# Testing Fisher, Neyman, Pearson, and Bayes

Ronald CHRISTENSEN

This article presents a simple example that illustrates the key differences and similarities between the Fisherian, Neyman-Pearson, and Bayesian approaches to testing. Implications for more complex situations are also discussed.

KEY WORDS: Confidence; Lindley's paradox; Most powerful test; $p$ values; Significance tests.

## 1. INTRODUCTION

One of the famous controversies in statistics is the dispute between Fisher and Neyman-Pearson about the proper way to conduct a test. Hubbard and Bayarri (2003) gave an excellent account of the issues involved in the controversy. Another famous controversy is between Fisher and almost all Bayesians. Fisher (1956) discussed one side of these controversies. Berger's Fisher lecture attempted to create a consensus about testing; see Berger (2003).

This article presents a simple example designed to clarify many of the issues in these controversies. Along the way many of the fundamental ideas of testing from all three perspectives are illustrated. The conclusion is that Fisherian testing is not a competitor to Neyman-Pearson (NP) or Bayesian testing because it examines a different problem. As with Berger and Wolpert (1984), I conclude that Bayesian testing is preferable to NP testing as a procedure for deciding between alternative hypotheses.

The example involves data that have four possible outcomes, $r = 1, 2, 3, 4$. The distribution of the data depends on a parameter $\theta$ that takes on values $\theta = 0, 1, 2$. The distributions are defined by their discrete densities $f(r|\theta)$ which are given in Table 1. In Section 2, $f(r|0)$ is used to illustrate Fisherian testing. In Section 3, $f(r|0)$ and $f(r|2)$ are used to illustrate testing a simple null hypothesis versus a simple alternative hypothesis although Subsection 3.1 makes a brief reference to an NP test of $f(r|1)$ versus $f(r|2)$. Section 4 uses all three densities to illustrate testing a simple null versus a composite alternative. Section 5 discusses some issues that do not arise in this simple example. For those who want an explicit statement of the differences between Fisherian and NP testing, one appears at the beginning of Section 6 which also contains other conclusions and comments. For readers who are unfamiliar with the differences, it seems easier to pick them up from the example than it would be from a general statement. Very briefly, however, a Fisherian test involves only one hypothesized model. The distribution of the data must be known and this distribution is used both to determine the test and to evaluate the outcome of the test. An NP test involves two hypotheses, a null hypothesis, and an alternative. A family of tests is considered that all have the same probability $\alpha$ of rejecting the null hypothesis when it is true. Within this family, a particular test is chosen so as to maximize the power, that is, the probability of rejecting the null hypothesis when the alternative hypothesis is true.

Finally, some truth in advertising. Fisher did not use the term Fisherian testing and certainly made no claim to have originated the ideas. He referred to "tests of significance." This often, but not always, gets contrasted with "tests of hypotheses." When discussing Fisherian testing, I make no claim to be expositing exactly what Fisher proposed. I am expositing a logical basis for testing that is distinct from Neyman-Pearson theory and that is related to Fisher's views.

## 2. FISHERIAN TESTS

The key fact in a Fisherian test is that it makes no reference to any alternative hypothesis. It can really be thought of as a model validation procedure. We have the distribution of the (null) model and we examine whether the data look weird or not.

For example, an $\alpha = .01$ Fisherian test of $H_0 : \theta = 0$ is based entirely on

| $r$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $f(r\|0)$ | .980 | .005 | .005 | .010 |

With only this information, one must use the density itself to determine which data values seem weird and which do not. Obviously, observations with low density are those least likely to occur, so they are considered weird. In this example, the weirdest observations are $r = 2, 3$ followed by $r = 4$. An $\alpha = .01$ test would reject the model $\theta = 0$ if either a 2 or a 3 is observed. An $\alpha = .02$ test rejects when observing any of $r = 2, 3, 4$.

In lieu of an $\alpha$ level, Fisher advocated using a $p$ value to evaluate the outcome of a test. The $p$ value is the probability of seeing

Table 1. Discrete Densities to be Tested

| $r$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $f(r\|0)$ | .980 | .005 | .005 | .010 |
| $f(r\|1)$ | .100 | .200 | .200 | .500 |
| $f(r\|2)$ | .098 | .001 | .001 | .900 |

Ronald Christensen is Professor, Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131 (E-mail: fletcher@stat.unm.edu). I thank the editor and associate editor for comments and suggestions that substantially improved this article.

