $z = 10$, seemed highly unscientific to both. Fisher also criticized Neyman–Pearson testing because of its need for an alternative hypothesis and for the associated difficulty of having to deal with a power function depending on (typically unknown) parameters.

Criticisms of p-values. Neyman criticized p-values for violating the frequentist principle, while Jeffreys felt that the logic of basing p-values on a tail area (as opposed to the actual data) was silly ["... a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred" (Jeffreys, 1961)]. More recently—and related to both these criticisms—there has been great concern that the too-common misinterpretation of p-values as error probabilities very often results in considerable overstatement of the evidence against H_0; compare Edwards, Lindman and Savage (1963), Gibbons and Pratt (1975), Berger and Sellke (1987), Berger and Delampady (1987), Delampady and Berger (1990) and even the popular press (Matthews, 1998).

Dramatic illustration of the nonfrequentist nature of p-values can be seen from the *applet* available at www.stat.duke.edu/∼berger. The applet assumes one faces a series of situations involving normal data with

unknown mean θ and known variance, and tests of the form $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. The applet simulates a long series of such tests and records how often H_0 is true for p-values in given ranges.

Use of the applet demonstrates results such as if, in this long series of tests, half of the null hypotheses are initially true, then, among the subset of tests for which the p-value is near 0.05, at least 22%—and typically over 50%—of the corresponding null hypotheses will be true. As another illustration, Sterne and Davey Smith (2001) estimated that roughly 90% of the null hypotheses in the epidemiology literature are initially true; the applet shows that, among the subset of such tests for which the p-value is near 0.05, at least 72%—and typically over 90%—of the corresponding null hypotheses will be true. The harm from the common misinterpretation of $p = 0.05$ as an error probability is apparent.

2.3 Impact on Science of the *Disagreement*

We do not address here the effect on statistics of having three (actually more) warring factions, except to say the obvious: it has not been good for our professional image. Our focus, instead, is on the effect that the ***disagreement*** concerning testing has had on the scientific community.

Goodman (1999a, b) and Hubbard (2000), elaborating on earlier work such as Goodman (1992, 1993) and Royall (1997), made a convincing case that the disagreement between Fisher and Neyman has had a significantly deleterious effect upon the practice of statistics in science, essentially because it has led to widespread confusion and inappropriate use of testing methodology in the scientific community. The argument is that testers—in applications—virtually always utilize p-values, but then typically interpret the p-values as error probabilities and act accordingly. The dangers in this are apparent from the discussion at the end of the last section. Note that this confusion is different from the confusion between a p-value and the posterior probability of the null hypothesis; while the latter confusion is also widespread, it is less common in serious uses of statistics.

Fisher and Neyman cannot be blamed for this situation: Neyman was extremely clear that one should use preexperimentally chosen error probabilities if frequentist validity is desired, while Fisher was very careful in distinguishing p-values from error probabilities.

Concerns about this (and other aspects of the inappropriate use of p-values) have repeatedly been raised in many scientific writings. To access at least some of the literature, see the following web pages devoted to the topic in various sciences:

Environmental sciences: www.indiana.edu/∼stigtsts
Social sciences: www.coe.tamu.edu/∼bthompson
Wildlife science:
www.npwrc.usgs.gov/perm/hypotest
www.cnr.colostate.edu/∼anderson/null.html.

It is natural (and common) in these sciences to fault the statistics profession for the situation, pointing out that common textbooks teach frequentist testing and then p-values, without sufficient warning that these are completely different methodologies (e.g., without showing that a p-value of 0.05 often corresponds to a frequentist error probability of 0.5, as indicated by the mentioned applet and conditional frequentist developments).

In contrast, the statistics profession mostly holds itself blameless for this state of affairs, observing that the statistical literature (and good textbooks) does have appropriate warnings. But we are not blameless in one sense: we have not made a concerted professional effort to provide the scientific world with a unified testing methodology (a few noble individual efforts—such as Lehmann, 1993—aside) and so we are tacit accomplices in the unfortunate situation. With a unified testing methodology now available, it is time to mount this effort and provide nonstatisticians with testing tools that they can effectively use and understand.

3. CONDITIONAL FREQUENTIST TESTING

3.1 Introduction to Conditioning

Conditional inference is one of the most important concepts in statistics, but often it is not taught in statistics courses or even graduate programs. In part this is because conditioning is automatic in the Bayesian paradigm—and hence not a subject of particular methodological interest to Bayesians—while, in the frequentist paradigm, there is no established general theory as to how to condition. Frequentists do condition automatically in various circumstances. For instance, consider a version of the famous Cox (1958) example, in which, say, an assay is sometimes run with a sample of size $n = 10$ and other times with a sample of size $n = 20$. If the choice of sample size does not depend on the unknowns under consideration in the assay (e.g., if it depends only on whether an employee is home sick or not), then virtually everyone would condition on the sample size, rather than, say, report an

error probability that is the average of the error probabilities one would obtain for the two sample sizes.

To be precise as to the type of conditioning we will discuss, it is useful to begin with a simple example, taken from Berger and Wolpert (1988) (which also discusses conditioning in general; see also Reid, 1995; Bjørnstad, 1996).

EXAMPLE 2. Two observations, X_1 and X_2, are to be taken, where

$$X_i = \begin{cases} \theta + 1, & \text{with probability } 1/2, \\ \theta - 1, & \text{with probability } 1/2. \end{cases}$$

Consider the confidence set for the unknown θ:

$$C(X_1, X_2) = \begin{cases} \text{the point } \{\frac{1}{2}(X_1 + X_2)\}, \\ \quad \text{if } X_1 \neq X_2, \\ \text{the point } \{X_1 - 1\}, \\ \quad \text{if } X_1 = X_2. \end{cases}$$

The (unconditional) frequentist coverage of this confidence set can easily be shown to be

$$P_\theta\big(C(X_1, X_2) \text{ contains } \theta\big) = 0.75.$$

This is not at all a sensible report, once the data are at hand. To see this, observe that, if $x_1 \neq x_2$, then we know for sure that their average is equal to θ, so that the confidence set is then actually 100% accurate. On the other hand, if $x_1 = x_2$, we do not know if θ is the data's common value plus 1 or their common value minus 1, and each of these possibilities is equally likely to have occurred.

To obtain sensible frequentist answers here, one can define the conditioning statistic $S = |X_1 - X_2|$, which can be thought of as measuring the strength of evidence in the data ($S = 2$ reflecting data with maximal evidential content and $S = 0$ being data of minimal evidential content). Then one defines frequentist coverage conditional on the strength of evidence S. For the example, an easy computation shows that this conditional confidence equals, for the two distinct cases,

$$P_\theta\big(C(X_1, X_2) \text{ contains } \theta \mid S = 2\big) = 1,$$
$$P_\theta\big(C(X_1, X_2) \text{ contains } \theta \mid S = 0\big) = \tfrac{1}{2}.$$

It is important to realize that conditional frequentist measures are fully frequentist and (to most people) clearly better than unconditional frequentist measures. They have the same unconditional property (e.g., in the above example one will report 100% confidence half the time and 50% confidence half the time, resulting in an "average" of 75% confidence, as must be the case for a frequentist measure), yet give much better indications of the accuracy for the type of data that one has actually encountered.

Exactly the same idea applies to testing. In the case of testing simple hypotheses $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, one determines a statistic $S(x)$, the magnitude of which indicates the strength of evidence in x. Then one computes conditional frequentist error probabilities of Type I and Type II, respectively, as

$$\alpha(s) = P_0(\text{reject } H_0 | S(x) = s) \quad \text{and}$$
$$\beta(s) = P_1(\text{accept } H_0 | S(x) = s). \tag{2}$$

A notational comment: a variety of other names are often given to conditioning quantities in the literature. Fisher often used the term "relevant subsets" to refer to subsets of the sample space upon which one should condition. In Example 2, these would be $\{(x_1, x_2) : x_1 = x_2\}$ and $\{(x_1, x_2) : x_1 \neq x_2\}$. Another common term (as in Lehmann, 1993) is "frame of reference," referring to the sample space (or subset thereof) that is actually to be used for the frequentist computation.

3.2 Brief History of Conditional Frequentist Testing

Fisher often used conditioning arguments in testing, as in the development of the Fisher exact test, wherein he chose S to be the marginal totals in a contingency table and then computed p-values conditional on these marginal totals. In addition, Fisher recommended that statisticians routinely condition on an ancillary statistic S (a statistic that has a distribution that does not depend on θ), when available. Fisher's arguments for conditioning were a mix of theory and pragmatism (cf. Savage, 1976; Basu, 1975, 1977), and led to a wide variety of conditioning arguments being developed in the *likelihood school* of statistics (see, e.g., Cox, 1958; Kalbfleish and Sprott, 1973; Reid, 1995).

The use of conditioning in the pure frequentist school was comparatively sporadic, perhaps because Neyman rarely addressed the issue (in spite of frequent criticism by Fisher concerning the supposed lack of conditioning in the frequentist school). The first extensive discussions of conditional frequentist testing were in Kiefer (1976, 1977) and Brown (1978). Among the many observations they made was that, from a frequentist perspective, any conditioning statistic—not just an ancillary statistic—could be employed. However, usual frequentist criteria did not seem to be useful in suggesting the conditioning statistic to use, so

the theory did not immediately lead to the development of statistical methodology. As late as 1993, Lehmann (1993) asserted, "This leaves the combined theory [of testing] with its most difficult issue: What is the relevant frame of reference?"

Berger, Brown and Wolpert (1994) approached the issue of choice of the conditioning statistic from the perspective of seeking a unification between conditional frequentist testing and Bayesian testing, and it is a version of the test proposed therein (as reformulated in Wolpert, 1996) that we will be discussing. That this test also provides a potential unification with Fisherian testing was only recently realized, however.

3.3 Recommended Conditioning Statistic and Test

Fisher argued that p-values are good measures of the strength of evidence against a hypothesis. A natural thought is thus to use p-values to define the conditioning statistic for testing. Thus, for $i = 0, 1$, let p_i be the p-value in testing H_i against the other hypothesis and define the conditioning statistic

$$S = \max\{p_0, p_1\}. \tag{3}$$

The use of this conditioning statistic is equivalent to deciding that data (in either the rejection or acceptance region) that have the same p-value have the same strength of evidence. Note that p-values are only being used in an ordinal sense; any strictly monotonic function of p, applied to both hypotheses, would lead to the same conditioning statistic.

The natural corresponding conditional test proceeds by:

- Rejecting H_0 when $p_0 \le p_1$, and accepting otherwise.
- Computing the Type I and Type II conditional error probabilities (CEPs) as in (2).

Using the results in Berger, Brown and Wolpert (1994), this can be shown to result in the test T^C, defined by

$$T^C = \begin{cases} \text{if } p_0 \le p_1, \\ \quad \text{reject } H_0 \text{ and report Type I CEP} \\ \quad \alpha(x) = \dfrac{B(x)}{1 + B(x)}, \\ \text{if } p_0 > p_1, \\ \quad \text{accept } H_0 \text{ and report Type II CEP} \\ \quad \beta(x) = \dfrac{1}{1 + B(x)}, \end{cases} \tag{4}$$

where $B(x)$ is the likelihood ratio (or Bayes factor).

EXAMPLE 3 (Taken from Sellke, Bayarri and Berger, 2001). It is desired to test

$$H_0 : X \sim \text{Uniform}(0, 1) \text{ versus } H_1 : X \sim \text{Beta}(1/2, 1).$$

The Bayes factor (or likelihood ratio) is then $B(x) = 1/(2\sqrt{x})^{-1} = 2\sqrt{x}$. Computation yields $p_0 = P_0(X \le x) = x$ and $p_1 = P_1(X \ge x) = 1 - \sqrt{x}$. Thus the conditioning statistic is $S = \max\{p_0, p_1\} = \max\{x, 1 - \sqrt{x}\}$ (so it is declared that, say, $x = \frac{3}{4}$ in the acceptance region has the same strength of evidence as $x = \frac{1}{16}$ in the rejection region, since they would lead to the same p-value in tests of H_0 and H_1, respectively).

The recommended conditional frequentist test is thus

$$T^C = \begin{cases} \text{if } x \le 0.382, \\ \quad \text{reject } H_0 \text{ and report Type I CEP} \\ \quad \alpha(x) = (1 + \frac{1}{2}x^{-1/2})^{-1}, \\ \text{if } x > 0.382, \\ \quad \text{accept } H_0 \text{ and report Type II CEP} \\ \quad \beta(x) = (1 + 2x^{1/2})^{-1}. \end{cases}$$

Note that the CEPs both vary with the strength of evidence in the data, as was one of the basic goals.

4. THE POTENTIAL *AGREEMENT*

We consider Neyman, Fisher and Jeffreys in turn, and discuss why T^C might—and might not—have appealed to them as a unifying test.

4.1 Neyman

The potential appeal of the test to Neyman is straightforward: it is fully compatible with the frequentist principle and hence is allowed within the frequentist paradigm. Neyman rarely discussed conditioning, in spite of considerable criticisms from Fisher in this regard, as noted above, and so it is difficult to speculate as to his reaction to use of the conditioning statistic in (3). The result—having a true frequentist test with error probabilities fully varying with the data—would have certainly had some appeal, if for no other reason than that it eliminates the major criticism of the Neyman–Pearson frequentist approach. Also, Neyman did use conditioning as a technical tool, for instance, in developments relating to similar tests (see, e.g., Neyman and Pearson, 1933), but in these developments the conditional Type I error always equalled the unconditional Type I error, so the fundamental issues involving conditioning were not at issue.

Neyman might well have been critical of conditioning that affected optimality properties, such as power.

This can occur if conditioning is used to alter the decision rule. The classic example of Cox (1958) is a good vehicle for discussing this possibility.

EXAMPLE 4. Suppose X is normally distributed as $\mathcal{N}(\theta, 1)$ or $\mathcal{N}(\theta, 4)$, depending on whether the outcome, Y, of flipping a fair coin is heads ($y = 1$) or tails ($y = 0$). It is desired to test $H_0 : \theta = -1$ versus $H_1 : \theta = 1$. The most powerful (unconditional) level $\alpha = 0.05$ test can then be seen to be the test with rejection region given by $x \geq 0.598$ if $y = 1$ and $x \geq 2.392$ if $y = 0$.

Instead, it seems natural to condition upon the outcome of the coin flip in the construction of the tests. Given $y = 1$, the resulting most powerful $\alpha = 0.05$ level test would reject if $x \geq 0.645$, while, given $y = 0$, the rejection region would be $x \geq 2.290$. This is still a valid frequentist test, but it is no longer unconditionally optimal in terms of power and Neyman might well have disapproved of the test for this reason. Lehmann (1993) provided an excellent discussion of the tradeoffs here.

