Could Fisher, Jeffreys, and Neyman Have Agreed on Testing? *

James O. Berger

Duke University

January 6, 2002

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 paper 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.

*This research was supported by the National Science Foundation, Grants DMS-9802261 and DMS-0103265. The paper is based on the Fisher lecture given by the author at the 2001 Joint Statistical Meetings.

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.

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 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 $p = .0037$).

- 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,σ) prior on the alternative.

- Neyman, had he pre-specified 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 of 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** – i.e., 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 paper. 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 is 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 papers 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) and Neyman (1961, 1977), and Jeffreys (1961). Other references and much useful historical discussion can be found, for instance, in Morrison and Henkel (1970), Speilman (1974, 1978), Carlson (1976), Savage (1976), Barnett (1982), Hall and Selinger (1986), Zabell (1992), Lehmann (1993), and Hubbard (2000). Furthermore, Fisher, Jeffreys and Neyman were statisticians of great depth and complexity, and their actual viewpoints towards 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, and 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, 1985b, contain discussions relating this 'practical' version to more common textbook definitions of frequentism.)

Frequentist Principle: In repeated 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)

$$\Pr(H_0|x) = \frac{B(x)}{1+B(x)} \quad \left(\text{or } \Pr(H_1|x) = \frac{1}{1+B(x)}\right), \tag{2.1}$$

  based on assigning equal prior probabilities of 1/2 to the two hypotheses and applying 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 ranges over the rejection or acceptance regions. Thus, reporting a pre-specified $\alpha = 0.05$ 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; cf. 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 non-frequentist 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 Smith (2001) estimate 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 The 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, 1999b) and Hubbard (2000), elaborating on earlier work such as Goodman (1992, 1993) and Royall (1997), make 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 subsection. Note that this confusion is different than 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 pre-experimentally 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 literatures. To access at least some of these literatures, 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).

In contrast, the statistics profession mostly holds itself blameless for this state of affairs, observing that the statistical literature (and good textbooks) do 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 are tacit accomplices in the unfortunate situation. With a unified testing methodology now available, it is time to mount this effort and provide non-statisticians 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 is often 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 automatically condition 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, and 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

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

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

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

This is not at all a sensible report, once the data is 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 one or their common value minus one, and each of these possibilities is equally likely to have occurred.

To obtain sensible frequentist answers here, one must 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,

$$\begin{aligned} P_\theta(C(X_1, X_2) \text{ contains } \theta \,|\, S = 2) &= 1 \\ P_\theta(C(X_1, X_2) \text{ contains } \theta \,|\, S = 0) &= \tfrac{1}{2}. \end{aligned}$$

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 $S(x)$, whose magnitude indicates the 'strength of evidence' in x. Then one computes conditional frequentist error probabilities of Type I and Type II, respectively, as

$$\alpha(s) = P_0(\text{Reject } H_0 | S(x) = s) \quad \text{and} \quad \beta(s) = P_1(\text{Accept } H_0 | S(x) = s). \tag{3.2}$$

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 A 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 whose distributions does not depend on θ), when available. Fisher's arguments for conditioning were a mix of theory and pragmatism (cf. Savage, 1976, and Basu, 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, 1974, and 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) asserts "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 that we will be discussing. That this test also provides a potential unification with Fisherian testing was only recently realized, however.

3.3 The 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

$$S = \max\{p_0, p_1\}. \tag{3.3}$$

The use of this conditioning statistic is equivalent to deciding that data (in either the rejection or acceptance regions) having the same p-value has 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 $p_0 \leq p_1$, and accepting otherwise;
- computing the Type I and Type II conditional error probabilities (CEPs) as in (3.2).

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

$$T^C = \begin{cases} \text{if } p_0 \leq p_1, & \text{reject } H_0 \text{ and report Type I CEP } \alpha(x) = \frac{B(x)}{1+B(x)}\,; \\ \text{if } p_0 > p_1, & \text{accept } H_0 \text{ and report Type II CEP } \beta(x) = \frac{1}{1+B(x)}, \end{cases} \tag{3.4}$$

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 $p_0 = P_0(X \leq x) = x$ and $p_1 = P_1(X \geq x) = 1 - \sqrt{x}$. 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:

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

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 of Fisher in this regard, and so it is difficult to speculate as to his reaction to use of the conditioning statistic in (3.3). 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 the fact 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 to 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) has 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 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 is at hand.)