DOI: 10.1198/000313005X20871

something as weird or weirder than you actually saw. (There is no need to specify that it is computed under the null hypothesis because there is only one hypothesis.) In this example, the weirdest observations are 2 and 3 and they are equally weird. Their combined probability is .01 which is the $p$ value regardless of which you actually observe. If you see a 4, both 2 and 3 are weirder, so the combined probability is .02.

| $r$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $f(r\|0)$ | .980 | .005 | .005 | .010 |
| $p$ value | 1.00 | .01 | .01 | .02 |

In Fisherian testing, the $p$ value is actually a more fundamental concept than the $\alpha$ level. Technically, an $\alpha$ level is simply a decision rule as to which $p$ values will cause one to reject the null hypothesis. In other words, it is merely a decision point as to how weird the data must be before rejecting the null model. If the $p$ value is less than or equal to $\alpha$, the null is rejected. Implicitly, an $\alpha$ level determines what data would cause one to reject $H_0$ and what data will not cause rejection. The $\alpha$ level rejection region is defined as the set of all data points that have a $p$ value less than or equal to $\alpha$. Note that in this example, an $\alpha = .01$ test is identical to an $\alpha = .0125$ test. Both reject when observing either $r = 2$ or 3. Moreover, the probability of rejecting an $\alpha = .0125$ test when the null hypothesis is true is *not* .0125, it is .01. However, Fisherian testing is not interested in what the probability of rejecting the null hypothesis *will be*, it is interested in what the probability *was* of seeing weird data.

The philosophical basis of a Fisherian test is akin to proof by contradiction. We have a model and we examine the extent to which the data contradict the model. The basis for suggesting a contradiction is actually observing data that are highly improbable under the model. The $p$ value gives an excellent measure of the extent to which the data do not contradict the model. (Large $p$ values do not contradict the model.) If an $\alpha$ level is chosen, for any semblance of a contradiction to occur, the $\alpha$ level must be small. On the other hand, even without a specific alternative, making $\alpha$ too small will defeat the purpose of the test, making it extremely difficult to reject the test for any reason. A reasonable view would be that an $\alpha$ level should never be chosen; that a scientist should simply evaluate the evidence embodied in the $p$ value.

As in any proof by contradiction, the results are skewed. If the data contradict the model, we have evidence that the model is invalid. If the data do not contradict the model, we have an attempt at proof by contradiction in which we got no contradiction. If the model is not rejected, the best one can say is that the data are consistent with the model. Not rejecting certainly does not prove that the model is correct, whence comes the common exhortation that one should never accept a null hypothesis.

## 3. SIMPLE VERSUS SIMPLE

Consider testing the simple null hypothesis $H_0 : \theta = 0$ versus the simple alternative hypothesis $H_A : \theta = 2$. This now appears to be a decision problem. We have two alternatives and we are deciding between them. This formulation as a decision problem is a primary reason that Fisher objected to NP testing, see Fisher (1956, chap. 4).

The relevant information for this testing problem is

| $r$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $f(r\|0)$ | .980 | .005 | .005 | .010 |
| $f(r\|2)$ | .098 | .001 | .001 | .900 |

Before examining formal testing procedures, look at the distributions. Intuitively, if we see $r = 4$ we are inclined to believe $\theta = 2$, if we see $r = 1$ we are quite inclined to believe that $\theta = 0$, and if we see either a 2 or a 3, it is still five times more likely that the data came from $\theta = 0$.

Although Fisherian testing does not use an explicit alternative, there is nothing to stop us from doing two Fisherian tests: a test of $H_0 : \theta = 0$ and then another test of $H_0 : \theta = 2$. The Fisherian tests both give perfectly reasonable results. The test for $H_0 : \theta = 0$ has small $p$ values for any of $r = 2, 3, 4$. These are all strange values when $\theta = 0$. The test for $H_0 : \theta = 2$ has small $p$ values when $r = 2, 3$. When $r = 4$, we do not reject $\theta = 2$; when $r = 1$, we do not reject $\theta = 0$; when $r = 2, 3$, we reject both $\theta = 0$ and $\theta = 2$. The Fisherian tests are not being forced to choose between the two distributions. Seeing either a 2 or a 3 is weird under both distributions.