Note, however, that the concern over power arises, not because of conditioning per se, but rather because the decision rule (rejection region) is allowed to change with the conditioning. One could, instead, keep the most powerful unconditional rejection region (so that the power remains unchanged), but report error probabilities conditional on Y. The resulting Type I error probabilities, conditional on $y = 1$ and $y = 0$, would be $\alpha(1) = 0.055$ and $\alpha(0) = 0.045$, respectively. The situation is then exactly the same as in Example 2, and there is no justification for reporting the unconditional $\alpha = 0.05$ in lieu of the more informative $\alpha(1) = 0.055$ or $\alpha(0) = 0.045$. (One can, of course, also report the unconditional $\alpha = 0.05$, since it reflects the chosen design for the experiment, and some people might be interested in the design, but it should be clearly stated that the conditional error probability is the operational error probability, once the data are at hand.)

We are not arguing that the unconditional most powerful rejection region is better; indeed, we agree with Lehmann's (1993) conclusion that conditioning should usually take precedence over power when making decisions. However, we are focusing here only on the inferential report of conditional error probabilities, in which case concerns over power do not arise.

Of course, we actually advocate conditioning in this article on (3) and not just on y. Furthermore, as we are following Fisher in defining the strength of evidence in the data based on p-values, we must define S separately for $y = 1$ and $y = 0$, so that we do condition on Y as well as S. The resulting conditional frequentist test is still defined by (4) and is easily seen to be

$$T^C = \begin{cases} \text{if } x \geq 0, \\ \quad \text{reject } H_0 \text{ and report Type I CEP} \\ \quad \alpha(x, y) = (1 + \exp\{2^{(2y-1)}x\})^{-1}, \\ \text{if } x < 0, \\ \quad \text{accept } H_0 \text{ and report Type II CEP} \\ \quad \beta(x, y) = (1 + \exp\{-2^{(2y-1)}x\})^{-1}. \end{cases}$$

Note that the answers using this fully conditional frequentist test can be quite different from the answers obtained by conditioning on Y alone. For instance, at the boundary of the unconditional most powerful rejection region ($x = 0.598$ if $y = 1$ and $x = 2.392$ if $y = 0$), the CEPs are $\alpha(0.598, 1) = \alpha(2.392, 0) = 0.232$. At, say, $x = 4.0$, the CEPs are $\alpha(4.0, 1) = 0.00034$ and $\alpha(4.0, 0) = 0.119$, respectively. Clearly these results convey a dramatically different message than the error probabilities conditioned only on Y (or the completely unconditional $\alpha = 0.05$).

Another feature of T^C that Neyman might have taken issue with is the specification of the rejection region in (4). We delay discussion of this issue until Section 5.1.

4.2 Fisher

Several aspects of T^C would likely have appealed to Fisher. First, the test is utilizing p-values to measure strength of evidence in data, as he recommended, and conditioning upon strength of evidence is employed. The resulting test yields error probabilities that fully vary with the strength of evidence in the data, a property that he felt was essential (and which caused him to reject Neyman–Pearson testing). In a sense, one can think of T^C as converting p-values into error probabilities, while retaining the best features of both.

One could imagine that Fisher would have questioned the use of (3) as a conditioning statistic, since it will typically not be ancillary, but Fisher was quite pragmatic about conditioning and would use nonancillary conditioning whenever it was convenient (e.g., to eliminate nuisance parameters, as in the Fisher exact test, or in fiducial arguments: see Basu, 1977, for discussion). The use of max rather than the more natural min in (3) might have been a source of concern to Fisher; we delay discussion of this issue until Section 5.2.

Fisher would have clearly disliked the fact that an alternative hypothesis is necessary to define the test T^C. We return to this issue in Section 5.3.

4.3 Jeffreys

The most crucial fact about the CEPs in (4) is that they precisely equal the objective Bayesian error probabilities, as defined in (1). Thus the conditional frequentist and objective Bayesian end up reporting the same error probabilities, although they would imbue them with different meanings. Hence we have **agreement** as to the reported numbers, which was the original goal. Jeffreys might have slightly disagreed with the rejection region specified in (4); we again delay discussion until Section 5.1.

Some statisticians (the author among them) feel that a statistical procedure is only on strong grounds when it can be justified and interpreted from at least the frequentist and Bayesian perspectives. That T^C achieves this unification is a powerful argument in its favor.

4.4 Other Attractions of T^C

The new conditional frequentist test has additional properties that might well have appealed to Fisher, Jeffreys and Neyman. A few of these are listed here.

4.4.1 *Pedagogical attractions.* Conditional frequentist testing might appear difficult, because of the need to introduce the conditioning statistic S. Note, however, that the test T^C is presented from a fully operational viewpoint in (4), and there is no mention whatsoever of the conditioning statistic. In other words, the test can be presented methodologically without ever referring to S; the conditioning statistic simply becomes part of the background theory that is often suppressed.

Another item of pedagogical interest is that teaching statistics suddenly becomes easier, for three reasons. First, it is considerably less important to disabuse students of the notion that a frequentist error probability is the probability that the hypothesis is true, given the data, since a CEP actually has that interpretation. Likewise, one need not worry to such an extent about clarifying the difference between p-values and frequentist error probabilities. Finally, in teaching testing, there is only one test—that given in (4). Moving from one statistical scenario to another requires only changing the expression for $B(x)$ (and this is even true when testing composite hypotheses).

4.4.2 *Simplifications that ensue.* The recommended conditional frequentist test results in very significant simplifications in testing methodology. One of the most significant, as discussed in Berger, Boukai and Wang (1997, 1999), is that the CEPs do not depend on the stopping rule in sequential analysis so that (i) their computation is much easier (the same as fixed sample size computations) and (ii) there is no need to "spend α" to look at the data. This last point removes the perceived major conflict between ethical considerations and discriminatory power in clinical trials; one sacrifices nothing in discriminatory power by evaluating (and acting upon) the evidence after each observation has been obtained.

A second simplification is that the error probabilities are computable in small sample situations, without requiring simulation over the sample space or asymptotic analysis. One only needs to be able to compute $B(x)$ in (4). An example of this will be seen later, in a situation involving composite hypotheses.

5. EXTENSIONS

5.1 Alternative Rejection Regions

A feature of T^C that is, at first, disconcerting is that the rejection region need not be specified in advance; it is predetermined as $\{x : p_0(x) \le p_1(x)\}$. This is, in fact, the minimax rejection region, that is, that which has unconditional error probabilities $\alpha = \beta$. The disconcerting aspect is that, classically, one is used to controlling the Type I error probability through choice of the rejection region, and here there seems to be no control. Note, however, that the unconditional α and β are not used as the reported error probabilities; the conditional $\alpha(x)$ and $\beta(x)$ in (4) are used instead. In Example 3, for instance, when $x = 0.25$, one rejects and reports Type I CEP $\alpha(0.25) = (1 + \frac{1}{2}(0.25)^{-1/2})^{-1} = 0.5$. While H_0 has formally been rejected, the fact that the reported conditional error probability is so high conveys the clear message that this is a very uncertain conclusion.

For those uncomfortable with this mode of operation, note that it is possible to, instead, specify an ordinary rejection region (say, at the unconditional $\alpha = 0.05$ level), find the "matching" acceptance region (which would essentially be the 0.05 level rejection region if H_1 were the null hypothesis), and name the region in the middle the *no-decision* region. The conditional test would be the same as before, except that one would now state "no decision" when the data are in the middle region. The CEPs would not be affected by this change, so that it is primarily a matter of preferred style of presentation (whether to give a decision with a high CEP or simply state no decision in that case).

A final comment here relates to a minor dissatisfaction that an objective Bayesian might have with T^C.

An objective Bayesian would typically use, as the rejection region, the set of potential data for which $P(H_0 \mid x) \leq 1/2$, rather than the region given in (4). In Berger, Brown and Wolpert (1994), this concern was accommodated by introducing a no-decision region consisting of the potential data that would lead to this conflict. Again, however, this is of little importance statistically (the data in the resulting no-decision region would be very inconclusive in any case), so simplicity argues for sticking with T^C.

5.2 Other Types of Conditioning

One could consider a wide variety of conditioning statistics other than that defined in (3). Sellke, Bayarri and Berger (2001) explored, in the context of Example 3, other conditioning statistics that have been suggested. A brief summary of the results they found follows.

Ancillary conditioning statistics rarely exist in testing and, when they exist, can result in unnatural conditional error probabilities. For instance, in Example 3, if one conditions on the ancillary statistic (which happens to exist in this example), the result is that $\beta(x) \equiv 1/2$ as the likelihood ratio, $B(x)$, varies from 1 to 2. This violates the desire for error probabilities that vary with the strength of evidence in the data.

Birnbaum (1961) suggested "intrinsic significance," based on a type of conditioning defined through likelihood concepts. Unfortunately, he found that it rarely works. Indeed, in Example 3, use of the corresponding conditioning statistic yields $\alpha(x) \equiv 1$ as $B(x)$ varies between 0 and $1/2$.

Kiefer (1977) suggested "equal probability continuum" conditioning, which yields the unnatural result, in Example 3, that $\beta(x) \to 0$ as $B(x) \to 2$; to most statisticians, a likelihood ratio of 2 would not seem equivalent to an error probability of 0.

In classical testing using p-values, the focus is usually on small p-values. It thus might seem more natural to condition on $S = \min\{p_0, p_1\}$ rather than $S = \max\{p_0, p_1\}$ when defining the conditional frequentist test. The motivation would be that instead of equating evidence *in favor* of the two hypotheses, one would equate evidence *against* them. In Example 3, however, this yields answers that are clearly unsatisfactory. For instance, the resulting conditional error probabilities are such that $\alpha(x) \to 1/3$ as $B(x) \to 0$, while $\beta(x) \to 0$ as $B(x) \to 2$, neither of which is at all sensible.

Of course, one example is hardly compelling evidence, but the example does show that conditioning statistics can easily lead to error probabilities that are counterintuitive. This is perhaps another reason that conditional frequentist testing has not been common in the statistical community, in spite of its considerable potential advantages. A chief attraction of the conditioning statistic in (3) is that it yields CEPs that can never be counterintuitive, since the resulting error probabilities must coincide with objective Bayesian error probabilities.

5.3 Calibrating p-Values When There Is No Alternative Hypothesis

Fisher often argued that it is important to be able to test a null hypothesis, even if no alternative hypothesis has been determined. The wisdom in doing so has been extensively debated: many statisticians have strong opinions pro and con. Rather than engaging this debate here, we stick to methodology and simply discuss how conditional frequentist testing can be done when there is no specified alternative.

The obvious solution to the lack of a specified alternative is to create a generic nonparametric alternative. We first illustrate this with the example of testing of fit to normality.

EXAMPLE 5. Berger and Guglielmi (2001) considered testing $H_0 : X \sim \mathcal{N}(\mu, \sigma)$ versus $H_1 : X \sim F(\mu, \sigma)$, where F is an unknown location–scale distribution that will be centered at the normal distribution. As mentioned above, the key to developing a conditional frequentist test is first to develop an objective Bayes factor, $B(x)$. This was done by choosing a Polya tree prior for F, centered at the $\mathcal{N}(\mu, \sigma)$ distribution, and choosing the right-Haar prior, $\pi(\mu, \sigma) = 1/\sigma$, for the location–scale parameters in each model. Berger and Guglielmi (2001) showed how to compute $B(x)$.

The recommended conditional frequentist test is then given automatically by (4). Because the null hypothesis has a suitable group invariance structure, the analysis in Dass and Berger (2003) can be used to show that the conditional Type I error is indeed $\alpha(x)$ in (4), while $\beta(x)$ is an average Type II error (see Section 5.4). It is interesting to note that this is an *exact* frequentist test, even for small sample sizes. This is in contrast to unconditional frequentist tests of fit, which typically require extensive simulation or asymptotic arguments for the determination of error probabilities.

Developing specific nonparametric alternatives for important null hypotheses, as above, can be arduous, and it is appealing to seek a generic version that

TABLE 1
Calibration of p-values as lower bounds on conditional error probabilities

p	0.2	0.1	0.05	0.01	0.005	0.001
$\alpha(p)$	0.465	0.385	0.289	0.111	0.067	0.0184

applies widely. To do so, it is useful to again follow Fisher and begin with a p-value for testing H_0. If it is a *proper* p-value, then it has the well-known property of being uniformly distributed under the null hypothesis. (See Bayarri and Berger, 2000, Robins, van der Vaart and Ventura, 2000, and the references therein for discussion and generalizations.) In other words, we can reduce the original hypothesis to the generic null hypothesis that $H_0 : p(X) \sim \text{Uniform}(0, 1)$.

For this p-value null, Sellke, Bayarri and Berger (2001) developed a variety of plausible nonparametric alternatives and showed that they yield a lower bound on the Bayes factor of $B(p) \geq -e\, p \log(p)$. Although each such alternative would result in a different test (4), it is clear that all such tests have

$$\alpha(p) \geq \left(1 + [-e\, p \log(p)]^{-1}\right)^{-1}. \tag{5}$$

This is thus a lower bound on the conditional Type I error (or on the objective posterior probability of H_0) and can be used as a "quick and dirty" calibration of a p-value when only H_0 is available.

Table 1, from Sellke, Bayarri and Berger (2001), presents various p-values and their associated calibrations. Thus $p = 0.05$ corresponds to a frequentist error probability of at least $\alpha(0.05) = 0.289$ in rejecting H_0.

While simple and revealing, the calibration in (5) is often a too-small lower bound on the conditional Type I error. Alternative calibrations have been suggested in, for example, Good (1958, 1992).

5.4 Other Testing Scenarios

For pedagogical reasons, we have only discussed tests of simple hypotheses here, but a wide variety of generalizations exist. Berger, Boukai and Wang (1997, 1999) considered tests of simple versus composite hypotheses, including testing in sequential settings. For composite alternatives, conditional Type II error is now (typically) a function of the unknown parameter (as is the unconditional Type II error or power function) so that it cannot directly equal the corresponding Bayesian error probability. Interestingly, however, a posterior average of the conditional Type II error function does equal the corresponding Bayesian error probability, so that one has the option of reporting the average Type II error or the average power instead of the entire function. This goes a long way toward answering Fisher's criticisms concerning the difficulty of dealing with power functions.

Dass (2001) considered testing in discrete settings and was able to construct the conditional frequentist tests in such a way that very little randomization is necessary (considerably less than for unconditional tests in discrete settings). Dass and Berger (2003) considered composite null hypotheses that satisfy an appropriate invariance structure and showed that essentially the same theory applies. This covers a huge variety of classical testing scenarios. Paulo (2002a, b) considered several problems that arise in sequential experimentation, including comparison of exponential populations and detecting the drift of a Brownian motion.

The program of developing conditional frequentist tests for the myriad of testing scenarios that are considered in practice today will involve collaboration of frequentists and objective Bayesians. This is because the most direct route to determination of a suitable conditional frequentist test, in a given scenario, is the Bayesian route, thus first requiring determination of a suitable objective Bayesian procedure for the scenario.

ACKNOWLEDGMENTS

This research was supported by National Science Foundation Grants DMS-98-02261 and DMS-01-03265. This article is based on the Fisher Lecture given by the author at the 2001 Joint Statistical Meetings.

REFERENCES

BARNETT, V. (1999). *Comparative Statistical Inference*, 3rd ed. Wiley, New York.