We are not arguing here that the unconditional most powerful rejection region is better; indeed, we agree with the Lehmann (1993) conclusion that conditioning should usually take precedence over power. However, we are focusing here primarily on the report of conditional error probabilities, so that concerns over power would seem to be obviated.

Of course, we actually advocate conditioning in this paper on (3.3), and not just on y. 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 (3.4), and is easily seen to be

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

Note that the answers using this fully conditional frequentist test can be quite different from the answers 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 $\alpha(1) = \alpha(0) = 0.232$. At, say, $x = 4.0$, the CEPs are $\alpha(1) = 0.00034$ and $\alpha(0) = 0.119$, respectively. Clearly these 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 (3.4). We delay discussion of this issue until Subsection 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 to be 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 (3.3) as a conditioning statistic, since it will typically not be ancillary, but Fisher was quite pragmatic about conditioning and would use non-ancillary 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 (3.3) might have been a source of concern to Fisher. Unfortunately, the minimum simply does not yield a sensible conditioning statistic, as indicated in Sellke, Bayarri and Berger (2001).

There is one aspect of T^C that Fisher clearly would have disliked: the fact that an alternative hypothesis is necessary to define the test. We return to this issue in Subsection 5.4.

4.3 Jeffreys

The most crucial fact about the CEPs in (3.4) is that they precisely equal the objective Bayesian error probabilities, as defined in (2.1). 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 (3.4); we again delay discussion of this issue until Subsection 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 (3.4), 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 two 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. Second, in teaching testing, there is only one test – that given in (3.4). 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 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 (3.4). An example of this will be seen later, in a situation involving composite hypotheses.

5 Extensions

5.1 Alternate 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 $\{x : p_0(x) \leq p_1(x)\}$. This is, in fact, the *minimax* rejection region, i.e., that which has unconditional error probabilities $\alpha = \beta$. The disconcerting aspect here 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 (3.4) 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 is 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').

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 $P(H_0 \,|\, x) \leq 1/2$, rather than the region given in (3.4). 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 that simplicity argues for sticking with T^C.

5.2 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 (1997a, 1997b) considered tests of simple versus composite hypotheses, including 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 'average power' instead of the entire function. This goes a long way towards answering Fisher's criticisms concerning the difficulty of dealing with power functions.

Dass (1998) 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 (2001) 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 (2002) considers 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.

5.3 Other types of conditioning

One could consider a wide variety of conditioning statistics, other than that defined in (3.3). 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 are as 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.

Of course, one example is hardly compelling evidence. But the example shows that conditioning statistics can easily lead to error probabilities that seem counterintuitive. A chief attraction of the conditioning statistic in (3.3) is that it yields CEPs that can never be counterintuitive, since the resulting error probabilities must coincide with objective Bayesian error probabilities.

5.4 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, with many statisticians having 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 to first develop an objective Bayes factor, $B(x)$. This was done by choosing a Polya tree prior for F, centered at H_0, and choosing the right-Haar prior, $\pi(\mu, \sigma) = 1/\sigma$, for the location-scale parameters in each model. Berger and Guglielmi (2001) show how to then compute $B(x)$.

The recommended conditional frequentist test is then given automatically by (3.4). Because the null hypothesis has a suitable group invariance structure, Dass and Berger (1998) can be used to show that the conditional Type I error is indeed $\alpha(x)$ in (3.4), while $\beta(x)$ is the 'average' Type II error, as mentioned in Subsection 5.2. 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 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) develop a variety of plausible nonparametric alternatives, and show 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 (3.4), it is clear that all such tests have

$$\alpha(p) \geq (1 + [-e\, p \log(p)]^{-1})^{-1}. \qquad (5.5)$$

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.

p	.2	.1	.05	.01	.005	.001
$\alpha(p)$	.465	.385	.289	.111	.067	.0184

Table 1: Calibration of p-values as lower bounds on conditional error probabilities.

While simple and revealing, the calibration in (5.5) is often a too-small lower bound on the conditional Type I error. Alternative calibrations have been suggested in, e.g., Good (1958, 1992).