### 3.1 Neyman-Pearson Tests

NP tests treat the two hypotheses in fundamentally different ways. A test of $H_0 : \theta = 0$ versus $H_A : \theta = 2$ is typically different from a test of $H_0 : \theta = 2$ versus $H_A : \theta = 0$. We examine the test of $H_0 : \theta = 0$ versus $H_A : \theta = 2$.

NP theory seeks to find the best $\alpha$ level test. $\alpha$ is the probability of rejecting $H_0$ when it is true. The rejection region is the set of data values that cause one to reject the null hypothesis, so under $H_0$ the probability of the rejection region must be $\alpha$. The best test is defined as the one with the highest power, that is, the highest probability of rejecting $H_0$ (observing data in the rejection region) when $H_A$ is true.

Defining the $\alpha$ level as the probability of rejecting the null hypothesis when it is true places an emphasis on repeated sampling so that the Law of Large Numbers suggests that about $\alpha$ of the time you will make an incorrect decision, provided the null hypothesis is true in all of the samples. Although this is obviously a reasonable definition prior to seeing the data, its relevance after seeing the data is questionable.

To allow arbitrary $\alpha$ levels, one must consider randomized tests. A randomized test requires a randomized rejection region. How would one perform an $\alpha = .0125$ test? Three distinct tests are: (a) reject whenever $r = 4$ and flip a coin, if it comes up heads, reject when $r = 2$; (b) reject whenever $r = 4$ and flip a coin, if it comes up heads, reject when $r = 3$; (c) reject whenever $r = 2$ or 3 and flip a coin twice, if both come up heads, reject when $r = 4$. It is difficult to convince anyone that these are reasonable practical procedures.

| $r$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $f(r\|0)$ | .980 | .005 | .005 | .010 |
| $f(r\|2)$ | .098 | .001 | .001 | .900 |
| $f(r\|2)/f(r\|0)$ | .1 | .2 | .2 | 90 |

As demonstrated in the famous Neyman-Pearson lemma (see Lehmann 1997, chap. 3), optimal NP tests are based on the likelihood ratio $f(r|2)/f(r|0)$. The best NP test rejects for the largest values of the likelihood ratio, thus the $\alpha = .01$ NP test rejects when $r = 4$. This is completely different from the Fisherian .01 test of $H_0$ that rejected when $r = 2, 3$. (On the other hand, the $\alpha = .02$ NP test coincides with the Fisherian test. Both reject when observing any of $r = 2, 3, 4$.) The power of the $\alpha = .01$ NP test is .9 whereas the power of the Fisherian $\alpha = .01$ test is only $.001 + .001 = .002$. Clearly the Fisherian test is not a good way to decide between these alternatives. But then the Fisherian test was not designed to decide between two alternatives. It was designed to see whether the null model seemed reasonable and, on its own terms, it works well. Although the meaning of $\alpha$ differs between Fisherian and NP tests, we have chosen two examples, $\alpha = .01$ and $\alpha = .02$, in which the Fisherian test (rejection region) also happens to define an NP test with the same numerical value of $\alpha$. Such a comparison would not be appropriate if we had examined, say, $\alpha = .0125$ Fisherian and NP tests.

NP testing and Fisherian testing are not comparable procedures, a point also made by Hubbard and Bayarri (2003). NP testing is designed to optimally detect some alternative hypothesis and Fisherian testing makes no reference to any alternative hypothesis. I might suggest that NP testers tend to want to have their cake and eat it too. By this I mean that many of them want to adopt the philosophy of Fisherian testing (involving $p$ values, using small $\alpha$ levels, and never accepting a null hypothesis) while still basing their procedure on an alternative hypothesis.

In particular, the motivation for using small $\alpha$ levels seems to be based entirely on the philosophical idea of proof by contradiction. Using a large $\alpha$ level would eliminate the suggestion that the data are unusual and thus tend to contradict $H_0$. However, NP testing cannot appeal to the idea of proof by contradiction. For example, in testing $H_0 : \theta = 1$ versus $H_A : \theta = 2$, the most powerful NP test would reject for $r = 4$, even though $r = 4$ is the most probable value for the data under the null hypothesis. (For any $\alpha < .5$, a randomized test is needed.) In particular, this example makes it clear that $p$ values can have no role in NP testing! See also Hubbard and Bayarri (2003) and discussion.