BASU, D. (1975). Statistical information and likelihood (with discussion). *Sankhyā Ser. A* **37** 1–71.

BASU, D. (1977). On the elimination of nuisance parameters. *J. Amer. Statist. Assoc.* **72** 355–366.

BAYARRI, M. J. and BERGER, J. (2000). P-values for composite null models (with discussion). *J. Amer. Statist. Assoc.* **95** 1127–1142, 1157–1170.

BERGER, J. (1985a). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York.

BERGER, J. (1985b). The frequentist viewpoint and conditioning. In *Proc. Berkeley Conference in Honor of Jack Kiefer and Jerzy Neyman* (L. Le Cam and R. Olshen, eds.) **1** 15–44. Wadsworth, Belmont, CA.

BERGER, J. and BERRY, D. (1988). Statistical analysis and the illusion of objectivity. *American Scientist* **76** 159–165.

BERGER, J., BOUKAI, B. and WANG, Y. (1997). Unified frequentist and Bayesian testing of a precise hypothesis (with discussion). *Statist. Sci.* **12** 133–160.

BERGER, J., BOUKAI, B. and WANG, Y. (1999). Simultaneous Bayesian–frequentist sequential testing of nested hypotheses. *Biometrika* **86** 79–92.

BERGER, J., BROWN, L. and WOLPERT, R. (1994). A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. *Ann. Statist.* **22** 1787–1807.

BERGER, J. and DELAMPADY, M. (1987). Testing precise hypotheses (with discussion). *Statist. Sci.* **2** 317–352.

BERGER, J. and GUGLIELMI, A. (2001). Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. *J. Amer. Statist. Assoc.* **96** 174–184.

BERGER, J. and MORTERA, J. (1999). Default Bayes factors for non-nested hypothesis testing. *J. Amer. Statist. Assoc.* **94** 542–554.

BERGER, J. and SELLKE, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence (with discussion). *J. Amer. Statist. Assoc.* **82** 112–139.

BERGER, J. and WOLPERT, R. L. (1988). *The Likelihood Principle*, 2nd ed. (with discussion). IMS, Hayward, CA.

BIRNBAUM, A. (1961). On the foundations of statistical inference: Binary experiments. *Ann. Math. Statist.* **32** 414–435.

BJØRNSTAD, J. (1996). On the generalization of the likelihood function and the likelihood principle. *J. Amer. Statist. Assoc.* **91** 791–806.

BRAITHWAITE, R. B. (1953). *Scientific Explanation*. Cambridge Univ. Press.

BROWN, L. D. (1978). A contribution to Kiefer's theory of conditional confidence procedures. *Ann. Statist.* **6** 59–71.

CARLSON, R. (1976). The logic of tests of significance (with discussion). *Philos. Sci.* **43** 116–128.

CASELLA, G. and BERGER, R. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem (with discussion). *J. Amer. Statist. Assoc.* **82** 106–111, 123–139.

COX, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29** 357–372.

DASS, S. (2001). Unified Bayesian and conditional frequentist testing for discrete distributions. *Sankhyā Ser. B* **63** 251–269.

DASS, S. and BERGER, J. (2003). Unified conditional frequentist and Bayesian testing of composite hypotheses. *Scand. J. Statist.* **30** 193–210.

DELAMPADY, M. and BERGER, J. (1990). Lower bounds on Bayes factors for multinomial distributions, with application to chi-squared tests of fit. *Ann. Statist.* **18** 1295–1316.

EDWARDS, W., LINDMAN, H. and SAVAGE, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review* **70** 193–242.

EFRON, B. and GOUS, A. (2001). Scales of evidence for model selection: Fisher versus Jeffreys (with discussion). In *Model Selection* (P. Lahiri, ed.) 208–256. IMS, Hayward, CA.

FISHER, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh (10th ed., 1946).

FISHER, R. A. (1935). The logic of inductive inference (with discussion). *J. Roy. Statist. Soc.* **98** 39–82.

FISHER, R. A. (1955). Statistical methods and scientific induction. *J. Roy. Statist. Soc. Ser. B* **17** 69–78.

FISHER, R. A. (1973). *Statistical Methods and Scientific Inference*, 3rd ed. Macmillan, London.

GIBBONS, J. and PRATT, J. (1975). P-values: Interpretation and methodology. *Amer. Statist.* **29** 20–25.

GOOD, I. J. (1958). Significance tests in parallel and in series. *J. Amer. Statist. Assoc.* **53** 799–813.

GOOD, I. J. (1992). The Bayes/non-Bayes compromise: A brief review. *J. Amer. Statist. Assoc.* **87** 597–606.

GOODMAN, S. (1992). A comment on replication, p-values and evidence. *Statistics in Medicine* **11** 875–879.

GOODMAN, S. (1993). P-values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology* **137** 485–496.

GOODMAN, S. (1999a). Toward evidence-based medical statistics. 1: The p-value fallacy. *Annals of Internal Medicine* **130** 995–1004.

GOODMAN, S. (1999b). Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine* **130** 1005–1013.

HACKING, I. (1965). *Logic of Statistical Inference*. Cambridge Univ. Press.

HALL, P. and SELINGER, B. (1986). Statistical significance: Balancing evidence against doubt. *Austral. J. Statist.* **28** 354–370.

HUBBARD, R. (2000). Minding one's p's and α's: Confusion in the reporting and interpretation of results of classical statistical tests in marketing research. Technical Report, College of Business and Public Administration, Drake Univ.

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Oxford Univ. Press.

JOHNSTONE, D. J. (1997). Comparative classical and Bayesian interpretations of statistical compliance tests in auditing. *Accounting and Business Research* **28** 53–82.

KALBFLEISH, J. D. and SPROTT, D. A. (1973). Marginal and conditional likelihoods. *Sankhyā Ser. A* **35** 311–328.

KIEFER, J. (1976). Admissibility of conditional confidence procedures. *Ann. Math. Statist.* **4** 836–865.

KIEFER, J. (1977). Conditional confidence statements and confidence estimators (with discussion). *J. Amer. Statist. Assoc.* **72** 789–827.

KYBURG, H. E., JR. (1974). *The Logical Foundations of Statistical Inference*. Reidel, Boston.

LAPLACE, P. S. (1812). *Théorie Analytique des Probabilités*. Courcier, Paris.

LEHMANN, E. (1993). The Fisher, Neyman–Pearson theories of testing hypotheses: One theory or two? *J. Amer. Statist. Assoc.* **88** 1242–1249.

MATTHEWS, R. (1998). The great health hoax. *The Sunday Telegraph*, September 13.

MORRISON, D. E. and HENKEL, R. E., eds. (1970). *The Significance Test Controversy. A Reader*. Aldine, Chicago.

NEYMAN, J. (1961). Silver jubilee of my dispute with Fisher. *J. Operations Res. Soc. Japan* **3** 145–154.

NEYMAN, J. (1977). Frequentist probability and frequentist statistics. *Synthese* **36** 97–131.

NEYMAN, J. and PEARSON, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. Roy. Soc. London Ser. A* **231** 289–337.

PAULO, R. (2002a). Unified Bayesian and conditional frequentist testing in the one- and two-sample exponential distribution problem. Technical Report, Duke Univ.

PAULO, R. (2002b). Simultaneous Bayesian–frequentist tests for the drift of Brownian motion. Technical Report, Duke Univ.
PEARSON, E. S. (1955). Statistical concepts in their relation to reality. *J. Roy. Statist. Soc. Ser. B* **17** 204–207.
PEARSON, E. S. (1962). Some thoughts on statistical inference. *Ann. Math. Statist.* **33** 394–403.
REID, N. (1995). The roles of conditioning in inference (with discussion). *Statist. Sci.* **10** 138–157, 173–199.
ROBINS, J. M., VAN DER VAART, A. and VENTURA, V. (2000). Asymptotic distribution of p values in composite null models (with discussion). *J. Amer. Statist. Assoc.* **95** 1143–1167, 1171–1172.
ROYALL, R. M. (1997). *Statistical Evidence: A Likelihood Paradigm*. Chapman and Hall, New York.
SAVAGE, L. J. (1976). On rereading R. A. Fisher (with discussion). *Ann. Statist.* **4** 441–500.
SEIDENFELD, T. (1979). *Philosophical Problems of Statistical Inference.* Reidel, Boston.
SELLKE, T., BAYARRI, M. J. and BERGER, J. O. (2001). Calibration of p-values for testing precise null hypotheses. *Amer. Statist.* **55** 62–71.
SPIELMAN, S. (1974). The logic of tests of significance. *Philos. Sci.* **41** 211–226.
SPIELMAN, S. (1978). Statistical dogma and the logic of significance testing. *Philos. Sci.* **45** 120–135.
STERNE, J. A. C. and DAVEY SMITH, G. (2001). Sifting the evidence—what's wrong with significance tests? *British Medical Journal* **322** 226–231.
WELCH, B. and PEERS, H. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. Roy. Statist. Soc. Ser. B* **25** 318–329.
WOLPERT, R. L. (1996). Testing simple hypotheses. In *Data Analysis and Information Systems* (H. H. Bock and W. Polasek, eds.) **7** 289–297. Springer, Heidelberg.
ZABELL, S. (1992). R. A. Fisher and the fiducial argument. *Statist. Sci.* **7** 369–387.

Comment

Ronald Christensen

I feel privileged to congratulate Jim Berger on his exemplary career leading to the Fisher lectureship, as well as this interesting work with his colleagues. I totally agree with the premise that there is vast confusion about the practical use of testing and I hope that this article puts one more nail into the coffin that Neyman–Pearson testing so richly deserves. However, in my view, except for the incorporation of p-values, this article has little to do with Fisherian testing. Ultimately, the key issue is to get the philosophical ideas down and to use methods that are appropriate to the problems being addressed.

In retrospect I believe that Neyman and Pearson performed a disservice by making traditional testing into a parametric decision problem. Frequentist testing is ill-suited for deciding between alternative parameter values. I think Berger and Wolpert (1984) ably demonstrated that in their wonderful book. For example, when deciding between two hypotheses, why would you reject a hypothesis that is 10 times more likely than the alternative just to obtain some preordained α level? It is a crazy thing to do unless you have prior knowledge that the probability of the alternative occurring is at least nearly 10 times larger. As to picking priors for scientific purposes, if you do not have enough data so that any "reasonable" prior gives the same answers in practice, you obviously cannot construct a scientific consensus and should admit that your results are your opinions.

Outside of Neyman–Pearson theory, testing is properly viewed as model validation. Either the model works reasonably or it does not. There is no parametric alternative hypothesis! To perform either Neyman–Pearson or Bayesian testing, you must have, or construct, a parametric alternative. If you are willing to construct an alternative, you should use one of those theories. (Nonparametric problems are properly thought of as having huge parameter sets.) But at some point we all have to stop dreaming up alternatives and either go on to other problems, retire or die. In model validation, there is a series of assumptions that constitutes the model. Data are obtained and a one-dimensional test statistic is chosen. Either the data, as summarized by the test statistic, appear to be consistent with the model or they do not. If they appear to be inconsistent, obviously it suggests something may be wrong with the model. (Proof by contradiction.) If they appear to be consistent, big deal! (No contradiction, no proof.) The model came from somewhere; one hopes from scientific experience. But we eventually show that all models are wrong. The important ques-

Ronald Christensen is Professor, Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico 87131 (e-mail: fletcher@stat.unm.edu).

tion is whether they are useful. If the data are consistent, they merely do not contribute to showing that this model is not useful.

The only general way to determine if data appear to be inconsistent with the model is to define as inconsistently odd those data values that have the smallest probability density under the model. (This is a major departure from Neyman–Pearson theory, which relies on comparing null and alternative densities.) The p-value is the probability of seeing something as odd or odder than you actually saw. If you see something odd, it gives you reason to doubt the validity of the model you are testing, but it does not suggest which particular aspect of the model is invalid. A parametric null hypothesis only raises its ugly head if one can validate all of the other assumptions that constitute the model, so that only this null hypothesis can be called in question.

If p-values or confidence coefficients do not mean what you want them to mean, you better do a Bayesian analysis, because short of developing fiducial inference, it is the only way I know to get numbers that people like to interpret. While it is interesting to see when p-values may or may not agree with posterior probabilities, if you are not willing to do a Bayesian analysis, you better learn what p-values and confidence coefficients really mean. We need to stop inviting people to misinterpret confidence coefficients which we do by presenting the long-run frequency "justification;" see Christensen (1995). The problem of misinterpreting p-values as posterior probabilities seems to be more limited.

Of course there remains the open question of how to choose a test statistic. Box (1980) suggested that in the subjective Bayesian context one use the marginal density for the potential data evaluated at the observed data. Such a test addresses the appropriateness of both the sampling distribution and the prior.

Comment

Wesley O. Johnson

1. INTRODUCTION

It is a privilege to comment on the weighty topic of statistical testing. Our profession suffers from a lack of coherence in its philosophical underpinnings. As a result, some statisticians use any method that helps them solve the problem at hand, while others rely on a single mode of inference. Many of us have, through the years, settled on some form of hybrid approach to hypothesis testing that involves p-values and/or Type I and Type II error considerations and/or Bayesian calculations. Scientists who use statistics in their research are left to the mercy of the statistics textbooks they have available. Given the differences in philosophy, we statisticians find it surprising when different approaches lead to similar "statistical practice." Nonstatisticians may find it shocking that they can disagree.

The perplexity caused by our differences of opinion has led Berger to try to alleviate some of the confusion by providing methods that are somehow consistent with all major philosophical camps. Berger has a long and distinguished history of attempting to reconcile the Bayesian and frequentist points of view. This article takes on the further challenge of attempting to improve the perception of statistics by the outside world. Consistent with his effort to reconcile, he has not discussed philosophical issues about the relative merits of this or that approach, but rather he focuses on discovering the common ground based on a particular synthesis of the views of Fisher, Neyman and Jeffreys. Professor Berger can only be applauded for taking on this lofty challenge. Moreover, the regimen prescribed is remarkable. He has reason to be proud for discovering it. This article will be read by many and should generate much discussion.

My own statistical foundation has grown over many years to lean, or perhaps topple, toward the subjective Bayesian mode, and this discussion reflects that fact. In doing so, it may depart somewhat from Berger's more lofty goals.

2. EPIDEMIOLOGY, SCREENING TESTS AND POINT NULLS

Berger points out the seemingly "surprising" result that the proportion of nulls that actually hold among

Wesley O. Johnson is Professor, Department of Statistics, University of California, Davis, California 95616 (e-mail: wojohnson@ucdavis.edu).

those times that a p-value is in a neighborhood of 0.05 can be quite large. Such results are quite familiar to epidemiologists. For example, consider a screening problem for some disease like HIV infection (see Gastwirth, Johnson and Reneau, 1991). Such screening tests are often quite accurate, so let us assume the sensitivity of the test (1 minus the false negative test rate) is about 0.99 and the specificity (1 minus the false positive rate) is about 0.995. Since the two error rates are so small, it might seem surprising to some that with an HIV population disease prevalence of 0.0004, for example, only about 7.5% of people with positive results will actually be infected.