References

[1] Barnett, V. (1982). *Comparative Statistical Inference.* Wiley, New York.

[2] Basu, D. (1975). Statistical information and likelihood (with discussion). *Sankhya*, **A 37**, 1–71.

[3] Basu, D. (1977). On the elimination of nuisance parameters. *J. Amer. Stat. Assoc.* , **72**, 355–366.

[4] Bayarri, M.J. and Berger, J. (2000). P-values for composite null models (with discussion). *J. Amer. Statist. Assoc.*, **95**, 1127–1142.

[5] Berger, J. (1985a). *Statistical Decision Theory and Bayesian Analysis, Second Edition.* Springer-Verlag, New York.

[6] Berger, J. (1985b). The frequentist viewpoint and conditioning. In *Proceedings of the Berkeley Conference in Honor of Jack Kiefer and Jerzy Neyman* (L. Le Cam and R. Olshen, Eds.). Wadsworth, Belmont.

[7] Berger, J. and Berry, D. (1988). Analyzing data: Is objectivity possible? *American Scientist*, **76**, 159–165.

[8] Berger, J., Boukai, B. and Wang, Y. (1997). Unified frequentist and Bayesian testing of a precise hypothesis (with discussion). *Statistical Science*, **12**, 133–160.

[9] Berger, J., Boukai, B. and Wang, Y. (1999). Simultaneous Bayesian-frequentist sequential testing of nested hypotheses. *Biometrika*, **86**, 79–92.

[10] Berger, J., Brown, L. and Wolpert, R. (1994). A unified conditional frequentist and Bayesian test for fixed and sequential hypothesis testing. *Ann. Statist.* **22**, 1787-1807.

[11] Berger, J. and Delampady, M. (1987). Testing precise hypotheses (with discussion). *Statist. Science*, **2**, 317–352.

[12] Berger, J. and Mortera, J. (1999). Default Bayes factors for non-nested hypothesis testing. *J. Amer. Statist. Assoc.*, **94**, 542–554.

[13] Berger, J. and Sellke, T. (1987). Testing of a point null hypothesis: The irreconcilability of significance levels and evidence (with discussion). *J. Amer. Statist. Assoc.*, **82**, 112–139.

[14] Berger, J. and Wolpert, R. L. (1988). *The Likelihood Principle* (second edition, with discussion). Institute of Mathematical Statistics, Hayward, CA.

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

[16] Birnbaum, A. (1961). On the foundation of statistical inference: binary experiments. *Ann. Math. Statist.*, **32**, 414–435.

[17] Braithwaite, R.B. (1953). *Scientific Exploration.* Cambridge Univ. Press, Cambridge.

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

[19] Carlson, R. (1976). The logic of tests of significance (with discussion). *Philosophy of Science*, **43**, 116–128.

[20] 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.

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

[22] Dass, S. (1998). *Unified Bayesian and Conditional Frequentist Testing Procedures.* Ph.D. Thesis, Purdue University.

[23] Dass, S. and Berger, J. (1998). Unified Bayesian and conditional frequentist testing of composite hypotheses. ISDS Discussion Paper 98-43.

[24] Delampady, M. and Berger, J. (1990). Lower bounds on posterior probabilities for multinomial and chi–squared tests. *Ann. Statist.*, **18**, 1295–1316.

[25] Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, **70**, 193–242.

[26] Efron, B., and Gous, A. (2001). Scales of evidence for model selection: Fisher versus Jeffreys (with discussion). In *Model Selection*, P. Lahiri, Ed., pp. 208–256. Institute of Mathematical Statistics Lecture Notes-Monograph Series Volume 38, Beachwood, Ohio.

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

[28] Fisher, R.A. (1935). The logic of inductive inference. *J. Roy. Statist. Soc.*, **98**, 39–54.

[29] Fisher, R.A. (1955). Statistical methods and scientific induction. *J. Roy. Statist. Soc.*, **B 17**, 69–78.

[30] Fisher, R.A. (1973). *Statistical Methods and Scientific Inference (3rd ed.)*. Macmillan, London.

[31] Gibbons, J. and Pratt, J. (1975). P-values: interpretation and methodology. *Amer. Statisician*, **29**, 20–25.