It seems that once you base the test on wanting a large probability of rejecting the alternative hypothesis, you have put yourself in the business of deciding between the two hypotheses. Even on this basis, the NP test does not always perform very well. The rejection region for the $\alpha = .02$ NP test of $H_0 : \theta = 0$ versus $H_A : \theta = 2$ includes $r = 2, 3$, even though 2 and 3 are five times more likely under the null hypothesis than under the alternative. Admittedly, 2 and 3 are weird things to see under either hypothesis, but when deciding between these specific alternatives, rejecting $\theta = 0$ (accepting $\theta = 2$) for $r = 2$ or 3 does not seem reasonable. The Bayesian approach to testing, discussed in the next subsection, seems to handle this decision problem well.

*Table 2. Posterior Probabilities of $\theta = 0,2$ for Two Prior Distributions a and b*

| Prior | r | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| | $f(r\mid 0)$ | .980 | .005 | .005 | .010 |
| | $f(r\mid 2)$ | .098 | .001 | .001 | .900 |
| $p_a(0) = 1/2$ | $p_a(0\mid r)$ | .91 | .83 | .83 | .01 |
| $p_a(2) = 1/2$ | $p_a(2\mid r)$ | .09 | .17 | .17 | .99 |
| $p_b(0) = 1/6$ | $p_b(0\mid r)$ | .67 | .50 | .50 | .002 |
| $p_b(2) = 5/6$ | $p_b(2\mid r)$ | .33 | .50 | .50 | .998 |

### 3.2 Bayesian Tests

Bayesian analysis requires us to have prior probabilities on the values of $\theta$. It then uses Bayes' theorem to combine the prior probabilities with the information in the data to find "posterior" probabilities for $\theta$ given the data. All decisions about $\theta$ are based entirely upon these posterior probabilities. The information in the data is obtained from the likelihood function. For an observed data value, say $r = r^*$, the likelihood is the function of $\theta$ defined by $f(r^*|\theta)$.

In our simple versus simple testing example, let the prior probabilities on $\theta = 0, 2$ be $p(0)$ and $p(2)$. Applying Bayes' theorem to observed data $r$, we turn these prior probabilities into posterior probabilities for $\theta$ given $r$, say $p(0|r)$ and $p(2|r)$. To do this we need the likelihood function which here takes on only the two values $f(r|0)$ and $f(r|2)$. From Bayes' theorem,

$$p(\theta|r) = \frac{f(r|\theta)p(\theta)}{f(r|0)p(0) + f(r|2)p(2)}, \qquad \theta = 0, 2.$$

Decisions are based on these posterior probabilities. Other things being equal, whichever value of $\theta$ has the larger posterior probability is the value of $\theta$ that we will accept. If both posterior probabilities are near .5, we might admit that we do not know which is right.

In practice, posterior probabilities are computed only for the value of $r$ that was actually observed, but Table 2 gives posterior probabilities for all values of $r$ and two sets of prior probabilities: (a) one in which each value of $\theta$ has the same probability, 1/2, and (b) one set in which $\theta = 2$ is five times more probable than $\theta = 0$.

As is intuitively reasonable, regardless of the prior distribution, if you see $r = 4$ the posterior is heavily in favor of $\theta = 2$, and if you see $r = 1$ the posterior substantially favors $\theta = 0$.

The key point is what happens when $r$ equals 2 or 3. With equal prior weight on the $\theta$'s, the posterior heavily favors $\theta = 0$, that is, with $r = 2$, $p_a(0|2) = .83$, $p_a(2|2) = .17$, and with $r = 3$, $p_a(0|3) = .83$, $p_a(2|3) = .17$. It is not until our prior makes $\theta = 2$ five times more probable than $\theta = 0$ that we wash out the evidence from the data that $\theta = 0$ is more likely, that is, $p_b(0|2) = p_b(2|2) = .50$ and $p_b(0|3) = p_b(2|3) = .50$. Given the prior, the Bayesian procedure is always reasonable.

The Bayesian analysis gives no special role to the null hypothesis. It treats the two hypotheses on an equal footing. That NP theory treats the hypotheses in fundamentally different ways is something that many Bayesians find disturbing.