Screening tests correspond to standard simple versus simple hypothesis tests. The null hypothesis is that the blood sample in question is infected, so Type I error constitutes falsely indicating that infected blood is not infected. The probability of a Type I error, α, is the false negative rate for the test, which is also 1 minus the test sensitivity, and the probability of a Type II error, β, is the false positive rate for the test, which is also 1 minus the test specificity. The prevalence of HIV infection in a population is analogous to the proportion of null hypotheses that are true. Predictive values (positive and negative) of the screening test are those proportions of correct outcomes for blood samples that were screened positive and negative, respectively.

Conversely, a simple versus simple hypothesis test also determines a screening test. Identify H_0 as the diseased state. Define a positive outcome to be when the p-value is less than the prescribed Type I error probability, namely $p \le \alpha$, and a negative outcome otherwise. Then all of the standard epidemiologic objects defined above apply. While Berger has focused on the probability of the null being true having observed the p-value in a neighborhood of the observed value, I continue the analogy between screening and testing by calculating the probabilities of H_0 when the p-value is either less than or greater than α, which correspond to 1 minus the predictive value positive and the predictive value negative, respectively. Berger also calculated the former.

By the Bayes theorem, the proportions of null hypotheses that are actually "true" given that $p \le \alpha$ and $p > \alpha$, respectively, are

$$\Pr(H_0|p \le \alpha) = \frac{\alpha \Pr(H_0)}{\alpha \Pr(H_0) + (1-\beta)(1-\Pr(H_0))}$$

and

$$\Pr(H_0|p > \alpha) = \frac{(1-\alpha)\Pr(H_0)}{(1-\alpha)\Pr(H_0) + \beta(1-\Pr(H_0))}.$$

If there are small Type I and Type II errors and if the prevalence of true null hypotheses is low, then the following approximations hold (as in Johnson and Gastwirth, 1991): $\Pr(H_0|p \le \alpha) \doteq \alpha \Pr(H_0)$ and $\Pr(H_0|p > \alpha) \doteq \Pr(H_0)/\{\Pr(H_0) + \beta\}$. Thus if the p-value is used to reject at level α and if a small proportion of nulls is a priori true, then the a posteriori proportion of nulls that holds, given $p \le \alpha$, will be α times the a priori proportion. If the a priori proportion of nulls were near 1, then $\Pr(H_0|p \le \alpha) \doteq \alpha/(\alpha + 1 - \Pr(H_0))$ and $\Pr(H_0|p > \alpha) \doteq 1$. Thus if 95% of all nulls were true a priori and $\alpha = 0.05$, then about 50% of the nulls will be true given $p \le \alpha$, while it is expected that virtually all nulls will be true when $p > \alpha$.

A classic problem in risk analysis involves testing that a particular population of individuals is completely disease-free (Suess, Gardner and Johnson, 2002; Hanson, Johnson, Gardner and Georgiadis, 2003). Having even a single individual in the population who is diseased is problematic. For example, in treating the potential for foot-and-mouth disease in cattle, as soon as an individual infected animal is discovered, entire herds in the vicinity of the discovered animal are generally eradicated (anonymous, Foot-and-mouth disease emergency disease guidelines, Animal and Plant Health Inspection Service, USDA, Hyattsville, MD, 1991). Moreover, in animal testing, there is rarely a perfect or "gold-standard" test that can be used for disease surveillance (Joseph, Gyorkos and Coupal, 1995; Enøe, Georgiadis and Johnson, 2000; Johnson, Gastwirth and Pearson, 2001). Without a perfect test, standard herd surveillance data lead to nonidentifiable models, making it virtually impossible to directly test for the absence of disease without either unrealistically assuming known accuracy for the screening tests or performing a subjective Bayesian analysis. Here, the composite null hypothesis is that there are infected units in the population (the prevalence of infection is positive) and the precise alternative states that there is none (the prevalence is zero). The Type I error involves the acceptance of the hypothesis of no infection when there is at least one infected animal in the population. It would appear that one cannot expect reconciliation for this kind of problem since a subjective approach seems warranted and, moreover, it is superficially unclear whether Berger's method will apply to composite nulls with precise alternatives.

3. SAMPLE SIZE

Many statisticians advise, "Don't do hypothesis testing." To a large extent, I subscribe to this philosophy.

Frequentist hypothesis tests present many dangers, not the least of which is the potential misinterpretation of results and confusion discussed by Berger. In my experience, the sample size problem is even more dangerous. It is common for scientists to implicitly accept null hypotheses by referring to factors that are not significant as essentially irrelevant, with no concern whether the estimated effect would be both statistically significant and practically meaningful if their sample size had been larger. For Bayesians, the moral of the famous Jeffreys–Lindley paradox is precisely that using ill-conceived reference priors can lead to inappropriately accepting null hypotheses. However, the problem of inappropriately accepting a null does not generally occur in a subjective Bayesian analysis since null and alternative hypotheses are treated symmetrically. Thus if there is a very high posterior certainty for the null, regardless of the sample size, it is appropriate to accept the null. If a Bayesian or other procedure is equivocal, the analyst must assess whether the observed effect would be of practical import were it real and thus whether a larger sample size could result in the conclusion that a meaningful effect exists.

Bayesian methods seem to offer a more balanced evaluation of the null model. For example, posterior probabilities for point nulls in standard normal models tend to be considerably larger than corresponding p-values (see, e.g., Berger, 1985a, Table 4.2). In Berger's (1985a) example (with $z = 1.96$, $n = 1000$), the p-value is 0.05 but the posterior probability of the null is 0.80. In the more extreme case (with $z = 3.291$, $n = 1000$), $p = 0.001$, but the posterior probability of the null is as large as 0.124. With smaller n and the same values for z, p-values remain unchanged while posterior probabilities decline but stay well above the corresponding p-values. This indicates a natural protection in Bayesian testing against the tendency to reject everything for large sample sizes in frequentist testing. Berger also calculates lower bounds (considering all possible priors under the alternative) on the posterior probability of the null, which turn out to be 0.127 and 0.0044 for these two z-values and for all choices of n. Regardless of the choice of prior under the alternative, it is more difficult to reject a point null using these proper Bayesian methods.

Perhaps with very large sample sizes, conditioning on the maximum p-value statistic results in conditional error rates that shed light on this problem. If so, I wonder what can be said about the situation with, say, 20 covariates in a regression model based on 10,000 observations where the p-values range from some small value less than 0.05 on down.

4. A SUBJECTIVIST'S VIEW VERSUS OBJECTIVE BAYES

The issue of performing an objective versus a subjective Bayesian analysis is a deep philosophical question. Conversations between individuals in opposite camps go on ad infinitum. I have often wondered why there is so much resistance within the statistical community to the use of subjective prior information. I suspect that many of us have little or no formal training in substantive areas of science, so when we approach a data set, it is without expert knowledge of the subject at hand. We often handle many different kinds of data and there is little time to become subject matter experts in even a small subset of these areas. In writing a statistical paper about a Bayesian approach, when selecting a prior distribution we often think that putting a real prior into the analysis would involve guesswork, so objectivity starts to sound very appealing. I could not agree more.

However, as a statistician working with scientists in several substantive areas who seem perfectly at ease with inputting their scientific knowledge in the form of prior distributions, it is difficult for me to understand the apparent reluctance of even Bayesians to consider using subjective prior information in their collaborative efforts. In my experience, it is quite the exception for a scientist to lack knowledge on the expected behavior of relevant observable quantities. It is also quite the exception for them to be reluctant to incorporate that information into the analysis (Joseph, Gyorkos and Coupal, 1995; Bedrick, Christensen and Johnson, 1997, 2000; Westfall, Johnson and Utts, 1997; Liu, Johnson, Gold and Lasley, 2003; Hanson, Johnson and Gardner, 2003; Hanson, Bedrick, Johnson and Thurmond, 2003; Georgiadis, Johnson, Gardner and Singh, 2003; Suess, Gardner and Johnson, 2002; Fosgate et al., 2002; McInturff, Johnson, Gardner and Cowling, 2003). Practical methods for eliciting and incorporating prior information for generalized regression models were developed by Bedrick, Christensen and Johnson (1996) and implemented in several of the above articles.

While I understand Professor Berger to be sympathetic to the use of subjective Bayesian methods and that here he is mainly concerned with attempting to find common ground among the majority of statisticians, it seems to me that this extraordinary effort to reconcile the masses may help to perpetuate the myth that subjectivism is a bad thing rather than a good thing. This is not a criticism of Berger's ef-

forts, but rather an observation. My attempt here is to encourage the use of subjective methods in substantive collaborative efforts. After all, is it not the case in science that there are different opinions about many issues? Do not good scientists disagree with each other while having scientifically sensible reasons for holding their separate opinions? It should not seem surprising when several analyses of the same data result in different conclusions because of different scientific models that were included in the analysis. The goal then is to collect enough of the right kind of data so as to result in very similar posteriors and, consequently, consensus—another type of reconciliation. Bayesian subjective methods provide an eloquent, and some would say coherent, method by which consensus ultimately can be achieved among people whose prior opinions are not so strong as to overwhelm any data.

5. OTHER TYPES OF RECONCILIATION

There are of course many other forms of inference to which the concept of reconciliation would apply. For a fairly general class of estimation problems, Samaniego and Reneau (1994), in particular, have characterized the circumstances when Bayesian point estimation is preferable, and when not, according to Bayes risk, a frequentist criterion. They point out that Bayes rules are often preferable provided the "prior distribution provides 'useful' information about the unknown parameter" and that it is not "overstated." Their conclusion about reconciliation is that both Bayesian and frequentist schools "are correct (and better than the other) under specific (and complementary) circumstances."

Another area that has been considered is the reconciliation of Bayes and frequentist methods for multiple testing problems. In particular, Westfall, Johnson and Utts (1997) discussed the issue of Bayesian multiplicity adjustment and presented a correspondence with the usual Bonferroni adjustment. More recently, Gönen, Westfall and Johnson (2003) developed informative Bayesian methods for multiple two-sample comparisons arising from general multivariate data. In this work, k simultaneous (correlated) point null hypotheses are considered. It is argued that independent point nulls with degree of belief 0.5 in each would result in $(0.5)^k$ degree of belief that all nulls were simultaneously true, which may be far too small to believe. A particular model is presented where the experimenter is asked to specify his or her subjective degree of belief that all nulls are simultaneously true as well as the marginal probabilities for single hypotheses. This induces a priori correlation between beliefs in individual nulls. It is not clear what would be an objective prior for this problem. The interplay between specifying the joint null probability and the marginal probabilities would seem to mitigate against this possibility, as it would seem to require scientific input and, consequently, informative prior information.

Comment

Michael Lavine

I want to begin by thanking Professer Berger for an extremely well written and thought provoking article. Its main points are to explain conditional error probabilities (CEPs) and argue that they can result in methodological unification. This discussion will focus on whether CEPs do, in fact, provide a methodology acceptable to everyone and whether such a unification is desirable.

Michael Lavine is Professor, Institute of Statistics and Decision Sciences, Duke University, Durham, North Carolina 27708-0251 (e-mail: michael@stat.duke.edu).

DO CEPS RESOLVE METHODOLOGICAL ISSUES?

Berger shows that CEPs look promising in each of his examples. Here we examine CEPs to see whether they look sensible when compared across several testing scenarios simultaneously. The idea of comparing across multiple scenarios was previously used by Gabriel (1969), Schervish (1996) and Lavine and Schervish (1999), who used the following idea of coherence. Another criterion for comparing across several scenarios appears in Schervish, Seidenfeld and Kadane (2003):

> Let $H' \subseteq H''$ be nested hypotheses. Any evidence for H' is necessarily evidence for

> H''; any evidence against H'' is necessarily evidence against H'. Therefore any measure m of support or strength-of-evidence should obey $m(H') \leq m(H'')$. A measure of support m is called *coherent* if $m(H') \leq m(H'')$ for all $H' \subseteq H''$ and *incoherent* otherwise. A testing procedure is called coherent if rejection of H'' entails rejection of H' for all $H' \subseteq H''$ and *incoherent* otherwise.

Apart from its intuitive appeal, Lavine and Schervish (1999) give a decision–theoretic argument for coherence.

An example from Lavine and Schervish (1999) illustrates the idea:

> [A consulting client] was comparing three modes of inheritance in the species *Astilbe biternata*. All three modes are represented by simple hypotheses concerning the distribution of the observable data. One hypothesis H_1 is called tetrasomic inheritance, while the other two hypotheses H_2 and H_3 (those which happen to have the largest and smallest likelihoods, respectively) together form a meaningful category, disomic inheritance. The Bayes factor in favor of H_2 will be larger than the Bayes factor in favor of $H_2 \cup H_3$ no matter what strictly positive prior one places over the three hypotheses because H_3 has the smallest likelihood.

Therefore, Bayes factors are incoherent measures of support.

What do CEPs say about the *Astilbe biternata* example? According to Berger (personal communication), in testing H_1 versus H_2, the CEP would be based on the likelihoods of H_1 and H_2. However, in testing H_1 versus $H_2 \cup H_3$, the CEP would use the likelihood based on equal a priori weights for H_2 and H_3. Therefore, CEPs would claim to have less evidence favoring $H_2 \cup H_3$ over H_1 than favoring H_2 over H_1 and would be incoherent.

A similar phenomenon can occur with nested linear models. Let $(x_1, y_1) = (-1, 3)$, $(x_2, y_2) = (1, 3)$ and consider the model $y_i = a + bx_i + \varepsilon_i$, where $\varepsilon_i \sim \mathrm{N}(0, 1)$. What do CEPs say about the hypotheses $H_0 : a = b = 0$, $H_1 : b = 0$ and H_2 : no restrictions? For testing H_0 versus H_1, CEPs would use the Bayes factor obtained from a standard Cauchy prior on a; this is also the prior that Jeffreys would have used. For testing H_0 versus H_2, CEPs would use a Zellner–Siow prior (Zellner and Siow, 1980) for (a, b) which for this data set is the standard two-dimensional Cauchy density proportional to $(1 + a^2 + b^2)^{-3/2}$. Both priors were suggested by Berger and Pericchi (2001). Calculations using these priors show $p(y_1, y_2|H_0) \approx 2 \times 10^{-5}$, $p(y_1, y_2|H_1) \approx 1 \times 10^{-2}$ and $p(y_1, y_2|H_2) \approx 3 \times 10^{-3}$. Because $p(y_1, y_2|H_1) > p(y_1, y_2|H_2)$, CEPs would claim more evidence for H_1 than for H_2 (as opposed to H_0) and would therefore be incoherent.

Berger states "The most crucial fact about the CEPs ... is that they precisely equal the objective Bayesian error probabilities ...," where "objective" means calculated using an objective Bayesian prior. It is precisely the objective Bayesian prior that makes CEPs incoherent. In both examples, when the hypotheses change, so do the priors. In particular, the prior mass or density assigned to each simple hypothesis changes according to which simple hypotheses are grouped together as a composite. This is in contrast to "subjective" Bayesian priors that reflect beliefs about the simple hypotheses, not the partitions by which they are grouped.