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

[33] Good, I. J. (1992). The Bayesian/non-Bayesian compromise: a brief review. *J. Amer. Statist. Assoc.*, **87**, 597–606.

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

[35] Goodman, S. (1993). P-values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *American J. Epidemiology*, **137**, 485–496.

[36] Goodman, S. (1999). Toward evidence-based medical statistics 1: the p-value fallacy. *Annals of Internal Medicine*, **130**, 995-1004.

[37] Goodman, S. (1999b). Toward evidence-based medical statistics 2: the Bayes factor. *Annals of Internal Medicine*, **130**, 1005–1013.

[38] Hacking, I. (1965). *The Logic of Statistical Inference.* Cambridge Univ. Press, New York.

[39] Hall, P., and Sellinger, B. (1986). Statistical significance: balancing evidence against doubt. *Australian J. Statistic.*, **28**, 354–370.

[40] 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 University.

[41] Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, London.

[42] Kalbfleish, J. D., and Sprott, D. A. (1974). Marginal and conditional likelihood. *Sankhya*, **A 35**, 311–328.

[43] Kiefer, J. (1976). Admissibility of conditional confidence procedures. *Ann. Math. Statist.*, **4**, 836–865.

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

[45] Kyburg, H.E. Jr. (1974). *The Logical Foundations of Statistical Inference.* Reidel, Boston.

[46] Laplace, P. S. (1812). *Théorie Analytique des Probabilités.* Courcier, Paris.

[47] Lehmann, E. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *J. Amer. Statist. Assoc.*, **88**, 1242–1249.

[48] Matthews, R. (1998). The great health hoax. In *The Sunday Telegraph*, September 13, 1998.

[49] Morrison, D.E., and Henkel, R.E. (1970). *The Significance Test Controversy.* Aldine, Chicago.

[50] Neyman, J. (1961). Silver jubilee of my dispute with Fisher. *J. Operations Research Society of Japan*, **3**, 145–154.

[51] Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthèse*, **36**, 97–131.

[52] Neyman, J., and Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Roy. Soc. of London*, **A, 231**, 289–337.

[53] Pearson, E.S. (1955). Statistical concepts in their relation to reality. *J. Roy. Statist. Soc.*, **B 17**, 204–207.

[54] Pearson, E.S. (1962). Some thoughts on statistical inference. *Ann. Math. Statist.*, **33**, 394–403.

[55] Reid, N. (1995). The roles of conditioning in inference. *Statist. Science*, **10**, 138–157.

[56] Robins, J. M., van der Vaart, A., and Ventura, V. (2000). The asymptotic distribution of p-values in composite null models. *J. Amer. Statist. Assoc.*, **95**, 1143–1156.

[57] Royall, R.M. (1997). *Statistical Evidence: A Likelihood Paradigm.* Chapman and Hall, New York.

[58] Savage, L. J. (1976). On rereading R. A. Fisher. *Ann. Statist.*, 4, 441–500.

[59] Seidenfeld, T. (1979). *Philosophical Problems of Statistical Inference.* Reidel, Boston.

[60] Sellke, T., Bayarri, M. J., and Berger, J. O. (2001). Calibration of p-values for testing precise null hypotheses. *Amer. Statistician*, **55**, 62–71.

[61] Spielman, S. (1974). The logic of tests of significance. *Philosophy of Science*, **41**, 211–226.

[62] Spielman, S. (1978). Statistical dogma and the logic of signifance testing. *Philosophy of Science*, **45**, 120–135.

[63] Sterne, J.A.C., and Smith, G.D. (2001). Sifting the evidence – what's wrong with significance tests? *BMJ*, **322**, 226–231.

[64] Welch, B., and Peers, H. (1963). On formulae for confidence points based on intervals of weighted likelihoods. *J. Roy. Statist. Soc.* **B 25**, 318–329.

[65] Wolpert, R. L. (1995). Testing simple hypotheses. In *Studies in Classification, Data Analysis, and Knowledge Organization*, Vol. 7. eds. H. H. Bock and W. Polasek, pp. 289–297. Springer-Verlag, Heidelberg.

[66] Zabell, S. (1992). R.A. Fisher and the fiducial argument. *Statist. Science*, **7**, 369–387.