If utilities are available, the Bayesian can base a decision on maximizing expected posterior utility. Berry (2004) discussed the practical importance of developing approximate utilities for designing clinical trials.

The absence of a clear source for the prior probabilities seems to be the primary objection to the Bayesian procedure. Typically, if we have enough data, the prior probabilities are not going to matter because the posterior probabilities will be substantially the same for different priors. If we do not have enough data, the posteriors will not agree but why should we expect them to? The best we can ever hope to achieve is that reasonable people (with reasonable priors) will arrive at a consensus when enough data are collected. In the example, seeing one observation of $r = 1$ or 4 is already enough data to cause substantial consensus. One observation that turns out to be a 2 or a 3 leaves us wanting more data.

## 4. SIMPLE VERSUS COMPOSITE

Now consider testing the simple null hypothesis $H_0 : \theta = 0$ versus the composite alternative hypothesis $H_A : \theta > 0$. Of course the composite alternative has only two values. Looking at the distributions in Table 1, the intuitive conclusions are pretty clear. For $r = 1$, go with $\theta = 0$. For $r = 4$, go with $\theta = 2$. For $r = 2, 3$, go with $\theta = 1$.

Fisherian testing has nothing new to add to this situation except the observation that when $\theta = 1$, none of the data are really weird. In this case, the strangest observation is $r = 1$ which has a $p$ value of .1.

The best thing that can happen in NP testing of a composite alternative is to have a uniformly most powerful test. With $H_A : \theta > 0$, let $\theta^*$ be a particular value that is greater than 0. Test the simple null $H_0 : \theta = 0$ against the simple alternative $H_A : \theta = \theta^*$. If, for a given $\alpha$, the most powerful test has the same rejection region regardless of the value of $\theta^*$, then that test is the uniformly most powerful test. It is a simple matter to see that the $\alpha = .01$ NP most powerful test of $H_0 : \theta = 0$ versus $H_A : \theta = 1$ rejects when $r = 4$. Because the most powerful tests of the alternatives $H_A : \theta = 1$ and $H_A : \theta = 2$ are identical, and these are the only permissible values of $\theta > 0$, this is the uniformly most powerful $\alpha = .01$ test. The test makes a "bad" decision when $r = 2, 3$ because with $\theta = 1$ as a consideration, you would intuitively like to reject the test. The $\alpha = .02$ uniformly most powerful test rejects for $r = 2, 3, 4$, which is in line with our intuitive evaluation, but recall from the previous section that this is the test that (intuitively) should not have rejected for $r = 2, 3$ when testing only $H_A : \theta = 2$.

An even-handed Bayesian approach might take prior probabilities that are the same for the null hypothesis and the alternative, that is, $\Pr[\theta = 0] = .5$ and $\Pr[\theta > 0] = .5$. Moreover, we might then put the same prior weight on every possible $\theta$ value within the alternative, thus $\Pr[\theta = 1|\theta > 0] = .5$ and $\Pr[\theta = 2|\theta > 0] = .5$. Equivalently, $p(0) = .5$, $p(1) = .25$, and $p(2) = .25$. The posterior probabilities are

| $r$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $p(0\|r)$ | .908 | .047 | .047 | .014 |
| $p(1\|r)$ | .046 | .948 | .948 | .352 |
| $p(2\|r)$ | .045 | .005 | .005 | .634 |
| $\Pr[\theta>0\|r]$ | .091 | .953 | .953 | .986 |

These agree well with the intuitive conclusions, even though the prior puts twice as much weight on $\theta = 0$ as on the other $\theta$'s.

The Bayesian approach to testing a simple null against a composite alternative can be recast as testing a simple null versus a simple alternative. Using the prior probability on the values of $\theta$ given that the alternative hypothesis is true, one can find the average distribution for the data under the alternative. With $\Pr[\theta = 1|\theta > 0] = .5$ and $\Pr[\theta = 2|\theta > 0] = .5$, the average distribution under the alternative is $.5f(r|1) + .5f(r|2)$. The Bayesian test of the $\theta = 0$ density $f(r|0)$ against this average density for the data under the alternative yields the posterior probabilities $p(0|r)$ and $\Pr[\theta > 0|r]$.