It seems that CEPs will not be attractive to subjective Bayesians or to anyone who values coherence, and they will not provide a methodological unification between such people, Fisherians and Neymanians.

IS METHODOLOGICAL UNIFICATION A GOOD THING?

Berger argues that statisticians' methodological ***disagreement*** has had a negative impact on science. I claim in contrast that ***disagreement*** is a good thing and we should not seek unification. The reason is that when statisticians formulate theories of hypothesis testing, they should distinguish between several possible scientific goals, any of which might be the scientist's aim in a particular application. A scientist comparing two composite hypotheses might be looking for any of the following.

1. *In which hypothesis does the truth lie?* Accepting for the moment that we can best distinguish among simple hypotheses by their likelihood, this question is loosely answered by finding regions of high likelihood in both the null and alternative hypotheses. A more severe answer is given by finding the likelihood ratio statistic and the maximum likelihood estimators in both hypotheses.
2. *Averaging over all the simple hypotheses in each composite, which composite explains the data better?* Accepting for the moment that we can specify

a prior (a weighting distribution for the purpose of averaging), this question is answered by Bayes factors.

3. *Summing or integrating over all the simple hypotheses in each composite, which composite has the most support?* Accepting again that we can specify a prior, this question is answered by posterior probabilities.
4. *Are our data in sufficient disaccord with H_0 that we should consider alternatives?* Setting aside the question of whether this is really a decision problem that requires specification of a loss function, perhaps this is the question being answered by testing H_0 without an explicit alternative.

By treating hypothesis testing as a single type of statistical problem, statisticians ignore that scientists may have these or other, fundamentally different questions which they, or we, call by the general term "hypothesis test." There is no reason to suppose that the different questions are all best answered by the same methodology and every reason to suppose the opposite. As a profession we are guilty of failing to ascertain what our clients want, of forcing too many of their disparate questions into the general rubric of hypothesis testing and of inappropriately applying our favorite methodology (different for different statisticians) to all of them. Perhaps some of the discomfort in the general scientific community with hypothesis tests is due to a vague or unconscious understanding that our methodolgy does not always answer the right question. If that is true, then seeking a unified methodology is counterproductive. We should instead emphasize the ***disagreement*** by distinguishing the different questions that different methodologies are targeted to answer.

Comment

Subhash R. Lele

Discussing a philosophical paper is always a tricky business; one has to balance between devil's advocacy and preaching to the choir. I am afraid I am going to play more the role of a devil's advocate than that of a preacher. I would like to commend Professor Berger for his valiant efforts in trying to unify the three competing schools of statistical thoughts. Unfortunately, I feel we are as far away from achieving that goal as the physicists are in formulating a satisfactory unified field theory. Moreover, I wonder if such a unification of approaches in statistics is even needed.

1. Professor Berger starts by saying that "while the debates over interpretation can be strident, statistical practice is little affected as long as the reported numbers are the same." He further puts a caveat that "We focus less on 'what is correct philosophically?' than on 'what is correct methodologically?' " Unless I am totally misunderstanding his words, I think the real crux of any statistical analysis is in the interpretation of the numbers. That is where science is conducted. How can something be incorrect philosophically, but correct methodologically? Just as numerical values without proper units are meaningless, output of a statistical procedure is meaningless without proper interpretation. I recall heated debates that ensued in the foundations of finite population sampling exactly on this point: what is correct inferentially versus what is correct probabilistically? Just because the numbers are the same, it does not mean one should be making the same scientific inference.

2. Professor Berger uses the p-value as a measure of the strength of evidence in the data. I am sure he is well aware of the arguments against this. See Royall (1997) for references as well as further discussion on this point. An important point (Royall, 1997; Lele, 1998) is that the strength of evidence is a comparative concept. Hacking (1965) suggested the use of the likelihood ratio as a measure of strength of evidence. This can be further generalized to achieve model robustness and outlier robustness, as well as the possibility of handling nuisance parameters, through the use of "evidence functions" (Lele, 1998). I am curious to know why Professor Berger uses controversial p-values instead of the likelihood ratio as a measure of the strength of evidence.

3. I wonder, when CEPs are interpreted as probabilities of a particular hypothesis being correct, are we im-

Subhash R. Lele is Professor, Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2G1 (e-mail: slele@ualberta.ca).

plicitly assuming that one of the hypotheses has to be correct? What if the truth is not contained in either of the hypotheses? Are we choosing the closest hypothesis in some sense?

4. It is clear that one can use various conditioning statistics (Section 5.2 and point 2 above) that purport to measure the "strength of evidence." How do we choose among the myriad of these statistics? If we follow Professor Berger's ideas, would we not end up in the same quagmire where we ended up when we tried to use ancillary statistics for conditioning?

5. One of the major uses of the power function of a test procedure is in deciding the sample size and the experimental design. How is the testing framework defined in this article useful toward this goal? Does a unification of the three schools exist in this context?

6. Should we not be interested more in the effect size and the confidence/credible intervals? Professor Berger seems to suggest that there is less of a problem in unifying these ideas than in unifying the testing framework. Why not solve the easier problem that is also more useful?

7. Professor Berger discusses the possibility of including the no-decision zone in his framework. This is quite closely related to the concept of the probability of weak evidence as introduced in Royall (2000). Of course, calculation of this probability will depend on the question of which experiment to repeat (Lehmann, 1993). I think this is really the crucial question that only scientists can address.

8. Calculation of p-values does not depend on the alternatives, but in Section 3.3 it seems that an explicit alternative is needed to compute it. Is it implicit in this article that the ratio of the p-values is a measure of strength of evidence? Does it not integrate over all other possible data that one could have observed? Would this not violate the likelihood principle?

9. As a side note, I would also like to point out that there seem to be some similarities between Professor Berger's approach and the approach discussed by Mayo (1996).

After reading the last sentence, "This is because the most direct route to determination of a suitable conditional frequentist test, in a given scenario, is the Bayesian route" (and, probably building upon my prior beliefs), I felt that Professor Berger is starting with an answer, that the Bayesian approach is inherently correct, and then trying to modify the frequentist approach so that it provides the Bayesian answers. I feel instead that we should start anew. Our first order of business should be the question of proper quantification of the evidence in the data. Once such quantification is attained, we should think in terms of decision-making procedures that may or may not involve subjective judgements. Since any decision-making procedure is bound to make errors, we should think in terms of reporting and controlling such errors, namely, probabilities of misleading evidence and weak evidence. These may then be used to design experiments effectively.

I would like to thank Professor Mark Taper for useful comments. I would also like to thank Professor Roger Berger for giving me the opportunity to comment on this interesting paper. This work was partially supported by a grant from the National Science and Engineering Research Council of Canada.

Comment

Deborah G. Mayo

1. INTRODUCTION

When two or more methodologies of inference recommend different appraisals of evidence there are two broad avenues that might be taken to "reconcile" them. The first is to take the position that the conflicting methodologies operate with distinct aims, goals, and assumptions, and that rather than insist that one or both be emended to achieve unified appraisals, one should articulate the intended interpretations of the key components of each so as to avoid confusions, and clarify the different domains in which one or the other methodology might be most appropriate. A second position is to maintain that one of the methodologies, say M_1, is superior or more plausible to that of alternative M_2, and advocate modifying M_2 so that it comes into line with M_1. However, it is equally open to adherents of M_2 to judge M_1 by dint of the standards of M_2.

Deborah G. Mayo is Professor, Department of Philosophy, Virginia Tech, Blacksburg, Virginia 24061-0126 (e-mail: mayod@vt.edu).

So in seeking a reconciliation by means of the second strategy, there is always a great danger of being guilty of begging the question against the alternative (often unwittingly). Without a careful scrutiny of equivocal terms, underlying assumptions and the rejoinders open to the alternative methodology, the fallacious nature of such arguments might go unnoticed and might even be celebrated as promoting the basic tenets of the rival school. Berger's reconciliation attempt falls under the second type of strategy. As a philosopher of statistics, my role will be to sketch some of the logical and epistemological concerns that a full scrutiny and response would require.

2. TERMINOLOGY: FREQUENTIST ERROR PROBABILITIES AND ERROR STATISTICS

Berger has so thoroughly coopted the terms of the Neyman–Pearson school, one hardly knows what terms are left to fall back on to articulate the equivocations. Since some such terminological distinctions are necessary, let us agree, for the purposes of this note at least, to retain the usual meaning of ***error probability*** as defined in Neyman–Pearson (N–P) statistical testing. Having been criticized by so many for its insistence that its error probabilities apply, not to statistical hypotheses, but only to procedures of testing (and estimation), one would think that school had earned the right to at least this much!

The probabilities of Type I and Type II errors, as well as p-values, are defined exclusively in terms of the *sampling distribution* of a test statistic $d(\mathbf{X})$, under a statistical hypothesis of interest—the familiar tail areas. In contrast, Berger's CEPs refer to the posterior probabilities of hypotheses under test, H_0, conditional on the observed $d(\mathbf{x})$. I limit error statistical accounts to those where probabilities are derived from sampling distributions (including both N–P and Fisherian significance tests).

3. HOW WOULD AN ERROR STATISTICIAN APPLY BERGER'S UNIFICATION?

It is not clear just what Berger recommends as a replacement for the current N–P and Fisherian tests, especially as his examples seem to fall outside both representatives of the error statistical paradigm. N–P tests require H_0 and H_1 to exhaust the space of hypotheses (within an assumed model), Fisherian tests are defined with only the null, and both approaches deliberately operate without prior probability assignments to hypotheses. We get some guidance from Example 1 and the discussion of the "applet" (Section 2.2). Clearly Berger intends it to show, even to a thoroughgoing error statistician, that there is something wrong with p-values, at least when used as data-dependent measures of evidence (without a CEP "correction"). In a Normal distribution test of $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$, "at least 22%—and typically over 50%—of the corresponding null hypotheses will be true" if we assume "half of the null hypotheses are initially true," conditional on a 0.05 statistically significant $d(\mathbf{x})$. Berger takes this to show it is dangerous to "interpret the p-values as error probabilities" (Section 2.3), but note the shift in meaning of "error probability." The alleged danger assumes the *correct* error probability is given by the proportion of true null hypotheses (in a chosen population of nulls), conditional on reaching an outcome significant at or near 0.05 (e.g., 22 or 50%). But why should an error statistician agree that the Bayesian definitions Berger advocates replace the error statistical ones?

Innocence by Association

Let us see how the error statistician is to proceed if he or she were to take seriously the apparent lesson of Berger's applet. A recent study that has gotten a great deal of attention reports statistically significant increases in blood clotting disorders and breast cancer among women using hormone replacement therapy (HRT) for 5 years or more. Let us suppose p is 0.02. The probability of observing so large an increase in disease rates when H_0 is true and HRT poses no increased risk is 0.02. Given the assumptions of the statistical model are met, the error statistical tester takes this to indicate a genuine increased risk, for example, approximately 2 additional cases, of breast cancer per 10,000 (higher, if HRT is taken for 10 years). Berger warns us that such low p-values may be overstating the evidence against the null (the discrepancy between p-values and CEPs increases with sample size and here there were over 16,000 women). To check this, it would seem, we must go on to consider a pool of null hypotheses from which H_0 may be seen to belong, and find the proportion of these that have been found to be true in the past. This serves as the prior probability for H_0. We are then to imagine repeating the current significance test over all of the hypotheses in the pool we have chosen. Then the posterior probability of H_0 (conditional on the observed result) will tell us whether the original assessment is misleading. But which pool of hypotheses should we use? Shall we look at all those asserting no increased risk or benefit of any sort? Or no

increased risk of specific diseases (e.g., clotting disorders, breast cancer)? In men and women? Or women only? Hormonal drugs or any treatments? The percentages "initially true" will vary considerably. Moreover, it is hard to see that we would ever know the proportion of true nulls rather than merely the proportion that have thus far not been rejected by other statistical tests!

Further, even if we agreed that there was a 50% chance of randomly selecting a true null hypothesis from a given pool of nulls, that would still not give the error statistician a frequentist prior probability of the truth of *this* hypothesis, for example, that HRT has no effect on breast cancer risks. Either HRT increases cancer risk or it does not. Conceivably, the relevant parameter, say the increased risk of breast cancer, could be modeled as a random variable, but its distribution would not be given by computing the rates of other apparently benign or useless treatments! Berger's Bayesian analysis advocates a kind of "innocence by association," wherein a given H_0 gets the benefit of having been drawn from a pool of true or not-yet-rejected nulls. Perhaps the tests have been insufficiently sensitive to detect risks of interest. Why should that be grounds for denying there is evidence of a genuine risk with respect to a treatment (e.g., HRT) that *does* show statistically significant risks? (The dangers are evident.)

The Assumption of "Objective" Bayesian Priors

Admittedly, Berger does not really seem to be recommending error statisticians calculate the above frequentist priors, but rather that they assume from the start the "objective" Bayesian prior of 0.5 to the null, the remaining 0.5 probability being spread out over the alternative parameter space. But seeing how much this influences the Bayesian CEP, which in turn licenses discounting the evidence of risk, should make us that much more leery of assuming them from the start. One can see why the Bayesian significance tester wishes to start with a fairly high prior to the null—else, a rejection of the null would be merely to claim that a fairly improbable hypothesis has become more improbable (Berger and Sellke, 1987, page 115). By contrast, it *is* informative for an error statistical tester to reject a null, even assuming it is not precisely true, because we can learn how false it is. (Some people deny point nulls are ever precisely true!) So, why should error statisticians agree to the charge that their interpretation of evidence is flawed because it disagrees with those based on a priori assumptions appropriate for a form of inference error statisticians reject?

4. BERGER'S FREQUENTIST PRINCIPLE

Berger's CEPs satisfy something he calls the frequentist principle (FP), which he alleges is Neyman's principle, only stated "in the form that is actually of clear practical value" (Section 2.1). The essential function of N–P tests, for Neyman, however, was to control at small values the probabilities of taking the action (or reaching the inference) associated with "reject H_0" when H_0 is true (Type I error) and at the same time control as far as possible the probability of taking the action associated with "accept H_0" when H_0 is false (Type II error), where accept H_0 may be regarded as a shorthand for "no evidence against H_0" (Neyman, 1976, page 749). Additionally, this error control has to hold regardless of prior probability assignments. Berger's FP, however, does not require controlling errors at small values and is highly dependent on prior probability assignments. So far as I can see, the only "clear practical value" of saddling Neyman with this vaguely worded principle (wherein the meaning of error rates is allowed to shift) is to convince us that CEPs satisfy the N–P error statistical philosophy. But they do not.

Berger has a problem regarding observed significance levels or p-values as legitimate error probabilities—perhaps because they are not predesignated—but neither Neyman nor Pearson felt this way. Neither, for that matter, did Fisher, only he denied that low error rates captured the essential justification of significance test reasoning for scientific, as opposed to "acceptance sampling," contexts (Fisher, 1973, page 45). The low p-value in the HRT study led to reject H_0 and assert that there is evidence of genuine (and quantifiable) increased risks in women taking HRT for 5 years or more. Only $p\%$ of such trials would lead to supposing one had gotten hold of a real, repeatable, effect were the outcomes merely the result of the chance assignments of HRT or placebo groups. The error probability holds whether it refers to an actual or hypothetical series of tests, and it is this hypothetical reasoning that matters for Fisher—but also for Pearson (and even for Neyman, in his inferential moods).

Turning the tables for a moment, consider how an error statistician might evaluate the error probabilities associated with *Berger's* procedure: construing a 0.05 significant result as little or no evidence of a discrepancy from the null hypothesis. The error of relevance

would be akin to a Type II error—denying there is evidence against H_0 when H_0 is false—and the probability of this error would be very high. [He would also fail to ensure low CEPs! Having calculated the CEP is 0.2 or 0.5, the Type II CEP would be 0.8 or 0.5. That is, 80 or 50% of the hypotheses (in the pool of nulls) would be false, when Berger has found little evidence against the null.] More generally, faced with conflicts between error probabilities and Bayesian CEPs, the error statistical tester may well conclude that the flaw lies with *the latter* measure. This is precisely what Fisher argued.

Fisher: The Function of the p-Value Is Not Capable of Finding Expression

Discussing a test of the hypothesis that the stars are distributed at random, Fisher takes the low p-value (about 1 in 33,000) to "exclude at a high level of significance any theory involving a random distribution" (Fisher, 1973, page 42). Even if one were to imagine that H_0 had an extremely high prior probability, Fisher continues—never minding "what such a statement of probability a priori could possibly mean"—the resulting high posteriori probability to H_0, he thinks, would only show that "reluctance to accept a hypothesis strongly contradicted by a test of significance" (ibid, page 44) ... "is not capable of finding expression in any calculation of probability a posteriori" (ibid, page 43). Indeed, if one were to consider the claim about the a priori probability to be itself a hypothesis, Fisher suggests, it would be rejected by the data.

The Core Principles of Error Statistics Are Not Satisfied

Thus, we must deny Berger's unification lives up to his claim to satisfy "the core principles" of error statistics. The two key claims he cites as attractive selling points of his unified account make this clear.

1. We need no longer warn students and researchers off supposing that "a frequentist error probability is the probability that the hypothesis is true" (Section 4.4.1), Berger assures us, since the CEP *is* the (objective) Bayesian posterior. This may be good news for those who concur that error probabilities "can never be counterintuitive [if they] coincide with objective Bayesian error probabilities" (Section 5.2), but what if we do not?

2. Error statisticians (who accept his unification) are also free from the complications of having to take into account the stopping rule used in sequential tests. Although a procedure that "tries and tries again," stopping only when a chosen small p-value is computed, is known to alter the N–P Type I error rate, CEPs are not affected. Pretty clearly then the unification is not controlling N–P error probabilities (Mayo and Kruse, 2001).

Berger's exhortations that we should focus not on "what is correct philosophically," but rather on "what is correct methodologically" presupposes the latter does not turn on the former when in fact it does—especially where different methods result in different numbers. Far from restoring the scientific credentials of statistics—a laudatory goal—this way of achieving a reconciliation seems only to be increasing the confusion and misunderstanding of the logic and value of error statistical methods.

5. A POSTDATA INTERPRETATION OF N–P TESTS (BASED ON A CONCEPT OF SEVERE TESTING)

A more promising route is to recognize, as did Egon Pearson, that there are two distinct philosophical traditions from which to draw in evaluating statistical methodology (Pearson, 1966, page 228). In one, probability is used to provide a postdata assignment of degree of probability, confirmation or belief in a hypothesis; in the second, probability is used to assess the probativeness, trustworthiness or severity of a test procedure (e.g., Peirce, Popper). Error statistical testing falls under the latter; Bayesian and other degree-of-confirmation accounts, the former. Nevertheless we can agree with Berger's charge that the central weakness of N–P tests is the need of a postdata assessment to reflect "the variation in evidence as the data range over the rejection or acceptance regions" (Section 2.2). Rather than supplement N–P tests with Bayesian CEPs, however, we can instead supplement them with data-specific assessments of the *severity* with which one or another inference passes. If it is possible, by means of such a severity interpretation of tests, to address the fallacies and "paradoxes" associated with N–P tests while articulating tests from the philosophical standpoint of the error statistician, then surely that would be desirable (Mayo, 1983, 1985, 1991, 1992, 1996, 1997; Mayo and Spanos, 2002). Here I can give only a bare sketch.

6. SEVERITY AND A POSTDATA INTERPRETATION OF N–P TESTS

The notion of severity stems from my attempt to improve upon Popper's philosophy of science. It reflects, I think, our ordinary, day-to-day notion of a probative exam: For a hypothesis or claim H to *have passed a severe test T with data $\boldsymbol{x}$, $\boldsymbol{x}$* must not only be in accord with H; such an accordance must be very improbable if, in fact, it is a mistake to regard $\boldsymbol{x}$ as evidence for H. That is, a hypothesis H passes a severe test T with data $\boldsymbol{x}$ if (**1**) $\boldsymbol{x}$ agrees with H and (**2**) with very high probability, test procedure T would have produced a result that accords *less* well with H than $\boldsymbol{x}$ does, if H were false or incorrect.

Although severity applies to assessing hypotheses in general, here my concern is with statistical tests. Consider test T_α, $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$, in the context of the simple Normal model with known variance, using the usual test statistic, $d(\mathbf{X}) = \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}$. Let $\mu_0 = 0$ and $\sigma = 1$. Since Berger focuses on p-values, let us set out the N–P test by reference to the observed significance level, $p(\boldsymbol{x}_0)$, that is, reject H_0 at level α iff $\mathrm{P}(d(\mathbf{X}) > d(\boldsymbol{x}_0);\ H_0) \leq \alpha$ for a small α, say, 0.03. $\boldsymbol{x}_0$ is the observed value of $\boldsymbol{X}$.

Although N–P tests are framed in terms of hypotheses being rejected or accepted, both correspond to "passing" some hypothesis, enabling a single notion of severity to cover both. "Reject H_0" in T_α licenses inferences about the extent of the positive discrepancy *indicated* by data $\boldsymbol{x} : \mu > \mu'$, whereas "accept H_0" corresponds to inferences about the discrepancies from H_0 *ruled out*, $\boldsymbol{x} : \mu \leq \mu'$, where $\mu' \geq 0$.

The Case of Rejecting H_0

For test T_α , define the severity with which hypothesis $\mu > \mu'$ passes when H_0 is rejected [i.e., $p(\boldsymbol{x}_0) \leq \alpha$]:

$$\begin{aligned}&\mathrm{Sev}\big(T_\alpha;\ \mu > \mu', d(\boldsymbol{x}_0)\big)\\ &\quad = \mathrm{P}\big(d(\boldsymbol{X}) \leq d(\boldsymbol{x}_0);\ \mu > \mu' \text{ is false}\big).\end{aligned}$$

In the special case where $\mu' = 0$, this is identical to 1 minus the p-value, as is plausible, but the assessment now has a clear post-data construal that varies appropriately with changing outcomes, sample sizes, and hypotheses of interest.

Since the primary goal of the severity interpretation is to avoid classic fallacies, as much attention is given to inferences that *are* licensed as to those that are not; any adequate assessment requires a report of both.

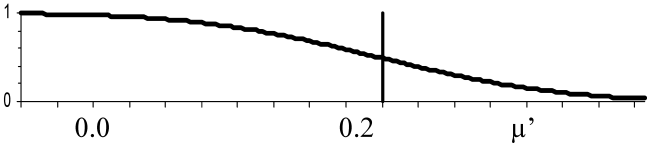

FIG. 1. *Severity with which test T_α passes $\mu > \mu'$ with $d(\boldsymbol{x}_0) = 2.0$.*

Suppose $n = 100$ ($\sigma_{\boldsymbol{x}} = 0.1$). If $d(\boldsymbol{x}_0) = 2$, then T_α rejects H_0 since we set α to 0.03. $H_0 : \mu > 0$ passes with severity about 0.977 [$\mathrm{P}(\overline{X} \leq 0.2; \mu = 0) = 0.977$], but we can also assess how well particular discrepancies pass, for example, $\mu > 0.1$ passes with severity about 0.84 [$\mathrm{P}(\overline{X} \leq 0.2;\ \mu = 0.1) = 0.84$].

In evaluating severity, we are not changing the hypotheses of the original test, but considering different postdata inferences that one might be interested in evaluating. Does $d(\boldsymbol{x}_0) = 2$ provide good evidence that $\mu > 0.2$? No, since our procedure would yield so large a $d(\boldsymbol{x}_0)$ 50% of the time even if μ were no greater than 0.2. One might either report extreme cases of inferences that are and are not warranted, or graph all severity values for a given observed $d(\boldsymbol{x}_0)$ as shown in Figure 1.

It is noticed that we obtain an inference that passes with severity 0.95 were we to form the one-sided 95% confidence interval that corresponds to test T_α (i.e., $\mu > 0.035$). However, confidence intervals (in addition to having a problematic postdata interpretation) treat all values in the interval on par, whereas we see that many (e.g., $\mu > 0.2$, $\mu > 0.3$) pass with low severity.

If the test yields a $d(\boldsymbol{x}_0)$ further into the rejection region, e.g., if $d(\boldsymbol{x}_0) = 4$, a different severity curve results wherein $\mu > 0$ passes with severity about 0.999, $\mu > 0.2$ passes with severity about 0.977, and so on.

Same p-Value, Two Sample Sizes

Although $\mu > 0.2$ passes with severity 0.977 with $d(\boldsymbol{x}_0) = 4.0$ when $n = 100$, suppose instead that $d(\boldsymbol{x}_0) = 4.0$ occurs in T_α with sample size 1600. Hypothesis $\mu > 0.2$ would now pass with nearly 0 severity! More generally, a statistically significant difference (at level p) is less indicative of a given discrepancy from the null the larger the n. The well-known "large n" problem in testing is thereby scotched.

Severity versus Power

At first glance severity may seem to be captured by the notion of a test's power, but in the case where

H_0 is rejected, severity is inversely related to power (as well as being data-dependent, unlike power). For example, with $n = 100$ the severity with which the inference $\mu > 0.4$ passes with $d(\boldsymbol{x}_0) = 2$ is low (i.e., 0.03). However, the power of the test against 0.4 is high, 0.98! The intuition that high severity is entailed by high power is true only when H_0 is *accepted*; that is, high power against μ' ensures that *failing* to reject H_0 warrants $\mu \leq \mu'$.

The Case of Accepting H_0

For test T_α the severity with which hypothesis $\mu \leq \mu'$ passes when H_0 is accepted [i.e., $p(\boldsymbol{x}_0) > \alpha$)] is:

$$\begin{aligned}&\mathrm{Sev}\big(T_\alpha;\ \mu \leq \mu', d(\boldsymbol{x}_0)\big)\\ &\quad = \mathrm{P}\big(d(\boldsymbol{X}) > d(\boldsymbol{x}_0);\ \mu \leq \mu' \text{ is false}\big).\end{aligned}$$

Suppose now that $n = 25$ in test T_α, and $d(\boldsymbol{x}_0) = 0.5$ (i.e., $\bar{X} = 0.1$, p-value $= 0.31$). The classic fallacy is to take a failure to reject the null as evidence of 0 discrepancy from the null: the problem is there are always discrepancies that the test had little capacity to detect. A popular gambit intended to avoid this would note that $d(\boldsymbol{x}_0) = 0.5$ is as close (statistically) to 0 as it is to 0.2—the "counternull" hypothesis (Rosenthal and Rubin, 1994). But this is yet to tell us which discrepancies we *are* warranted in ruling out, whereas severity does. In particular, $\mu \leq 0.4$ and $\mu \leq 0.5$ pass with high severity (0.93 and 0.97).

The post-data severity assessments are still based on standard error probabilities, but they are evaluated relative to the observed value of the test statistic. Viewing N–P tests from the severe testing perspective, we see that the real value of being able to control error probabilities at small values is not the desire to have a good track record in the long run—although such a long-run justification is still available (and in several contexts may be perfectly apt). It is, rather, because of how this lets us severely probe, and thereby understand correctly, the process underlying the data in front of us.

ACKNOWLEDGMENT

I thank Professor Aris Spanos for many useful comments.

Comment

Luis R. Pericchi

Professor Berger, in his 2001 Fisher Lecture, proposes a new foundation to the ubiquitous subject of precise hypothesis testing. The thrusts of the article are both philosophical and scientific, proposing a collective search that will hopefully embrace statisticians from the main philosophical schools, and one that incorporates fundamental elements of each. The emerging new synthesis would produce methodology and measures of evidence over which statisticians might agree and that, most importantly, will move scientific practice to a superior level, improving the service of statistics to science on the subject of testing, on which arguably there is most confusion and malpractice. My view is that the author makes a compelling case for the simplest case of testing: a simple hypothesis versus a simple alternative. Far from solving all open questions for more complicated situations, the author puts forward an eclectic research program which is both promising and timely, and which will be beneficial for statistics as a whole and for science as a consequence. The building block is the conditional viewpoint of statistics and Jeffreysian measures of error are the key to choose on which statistics to condition. If this program is successful in this most controversial subject, a new face of statistics will emerge in the form of a superior synthesis, with an enlarged Bayesian and particularly Jeffreysian component, but one that will accommodate Fisherian and Neymanian aspects as well. I will elaborate on some points. Overall I envision that what Professor Berger is proposing at this point will be theory and practice in the not so far future.

Luis R. Pericchi is Professor, Department of Mathematics and Computer Science, University of Puerto Rico, Rio Piedras, Puerto Rico 00931-3355 (e-mail: pericchi@goliath.cnnet.clu.edu).

TWO HISTORICAL AGREEMENTS

Shortly after the final edition of Jeffreys' (1961) *Theory of Probability* book, Welch and Peers (1963) produced an agreement: They proved that for univariate

regular parametric likelihoods, the Jeffreys rule for selecting an objective (typically improper) prior for estimation is the optimal prior in the sense that confidence level and posterior probability for intervals will agree fastest as the sample size grows. This result set up an agenda for eclectic work in more complex situations, and the now established theory of frequentist matching priors is still being developed. Whereas, for multidimensional priors, it is not the Jeffreys rule which gives the optimum, but refinements of them like reference and other priors, the result gave an unexpected new justification of the Jeffreys rule in unidimensional situations (originally developed by Jeffreys on parameter invariance requirements) and for objective Bayesian thinking in general. A second agreement was reported by Berger, Brown and Wolpert (1994), who assumed a conditional frequentist viewpoint for simple versus simple testing and showed that by conditioning on the likelihood ratio (equal to the Bayes factor in this setting), the conditional frequentist error probabilities equal the Bayesian posterior probabilities of error. As with Welch and Peers, this result set up an agenda for new agreement: even if it is bound to require subtleties and modifications for more general situations, the agenda is there. Subsequent articles have advanced the theory somewhat, but there is still a lot to be done. I believe that the agenda will gain momentum for two reasons. First, the disagreement in testing is far more important than in interval estimation. Second, because the schools of statistics now talk much more to each other and fight less.