It might also be reasonable to put equal probabilities on every $\theta$ value. In decision problems like this, where you know the (sampling) distributions, the only way to get unreasonable Bayesian answers is to use an unreasonable prior.

## 5. GENERAL MATTERS

### 5.1 Fisherian Testing

One thing that the example in Section 2 does not illustrate is that in a Fisherian test, it is not clear what aspect of the model is being rejected. If $y_1, y_2, \ldots, y_n$ are independent $N(\mu, \sigma^2)$ and we perform a $t$ test of $H_0 : \mu = 0$, a rejection could mean that $\mu \neq 0$, or it could mean that the data are not independent, or it could mean that the data are not normal, or it could mean that the variances of the observations are not equal. In other words, rejecting a Fisherian test suggests that something is wrong with the model. It does not specify what is wrong.

The example of a $t$ test raises yet another question. Why should we summarize these data by looking at the $t$ statistic,

$$\frac{\bar{y} - 0}{s/\sqrt{n}} \ ?$$

One reason is purely practical. In order to perform a test, one must have a known distribution to compare to the data. Without a known distribution there is no way to identify which values of the data are weird. With the normal data, even when assuming $\mu = 0$, we do not know $\sigma^2$ so we do not know the distribution of the data. By summarizing the data into the $t$ statistic, we get a function of the data that has a known distribution, which allows us to perform a test. Another reason is essentially: why not look at the $t$ statistic? If you have another statistic you want to base a test on, the Fisherian tester is happy to oblige. To quote Fisher (1956, p. 49), the hypothesis should be rejected "if any relevant feature of the observational record can be shown to [be] sufficiently rare." After all, if the null model is correct, it should be able to withstand any challenge. Moreover, there is no hint in this passage of worrying about the effects of performing multiple tests. Inflating the probability of Type I error (rejecting the null when it is true) by performing multiple tests is not a concern in Fisherian testing because the probability of Type I error is not a concern in Fisherian testing.

The one place that possible alternative hypotheses arise in Fisherian testing is in the choice of test statistics. Again quoting Fisher (1956, p. 50), "In choosing the grounds upon which a general hypothesis should be rejected, personal judgement may and should properly be exercised. The experimenter will rightly consider all points on which, in the light of current knowledge, the hypothesis may be imperfectly accurate, and will select tests, so far as possible, sensitive to these possible faults, rather than to others." Nevertheless, the logic of Fisherian testing in no way depends on the source of the test statistic.

There are two final points to make on how this approach to testing impacts standard data analysis.

First, $F$ tests and $\chi^2$ tests are typically rejected only for large values of the test statistic. Clearly, in Fisherian testing, that is inappropriate. Finding the $p$ value for an $F$ test should involve finding the density associated with the observed $F$ statistic and finding the probability of getting any value with a lower density. This will be a two-tailed test, rejecting for values that are very large or very close to 0. As a practical matter, it is probably sufficient to always remember that "one-sided $p$ values" very close to 1 should make us as suspicious of the model as one-sided $p$ values near 0. Christensen (2003) discusses situations that cause $F$ statistics to get close to 0.

Second, although Fisher never gave up on his idea of fiducial inference, one can use Fisherian testing to arrive at "confidence regions" that do not involve either fiducial inference or repeated sampling. A $(1-\alpha)$ confidence region can be defined simply as a collection of parameter values that would not be rejected by a Fisherian $\alpha$ level test, that is, a collection of parameter values that are consistent with the data as judged by an $\alpha$ level test. This definition involves no long run frequency interpretation of "confidence." It makes no reference to what proportion of hypothetical confidence regions would include the true parameter. It does, however, require one to be willing to perform an infinite number of tests without worrying about their frequency interpretation. This approach also raises some curious ideas. For example, with the normal data discussed earlier, this leads to standard $t$ confidence intervals for $\mu$ and $\chi^2$ confidence intervals for $\sigma^2$, but one could also form a joint 95% confidence region for $\mu$ and $\sigma^2$ by taking all the pairs of values that satisfy

$$\frac{|\overline{y}-\mu|}{\sigma/\sqrt{n}} < 1.96.$$

Certainly all such $\mu$, $\sigma^2$ pairs are consistent with the data as summarized by $\bar{y}$.