EFFECT OF THE DISAGREEMENT

A lot has been written to criticize unconditional tests (as summaries of evidence) and p-values (as error probabilities) as compared with too little that has been made, particularly in textbooks that are actually used at an elementary level, to improve these widespread practices. I believe that the reasons for this are basically (1) the nonexistence of general methodological alternatives and, importantly, (2) what Professor Berger calls the disagreement. (The second reason may be claimed to be even more important, since it may prevent a collective search for general alternatives and keep stock with the old practices.) Other suggestions for agreement have been discussed, for example, by Cox and Hinkley (1974), that either change the prior or change the α-level of the test with the sample size. But Professor Berger's proposal seems to be the first to incorporate simultaneously essential concepts from the three main schools of statistics.

ASPECTS OF THE LECTURE

The aim of the Lecture appears to be nothing less than to change the whole subject of statistical testing of a precise hypothesis by all schools of statistical thought via a new and superior synthesis. Only if this is achieved can the general practice of statistical evidence of scientific hypothesis be changed, yielding conclusions on which statisticians and—later, perhaps much later—scientists may agree. An avenue of basic agreement must be found prior to the possibility of changing widespread inferior practice.

The basic point of view is that the conditional viewpoint of statistics is correct, both from a Bayesian and from a frequentist perspective. It is, therefore, the fundamental concept that may yield a unifying, all-embracing method for agreement in general methodology, but that in the case of a precise null hypothesis has a most difficult trial. In fact, Professor Berger called the conditional Bayesian approach the most useful viewpoint in his book (Berger, 1985a). Bayesian is certainly not the only conditional approach, but it is the most structured one and thus the least complicated to implement. This fact, outlined in Berger, Brown and Wolpert (1994) and masterfully discussed in this Lecture, is a landmark that opens up a promising avenue for unification on this (conceptually) difficult problem.

The style of the Lecture is the style attributed to Hegel by Goethe: To place yourself in the horizon of your adversaries and refute them with their own words and concepts. In the important and ubiquitous scientific problem of testing a precise hypothesis, the concepts are more important than the conclusions; this differs from estimation. It has to differ if unification is deemed possible, since the conclusions in testing are in fundamental disagreement. The adversaries are to be found, however, in several camps: pure Fisherian, Neymanian and even Bayesians.

An original mathematical aspect of the Lecture is to condition on the maximum of p-values, rather than the traditional motivation of conditioning on the value of the likelihood ratio. This makes it even more attractive and compelling in the simple versus simple hypothesis case.

I anticipate that immediate effects of this Lecture will be to influence Bayesians and to give more credibility to Bayes factors. Some Bayesians at some point

TABLE 1
The conditional test at work, showing sensible decisions and conditional error probabilities which are equal to the posterior probabilities of the rejected model

		Unconditional			Conditional			p-values	
$\bar{x} = 1.645/\sqrt{n}$	n	Decision	α	β	Decision	α_c	β_c	p_0	p_1
0.74	5	Reject	0.05	0.28	Reject	0.23		0.05	0.28
0.52	10	Reject	0.05	0.06	Reject	0.45		0.05	0.06
0.42	15	Reject	0.05	0.013	Accept		0.23	0.05	0.013
0.37	20	Reject	0.05	0.002	Accept		0.07	0.05	0.002
0.33	25	Reject	0.05	0.000	Accept		0.01	0.05	0.000

in their lives have misunderstood the so-called Lindley paradox. The reasoning, even made by quite a few Bayesians, is that the Lindley paradox is caused by the often unjustifiable assumption of a positive probability of a set of measure zero—in this case the null hypothesis. Thus, the reasoning goes, the Bayes factor is at fault and should be replaced by a different statistic with a different asymptotic behavior. (The disagreement is also present in Bayesian quarters!) This reasoning is simply wrong. A simple counterexample is the following: Assume a random Normal sample with variance 1 and $H_0 : \text{mean} = 0$ versus $H_1 : \text{mean} = 1$.

Here the prior is the natural prior $\mathrm{P}(H_0) = \mathrm{P}(H_1) = 1/2$. The conditional test (see Table 1) and the Bayes factor naturally select H_0 iff $\bar{x} < 1/2$, the midpoint between the hypotheses. The conditional test appears to be more sensible from both frequentist and Bayesian outlooks. This is Lindley's paradox: As evidence accumulates, Bayesian posterior probabilities (and now also conditional frequentist testing) will tend to differ more with unconditional testing both in the decision taken and the errors reported. Furthermore, it is clear in the example that it is the conditional frequentist and Bayesian posterior probability output which make sense. Certainly this is a simplistic example, but if an approach fails in the simplest of examples, is it suspect in a more complex situation?

Bayesian positions might be classified broadly into two categories (still another classification of Bayesians): those who envision that the world eventually will be 100% Bayesian and those who more humbly (and certainly with more political instinct) insist that a new synthesis (sometimes called a compromise) will be the world of the future. This new synthesis, is bound to have a fundamental, if not dominant, Bayesian basic reasoning, but with varied frequentist aspects incorporated. This Lecture gives substantial weight to the latter position. In itself, this Lecture gives partial proof that a new and still unwritten synthesis is the future.

SOME OPEN PROBLEMS

I mentioned before that a main reason for lack of a conditional replacement for unconditional frequentist testing and p-values is the nonexistence of a general conditional alternative. Since the shortest route to this alternative is the development of general approaches to objective priors for testing, it follows that developing objective priors for Bayes factors is the most serious hurdle. Once resolved, conditioning on them would be the natural extension of this line of work. Recent developments in intrinsic, fractional and EP priors seem to me to be the most natural candidates to pursue. Another point that I have not mentioned is the consequences that conditional testing will have in the stopping rule principle. I think that extensions to sequential testing in general settings are bound to be more difficult to foresee at this point. One reason for this is that, even though for precise hypotheses the resulting conditional test obeys the stopping rule principle, accepted priors as reference priors in nonsequential settings actually depend on the stopping rule. Thus sequential testing under this fresh thinking, masterfully exposed by Professor Berger, is going to be very exciting!

A FINAL COMMENT

There seems to be some room for approximations here. Schwarz's Bayesian information criterion (BIC) is a rough approximation to Bayes factors, and is simple and easy to use. For the class of problems on which the BIC is safe, using it instead of a proper Bayes factor may produce an easily reachable substantial extension of the method.

Comment

N. Reid

Methods to reconcile different approaches to inference are needed and much welcomed. As researchers in substantive areas increasingly use more complex statistical analyses, and become more knowledgeable in statistics, the perceived discord between Bayesian and frequentist approaches seems to loom larger than it does in the statistical community. I think this causes confusion and frustration among researchers, much as the introduction to inference causes confusion and frustration among students. In my view, the most promising route to a compromise is to derive Bayesian inference procedures that can also be justified by their behavior in repeated sampling from the model. This article outlines such a route in the context of testing a null hypothesis. Recent work on so-called matching priors attempts to identify Bayesian posterior probability intervals that also have an interpretation as confidence intervals.

The Neyman approach to testing is a mathematical device designed, I believe, to generate test statistics for highly structured settings and should not be used as a basis for inference in a particular problem. It is unfortunate that there are still elementary textbooks recommending rejection of null hypotheses at level 0.05, but I do not believe Neyman would have used his work in such a prescriptive and dogmatic way. On the other hand, p-values have always seemed to me a sensible and constructive way to assess whether or not the data at hand are broadly consistent with a hypothesized model. The fact that they may be incorrectly interpreted by a broader public is dismaying, but not necessarily their death knell. In recent work with Don Fraser, we emphasized the role of the p-value regarded as a function of a parameter of interest in constructing confidence intervals at any desired level of confidence. In fact I think that the main role of testing should be in providing confidence sets or regions through inversion, although I appreciate that many practitioners do not use p-values, or testing, in this way.

For these reasons I was personally interested in the prospect of Fisher/Jeffreys agreement and less concerned with the possibility of constructing conditional Type I errors. In fact I do not agree with Berger that p-values are misinterpreted as Type I errors, although I do believe they are misinterpreted as posterior probabilities that the null hypothesis is true. Thus a method that more closely aligns p-values with posterior probabilities is to be welcomed.

As far as I can tell, though, the Fisher/Jeffreys agreement is essentially to have Fisher acknowledge Jeffreys was correct. In the highly artificial case of testing a simple null hypothesis against a simple alternative, Berger argues that the posterior probability of the null can be interpreted as the maximum of two p-values. While an interesting link, it seems difficult to generalize this to more realistic settings. It is argued here and in related work that one solution is to recalibrate p-values so that they have an interpretation as posterior probabilities; in the absence of other information about the alternative, to use the recalibration that replaces the p-value with $\alpha(p) = \Pr(H_0|x)$. It seems to me that this means that the result of applying T^C to Example 1 is to replace the p-value of 0.021 or 0.0037 by the posterior probabilities 0.28 or 0.11, or perhaps some probabilities slightly different, obtained by using a slightly different prior, but roughly the same order of magnitude. Extending the method to composite null hypotheses seems to require basically a full Bayesian analysis, and the connection to p-values becomes even more remote.

N. Reid is Professor, Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3 (e-mail: reid@utstat.toronto.ca).

Rejoinder

James O. Berger

I enjoyed reading the discussions and am grateful to the discussants for illuminating the problem from interestingly different perspectives. Surprisingly, there was little overlap in the comments of the discussants, and so I will simply respond to their discussions in order. As usual, I will primarily restrict my comments to issues of disagreement or where elaboration would be useful.

RESPONSE TO PROFESSOR CHRISTENSEN

Christensen argues for the Bayesian and likelihood approaches to testing when one has an alternative hypothesis, and I do not disagree with what he says. Indeed, one of the purposes of this article was to show that frequentists, through the conditional approach, can also enjoy some of the benefits of better interpretability to which Christensen refers.

Christensen mostly discusses the interesting issue of model validation when a parametric alternative hypothesis is not available. In Section 5 of the article, I discussed two ways to approach this problem, designed to overcome the difficulty of seemingly having to depend on p-values in such situations. Christensen also notes the difficulty in choosing a test statistic for model validation; see Bayarri and Berger (2000) for relevant discussion on this point.

RESPONSE TO PROFESSOR JOHNSON

Johnson reminds us that, in many problems such as screening tests, it is not uncommon for nulls to be true—even when their p-values are small—because of the magnitude of the prior probabilities of hypotheses that are typically encountered in the area. This is indeed important to keep in mind, but the misleading nature of p-values is apparent even if hypotheses have equal prior probabilities.

Johnson next mentions an interesting problem in risk analysis in which the null hypothesis is composite and the alternative is simple. As briefly mentioned in Section 5.4, handling this within the testing framework of the article would require use of a prior distribution on the composite model and would result in the posterior probability of the null being equal to an "average conditional Type I error." Use of an average Type I error is not common frequentist practice, so adoption of the suggested procedure by frequentists, in this situation, would likely be problematical. Of course, reporting Type I error as a function of the parameter is not common either and is not practically appealing. (Johnson's example is one in which taking the sup of the Type I error over the null parameter space would also not be practically appealing.) If one did so, it would seem necessary to indicate which parameter values were deemed to be of particular interest, and it is then a not-so-big step to write down a distribution (call it a prior or a weighting function) to reflect the parameters of interest, implement the conditional frequentist test and report the average Type I error.

Johnson also raises the important issue that the misleading nature of p-values, from a Bayesian perspective, becomes more serious as the sample size increases. One nice feature of the conditional frequentist approach is its demonstration of this fact purely from the frequentist perspective (since the conditional frequentist Type I error probability equals the Bayesian posterior probability). Johnson wonders if this can also be applied to a regression model with large sample size and 20 covariates. The answer is, unfortunately, no, in that efforts to develop an analog of the conditional frequentist testing methodology for multiple hypotheses have not been successful. Indeed, Gönen, Westfall and Johnson (2003) indicated one of the problems in attempting to do this, namely, the crucial and delicate way that the prior probabilities of the multiple hypotheses can enter into the analysis.

Johnson reminds us that, while objective statistical methodology certainly can have its uses, we would often be better off to embrace the subjective Bayesian approach in practice. I agree, although my own practical experience is that a mixed approach is typically needed; it is often important to introduce some subjective information about key unknowns in a problem, but other unknowns have to be treated in a default or objective fashion.

RESPONSE TO PROFESSOR LAVINE

Lavine presents several interesting examples related to the incoherence of objective Bayesian testing, when "objective" is defined to mean, for instance, that each hypothesis is given equal prior probability. Incoherencies can then arise when one of the hypotheses is a union of other hypotheses, and these hypotheses are subsequently tested separately, without the prior mass for the original hypothesis being divided among the subhypotheses.

Within objective Bayesian testing, this is not a serious practical problem, in that it is understood that objective Bayesians may need to be more sophisticated than using the naive "equal prior probability of hypotheses" assumption (in much the same way that it is well understood that always using a constant prior density for parameters is not good objective Bayesian practice). Alas, the "cure for incoherency" for conditional frequentist testing is not so simple and, indeed, may not be possible. This is because the frequentist–Bayesian unification for testing two hypotheses seems to work well only with equal prior probabilities of hypotheses (see Berger, Brown and Wolpert, 1994) and, as mentioned earlier, effectively dealing with more than two hypotheses in the conditional frequentist testing paradigm has proven to be elusive. My current view on this issue is that the conditional frequentist approach eliminates the greatest source of incoherency in frequentist testing and hence is much better in practice, but does not eliminate all incoherency.

Lavine asks, "Is methodological unification a good thing?", and suggests that it is not. However, he is referring to the issue that there can be a variety of conceptually quite different testing goals and that each separate goal might require a different analysis. This is very different from saying that, for testing with a particular goal in mind, it is okay to have methodologies that yield very different answers; this last, I argue, is highly undesirable for statistics. Now it could be that each of the different testing methodologies is the right answer for one of the particular testing goals, but I do not think so. Thus, even accepting Lavine's thesis that there are four distinct testing scenarios, I would argue that each should ideally have its own unified testing methodology.

RESPONSE TO PROFESSOR LELE

Lele suggests that the unification of having different statistical approaches produce the same numbers is not satisfactory, when the interpretations of these numbers are quite different. As a general point this might be true, but let us focus on the unified testing situation: The conditional frequentist will choose to interpret an error probability of 0.04 in terms of a long-run frequency and the Bayesian, in terms of posterior probability. Producing the same number simply means that either interpretation is valid for the given test. My view is that inferential statements that have two (or more) powerful supporting interpretations are considerably stronger than inferences that can be justified only from one perspective.