### 5.2 Neyman-Pearson Tests

To handle more general testing situations, NP theory has developed a variety of concepts such as unbiased tests, invariant tests, and $\alpha$ similar tests; see Lehmann (1997). For example, the two-sided $t$ test is not a uniformly most powerful test but it is a uniformly most powerful unbiased test. Similarly, the standard $F$ test in regression and analysis of variance is a uniformly most powerful invariant test.

The NP approach to finding confidence regions is also to find parameter values that would not be rejected by a $\alpha$ level test. However, just as NP theory interprets the size $\alpha$ of a test as the long run frequency of rejecting an incorrect null hypothesis, NP theory interprets the confidence $1-\alpha$ as the long run probability of these regions including the true parameter. The rub is that you only have one of the regions, not a long run of them, and you are trying to say something about this parameter based on these data. In practice, the long run frequency of $\alpha$ somehow gets turned into something called "confidence" that this parameter is within this particular region.

Although I admit that the term "confidence," as commonly used, feels good, I have no idea what "confidence" really means as applied to the region at hand. Hubbard and Bayarri (2003) made a case, implicitly, that an NP concept of confidence would have no meaning as applied to the region at hand, that it only applies to a long run of similar intervals. Students, almost invariably, interpret confidence as posterior probability. For example, if we were to flip a coin many times, about half of the time we would get heads. If I flip a coin and look at it but do not tell you the result, you may feel comfortable saying that the chance of heads is still .5 even though I know whether it is heads or tails. Somehow the probability of what is going to happen in the future is turning into confidence about what has already happened but is unobserved. Since I do not understand how this transition from probability to confidence is made (unless one is a Bayesian in which case confidence actually is probability), I do not understand "confidence."

### 5.3 Bayesian Testing

Bayesian tests can go seriously wrong if you pick inappropriate prior distributions. This is the case in Lindley's famous paradox in which, for a seemingly simple and reasonable testing situation involving normal data, the null hypothesis is accepted no matter how weird the observed data are relative to the null hypothesis. The datum is $X|\mu \sim N(\mu, 1)$. The test is $H_0 : \mu = 0$ versus $H_A : \mu > 0$. The priors on the hypotheses do not really matter, but take $\Pr[\mu = 0] = .5$ and $\Pr[\mu > 0] = .5$. In an attempt to use a noninformative prior, take the density of $\mu$ given $\mu > 0$ to be flat on the half line. (This is an improper prior but similar proper priors lead to similar results.) The Bayesian test compares the density of the data $X$ under $H_0 : \mu = 0$ to the average density of the data under $H_A : \mu > 0$. (The latter involves integrating the density of $X|\mu$ times the density of $\mu$ given $\mu > 0$.) The average density under the alternative makes any $X$ you could possibly see infinitely more probable to have come from the null distribution than from the alternative. Thus, anything you could possibly see will cause you to accept $\mu = 0$. Attempting to have a noninformative prior on the half line leads one to a nonsensical prior that effectively puts all the probability on unreasonably large values of $\mu$ so that, by comparison, $\mu = 0$ always looks more reasonable.

## 6. CONCLUSIONS AND COMMENTS

The basic elements of a Fisherian test are: (1) There is a probability model for the data. (2) Multidimensional data are summarized into a test statistic that has a known distribution. (3) This known distribution provides a ranking of the "weirdness" of various observations. (4) The $p$ value, which is the probability of observing something as weird or weirder than was actually observed, is used to quantify the evidence against the null hypothesis. (5) $\alpha$ level tests are defined by reference to the $p$ value.

The basic elements of an NP test are: (1) There are two hypothesized models for the data: $H_0$ and $H_A$. (2) An $\alpha$ level is chosen which is to be the probability of rejecting $H_0$ when $H_0$ is true. (3) A rejection region is chosen so that the probability of data falling into the rejection region is $\alpha$ when $H_0$ is true. With discrete data, this often requires the specification of a randomized rejection region in which certain data values are randomly assigned to be in or out of the rejection region. (4) Various tests are evaluated based on their power properties. Ideally, one wants the most powerful test. (5) In complicated problems, properties such as unbiasedness or invariance are used to restrict the class of tests prior to choosing a test with good power properties.