Lele is concerned with the use of p-values to measure the "strength of evidence in the data" and refers to some of the many arguments in the literature which indicate that p-values are poor measures of evidence. Indeed, perhaps the primary motivation for this article is precisely that p-values are poor measures of evidence about the comparative truth of hypotheses, which is what is addressed in the literature to which Lele refers. In this article, p-values are used in a quite different fashion, however—not to compare hypotheses, but rather to measure the strength of the generic information content in the data within a specific test: Saying that data for which $p_0 = 0.04$ has the same generic strength of evidence as the data for which $p_1 = 0.04$, in a specific test under consideration, is a comparatively mild evidential statement. (This is like saying, in estimation of a normal mean μ, that the strength of evidence in the data is measured by $S/\sqrt{n}$; it says nothing directly about μ, the quantity of interest.) In response to another of Lele's questions, the ratio of p-values has no role in the analysis.

Of course, Lele is correct that other measures of strength of evidence in the data, such as likelihood ratio, could be used to develop conditioning statistics. Indeed, I mentioned a variety of these possibilities in Section 5.2. I specifically did mention Birnbaum's attempt to use likelihood ratio to define a conditioning statistic and I pointed out that it often fails to give satisfactory answers, as Birnbaum himself noted. (Likelihood ratio is a great measure of the comparative support that the data has for hypotheses, but fails to provide adequate conditioning statistics in the conditional frequentist paradigm.) Lele further asks how to choose from among the myriad possible conditioning statistics. The main point of the article is that one should use the p-value conditioning statistic, because it is the only choice that achieves the unification of viewpoints.

Here are answers to a number of Lele's other questions.

- The development of conditional error probabilities implicitly assumes that one of the hypotheses is correct. Bayesian testing can be given an interpretation in terms of which hypothesis is closest to the true hypothesis, but I do not know of any such interpretation for conditional frequentist testing.
- Dass and Berger (2003) indicated how sample size and design questions should be addressed in the conditional frequentist framework. Central is the notion that one should design so as to achieve conditional frequentist (or Bayesian) inferential goals.
- There is already a vast literature on unification of frequentist and Bayesian confidence sets, as mentioned in the discussions by Pericchi and Reid, so there was no reason to look at this problem first, as Lele proposes.
- The use of the alternative hypothesis, in our definition of p-values, is limited to utilization of the likelihood ratio test statistic to define the p-values.
- Since the proposed conditional frequentist error probabilities equal the objective Bayesian posterior probabilities of hypotheses, they clearly are compatible with the likelihood principle. However, there is a slight violation of the likelihood principle in that the critical value for the test will depend on the full sampling models under consideration. This has very little practical import, however, in that the CEPs for data near the critical value will be large, leading to the clear conclusion that there is no substantial evidence in favor of either of the hypotheses for such data.
- Lele suggests that the unification achieved here is simply an attempt to modify frequentist theory so that it agrees with Bayesian theory. That is not an accurate characterization, in that unification of conditional frequentist and Bayesian methodology is always essentially unique, and the goal of this line of research (also mentioned by Pericchi and Reid) is to discover an essentially unique unified methodology (if it exists at all). It is interesting that, until Berger, Brown and Wolpert (1994), it was felt that unification in the testing domain was not possible.

RESPONSE TO PROFESSOR MAYO

I like Mayo's phrase "innocence by association." Alas, her discussion reflects the more standard "guilt by association." I have, in the past, often written about difficulties with p-values and unconditional error probabilities, and instead advocated use of posterior probabilities of hypotheses or Bayes factors. It is perhaps because of this history that Mayo begins the substantive part of her discussion with the statement that, "In contrast [to frequentist error probabilities], Berger's CEPs refer to the posterior probabilties of hypotheses under test ...."

In actuality, all the CEPs in the article are found by a purely frequentist computation, involving only the sampling distribution. It is noted in the article that these fully frequentist error probabilities happen to equal the objective Bayesian posterior probabilities, but this does not change their frequentist nature in any respect. (Likewise, it would not be reasonable to reject all standard frequentist confidence sets in the linear model just because they happen to coincide with objective Bayesian credible sets.) As another way of saying this, note that one could remove every reference to Bayesian analysis in the article and what would be left is simply the pure frequentist development of CEPs. Indeed, I originally toyed with writing the article this way—bringing in the relationship to Bayesian analysis only at the end—to try to reduce what I feared would be guilt by association.

Mayo's discussion then turns to a critique of Bayesian testing. Were this a Bayesian article, rather than an article primarily about a frequentist procedure, I would happily defend Bayesian analysis from these criticisms. I will refrain from doing so here, however, since such a defense would inevitably distract from the message that pure frequentist reasoning should result in adoption of the recommended CEPs. Many of Mayo's other comments also reflect this confusion about the frequentist nature of CEPs, and it would be repetitive if I responded to each. Hence I will confine myself to responding to a few other comments that Mayo makes.

- Why should the frequentist school have exclusive right to the term "error probability?" It is not difficult to simply add the designation "frequentist" (or Type I or Type II) or "Bayesian" to the term to differentiate between the schools.
- The applet is mentioned mainly as a reference for those who seek to improve their intuition concerning the behavior of p-values. (To paraphrase Neyman, can it be wrong to study how a concept works in repeated use?) In particular, none of the logic leading to CEPs is based on the applet.
- Mayo finds the stated frequentist principle to be vaguely worded and indeed it is. It does, however, convey what I believe to be the essence of the principle; see, for instance, Section 10 of Neyman (1977),

which gives a considerably expanded discussion of this version of the principle. I neglected to say that the frequentist principle can be applied separately to Type I errors and Type II errors, which is precisely what is done by CEPs.

- Mayo asserts that Neyman, Pearson and Fisher all thought that p-values are "legitimate error probabilities" (which, because of my first listed comment above, presumably means "frequentist error probabilities"). My reading of the literature is quite the opposite—that this was perhaps the most central element of the Neyman–Fisher debate, with Neyman opposing p-values because they are not predesignated (and hence cannot have a long-run frequency interpretation in actual use) and Fisher asserting that insistence on predesignated error probabilities is misguided in science.
- Mayo finishes with an introduction to "severity and a postdata interpretation of N–P tests," a development apparently aimed at bringing postdata assessment into N–P testing. Since CEPs provide postdata frequentist error probabilities based on essentially standard concepts (e.g., Type I and Type II error and conditioning), I do not see a need for anything more elaborate.

RESPONSE TO PROFESSOR PERICCHI

I certainly agree with Pericchi's historical perspective and elaborations on the need for unification in testing. I also agree with his assessment that a complete overhaul of statistical testing is necessary, with unconditional tests (and/or p-values) being replaced by conditional tests. It would be nice if the conditional frequentist paradigm would itself be sufficient for this retooling of testing, in that the task would then not be diverted by ideology. Unfortunately, the conditional frequentist testing theory is hard to extend in many ways (e.g., to the case of multiple hypotheses).

Pericchi does point out two scenarios where there is real potential for progress on the conditional frequentist side: sequential testing (Paulo, 2002b, is relevant here) and use of approximations such as BIC. However, in general, I suspect that the main use of conditional frequentist arguments will be to demonstrate that objective Bayesian testing does have a type of frequentist validity, thus making it also attractive to frequentists who recognize the centrality of conditioning.

RESPONSE TO PROFESSOR REID

Reid also emphasizes the value in a Bayesian–frequentist unification, and properly observes the importance of p-values as a technical tool for a wide variety of statistically important calculations. I quite agree; indeed, the article demonstrates another important technical use of p-values, in defining the conditioning statistic for the proposed conditional frequentist tests.

It is interesting that Reid has not observed frequent misinterpretation of p-values as Type I error probabilities, but rather has observed their frequent misinterpretation as posterior probabilities. Individuals' experiences are quite different in this regard; for instance, Hubbard (2000) recounts that the main problem in the management science literature is the misinterpretation of p-values as Type I error probabilities.

Reid mentions the issue of extending the analysis to composite null hypotheses, and worries that it requires essentially a full Bayesian analysis. Luckily, most classical composite null hypotheses have an invariance structure that allows reduction to a point null for conditional frequentist testing, as shown in Dass and Berger (2003).

ADDITIONAL REFERENCES

BEDRICK, E. J., CHRISTENSEN, R. and JOHNSON, W. O. (1996). A new perspective on priors for generalized linear models. *J. Amer. Statist. Assoc.* **91** 1450–1460.

BEDRICK, E. J., CHRISTENSEN, R. and JOHNSON, W. O. (1997). Bayesian binomial regression: Predicting survival at a trauma center. *Amer. Statist.* **51** 211–218.

BEDRICK, E. J., CHRISTENSEN, R. and JOHNSON, W. O. (2000). Bayesian accelerated failure time analysis with application to veterinary epidemiology. *Statistics in Medicine* **19** 221–237.

BERGER, J. and PERICCHI, L. (2001). Objective Bayesian methods for model selection: Introduction and comparison (with discussion). In *Model Selection* (P. Lahiri, ed.) 135–207. IMS, Hayward, CA.

BERGER, J. O. and WOLPERT, R. (1984). *The Likelihood Principle*. IMS, Hayward, CA.

BOX, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J. Roy. Statist. Soc. Ser. A* **143** 383–430.

CHRISTENSEN, R. (1995). Comment on Inman (1994). *Amer. Statist.* **49** 400.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

ENØE, C., GEORGIADIS, M. P. and JOHNSON, W. O. (2000). Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Preventive Veterinary Medicine* **45** 61–81.

FOSGATE, G. T., ADESIYUN, A. A., HIRD, D. W., JOHNSON, W. O., HIETALA, S. K., SCHURIG, G. G. and RYAN, J. (2002). Comparison of serologic tests for detection of Brucella infections in cattle and water buffalo (*Bubalus bubalis*). *American Journal of Veterinary Research* **63** 1598–1605.

GABRIEL, K. R. (1969). Simultaneous test procedures—some theory of multiple comparisons. *Ann. Math. Statist.* **40** 224–250.

GASTWIRTH, J. L., JOHNSON, W. O. and RENEAU, D. M. (1991). Bayesian analysis of screening data: Application to AIDS in blood donors. *Canad. J. Statist.* **19** 135–150.

GEORGIADIS, M. P., JOHNSON, W. O., GARDNER, I. A. and SINGH, R. (2003). Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests. *Applied Statistics* **52** 63–76.

GÖNEN, M., WESTFALL, P. H. and JOHNSON, W. O. (2003). Bayesian multiple testing for two-sample multivariate endpoints. *Biometrics*. To appear.

HANSON, T. E., BEDRICK, E. J., JOHNSON, W. O. and THURMOND, M. C. (2003). A mixture model for bovine abortion and fetal survival. *Statistics in Medicine* **22** 1725–1739.

HANSON, T., JOHNSON, W. O. and GARDNER, I. A. (2003). Hierarchical models for estimating disease prevalence and test accuracy in the absence of a gold-standard. *Journal of Agricultural, Biological and Environmental Statistics*. To appear.

HANSON, T. E., JOHNSON, W. O., GARDNER, I. A. and GEORGIADIS, M. (2003). Determining the disease status of a herd. *Journal of Agricultural, Biological and Environmental Statistics*. To appear.

JOHNSON, W. O. and GASTWIRTH, J. L. (1991). Bayesian inference for medical screening tests: Approximations useful for the analysis of Acquired Immune Deficiency Syndrome. *J. Roy. Statist. Soc. Ser. B* **53** 427–439.

JOHNSON, W. O., GASTWIRTH, J. L. and PEARSON, L. M. (2001). Screening without a gold standard: The Hui–Walter paradigm revisited. *American Journal of Epidemiology* **153** 921–924.

JOSEPH, L., GYORKOS, T. W. and COUPAL, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology* **141** 263–272.

LAVINE, M. and SCHERVISH, M. J. (1999). Bayes factors: what they are and what they are not. *Amer. Statist.* **53** 119–122.

LELE, S.R. (1998). Evidence functions and the optimality of the likelihood ratio. Paper presented at the Ecological Society of America Symposium on the Nature of Scientific Evidence, August 3, 1998, Baltimore.

LIU, Y., JOHNSON, W. O., GOLD, E. B. and LASLEY, B. L. (2003). Bayesian analysis of the effect of risk factors on probabilities of anovulation in cycling women. Unpublished manuscript.

MAYO, D. (1983). An objective theory of statistical testing. *Synthese* **57** 297–340.

MAYO, D. (1985). Behavioristic, evidentialist, and learning models of statistical testing. *Philos. Sci.* **52** 493–516.

MAYO, D. (1991). Sociological vs. metascientific theories of risk assessment. In *Acceptable Evidence: Science and Values in Risk Management* (D. Mayo and R. Hollander, eds.) 249–279. Oxford Univ. Press.

MAYO, D. (1992). Did Pearson reject the Neyman–Pearson philosophy of statistics? *Synthese* **90** 233–262.

MAYO, D. (1996). *Error and the Growth of Experimental Knowledge*. Univ. Chicago Press.

MAYO, D. (1997). Response to C. Howson and L. Laudan. *Philos. Sci.* **64** 323–333.

MAYO, D. and KRUSE, M. (2001). Principles of inference and their consequences. In *Foundations of Bayesianism* (D. Corfield and J. Williamson, eds.) 381–403. Kluwer, Dordrecht.

MAYO, D. and SPANOS, A. (2002). A severe testing interpretation of Neyman–Pearson tests. Working paper, Virginia Tech.

MCINTURFF, P., JOHNSON, W. O., GARDNER, I. A. and COWLING, D. W. (2003). Bayesian modeling of risk based on outcomes that are subject to error. *Statistics in Medicine*. To appear.

NEYMAN, J. (1976). Tests of statistical hypotheses and their use in studies of natural phenomena. *Comm. Statist. Theory Methods* **A5** 737–751.

PEARSON, E. S. (1966). On questions raised by the combination of tests based on discontinuous distributions. In *The Selected Papers of E. S. Pearson* 217–232. Cambridge Univ. Press. First published in *Biometrika* **37** 383–398 (1950).

ROSENTHAL, R. and RUBIN, D. B. (1994). The counternull value of an effect size: A new statistic. *Psychological Science* **5** 329–334.

ROYALL, R. M. (2000). On the probability of observing misleading statistical evidence (with discussion). *J. Amer. Statist. Assoc.* **95** 760–780.

SAMANIEGO, F. J. and RENEAU, D. M. (1994). Toward a reconciliation of the Bayesian and frequentist approaches to point estimation. *J. Amer. Statist. Assoc.* **89** 947–957.

SCHERVISH, M. J. (1996). P-values: What they are and what they are not. *Amer. Statist.* **50** 203–206.

SCHERVISH, M. J., SEIDENFELD, T. and KADANE, J. B. (2003). Measures of incoherence: How not to gamble if you must (with discussion). In *Bayesian Statistics* **7** (J. M. Bernardo et al., eds.). Oxford Univ. Press.

SUESS, E., GARDNER, I. A. and JOHNSON, W. O. (2002). Hierarchical Bayesian model for prevalence inferences and determination of a country's status for an animal pathogen. *Preventive Veterinary Medicine* **55** 155–171.

WESTFALL, P. H., JOHNSON, W. O. and UTTS, J. M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika* **84** 419–427.

ZELLNER, A. and SIOW, A. (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics* (J. M. Bernardo et al., eds.) 585–603. Oxford Univ. Press.