Fisherian testing seems to be a reasonable approach to model validation. In fact, Box (1980) suggested Fisherian tests, based on the marginal distribution of the data, as a method for validating Bayesian models. Fisherian testing is philosophically based on the idea of proof by contradiction in which the contradiction is not absolute.

Bayesian testing seems to be a reasonable approach to making a decision between alternative hypotheses. The results are influenced by the prior distributions, but one can examine a variety of prior distributions.

Neyman-Pearson testing seems to be neither fish nor fowl. It seems to mimic Fisherian testing with its emphasis on the null hypothesis and small $\alpha$ levels, but it also employs an alternative hypothesis, so it is not based on proof by contradiction as is Fisherian testing. Because NP testing focuses on small $\alpha$ levels, it often leads to bad decisions between the two alternative hypotheses. Certainly, for simple versus simple hypotheses, any problems with NP testing vanish if one is not philosophically tied down to small $\alpha$ values. For example, any reasonable test (as judged by frequentist criteria) must be within both the collection of all most powerful tests and the collection of all Bayesian tests, see Ferguson (1967, p. 204).

Finally, there is the issue of whether $\alpha$ is merely a measure of how weird the data are, or whether is should be interpreted as the probability of making the wrong decision about the null. If $\alpha$ is the probability of making an incorrect decision about the null, then performing multiple tests to evaluate a composite null causes problems because it changes the overall probability of making the wrong decision. If $\alpha$ is merely a measure of how weird the data are, it is less clear that multiple testing inherently causes any problem. In particular, Fisher (1935, chap. 24) did not worry about the experimentwise error rate when making multiple comparisons using his "least significant difference" method in analysis of variance. He did, however, worry about drawing inappropriate conclusions by using an invalid null distribution for tests determined by examining the data.

In recent years, my own classroom presentations have largely abandoned NP ideas when teaching regression, analysis of variance, statistical methods, or almost any applied course. I now teach Fisherian testing and confidence intervals based on Fisherian tests. In theory courses I teach some NP testing because of its historical role and the fact that other statisticians expect students to know it. If I could get away with it, I would teach introductory statistics from a Bayesian point of view. The idea of teaching Bayesian statistics to introductory students was examined by Albert (1997), Berry (1997), and Moore (1997). I firmly believe that Bayesian ideas are easier for students to understand than tests and confidence intervals based on the idea of proof by contradiction, which in turn is easier to understand than NP ideas.

*[Received September 2004. Revised November 2004.]*

## REFERENCES

Albert, J, (1997), "Teaching Bayes' Rule: A Data-Oriented Approach," *The American Statistician*, 51, 247–253.

Berger, J. O. (2003), "Could Fisher, Jeffreys and Neyman have Agreed on Testing?" *Statistical Science*, 18, 1–32.

Berger, J. O., and Wolpert, R. (1984), *The Likelihood Principle*, Hayward, CA: Institute of Mathematical Statistics.

Berry, D. A. (1997), "Teaching Elementary Bayesian Statistics With Real Applications in Science," *The American Statistician*, 51, 241–246.

——— (2004), "Bayesian Statistics and the Efficiency and Ethics of Clinical Trials," *Statistical Science*, 19, 175–187.

Box, G. E. P. (1980), "Sampling and Bayes' Inference in Scientific Modelling and Robustness," *Journal of the Royal Statistical Society*, Ser. A, 143, 383–404.

Christensen, R. (2003). "Significantly Insignificant $F$ Tests," *The American Statistician*, 57, 27–32.

Ferguson, T. S. (1967), *Mathematical Statistics: A Decision Theoretical Approach*, New York: Academic Press.

Fisher, R. A. (1935), *The Design of Experiments* (9th ed., 1971), New York: Hafner Press.

——— (1956), *Statistical Methods and Scientific Inference* (3rd. ed., 1973), New York: Hafner Press.

Hubbard, R., and Bayarri, M. J. (2003), "Confusion Over Measures of Evidence ($p$'s) Versus Errors ($\alpha$'s) in Classical Statistical Testing," *The American Statistician*, 57, 171–177.

Lehmann, E. L. (1997), *Testing Statistical Hypotheses* (2nd ed.), New York: Springer.

Moore, D. (1997), "Bayes for Beginners? Some Reasons to Hesitate," *The American Statistician*, 51, 254–261.