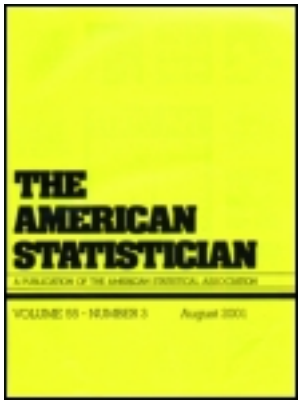
This article was downloaded by: [Yale University Library]
On: 24 February 2013, At: 07:40
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



The American Statistician
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/utas20

P Values: What They are and What They are Not

Mark J. Schervish [a]
[a] Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
Version of record first published: 17 Feb 2012.

To cite this article: Mark J. Schervish (1996): P Values: What They are and What They are Not, The American Statistician, 50:3, 203-206

To link to this article: http://dx.doi.org/10.1080/00031305.1996.10474380

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.tandfonline.com/page/terms-and-conditions

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

P Values: What They Are and What They Are Not

Mark J. SCHERVISH

P values (or significance probabilities) have been used in place of hypothesis tests as a means of giving more information about the relationship between the data and the hypothesis than does a simple reject/do not reject decision. Virtually all elementary statistics texts cover the calculation of P values for one-sided and point-null hypotheses concerning the mean of a sample from a normal distribution. There is, however, a third case that is intermediate to the one-sided and point-null cases, namely the interval hypothesis, that receives no coverage in elementary texts. We show that P values are continuous functions of the hypothesis for fixed data. This allows a unified treatment of all three types of hypothesis testing problems. It also leads to the discovery that a common informal use of P values as measures of support or evidence for hypotheses has serious logical flaws.

KEY WORDS: Evidence; Interval hypothesis; Measure of support; One-sided hypothesis; Point-null hypothesis; Significance probability.

Downloaded by [Yale University Library] at 07:40 24 February 2013

1. INTRODUCTION

Consider an observation X that is thought to have a normal distribution with mean μ and variance 1, denoted $N(\mu, 1)$, conditional on a parameter $\mathrm{M} = \mu$. The usual types of hypotheses concerning M in which one is interested include H_1: $\mathrm{M} = \mu_0$ versus A_1: $\mathrm{M} \neq \mu_0$ (called point-null hypotheses), H_2: $\mathrm{M} \leq \mu_0$ versus A_2: $\mathrm{M} > \mu_0$, and H_3: $\mathrm{M} \geq \mu_0$ versus A_3: $\mathrm{M} < \mu_0$ (called one-sided hypotheses). Other types of hypotheses include interval hypotheses of the form H_4: $\mathrm{M} \in [\mu_1, \mu_2]$ versus A_4: $\mathrm{M} \notin [\mu_1, \mu_2]$. Both point-null and one-sided hypotheses are limits of interval hypotheses either as the endpoints move together or as one endpoint becomes infinite. In Section 2 we show how the P value (or significance probability) is continuous as a function of the hypothesis on the class of all point-null, one-sided, and interval hypotheses. This observation allows us to treat all of the above types of hypotheses as versions of the same kind of hypothesis. In particular, any interpretation that one chooses to give to P values ought to be consistent across the different types of hypotheses because they really are not as different as they might have seemed.

One common interpretation of a P value is that it measures the degree to which the observation $X = x$ supports H or the amount of evidence in favor of H in the data. (See, for example, Lehmann 1975, p. 11. Berkson (1942) and Blyth and Staudte (1995, sec. 4) argue against this interpretation for very different reasons, both of which are different from those given in this article.) In Section 3 we introduce a simple logical condition that should be satisfied by a measure of support, namely that if hypothesis H implies hypothesis H', then there should be at least as much support for H' as there is for H. We then show that the P value fails to meet this condition. In Section 4 we explore ways to modify the interpretation of P values as measures of support.

2. *P* VALUES ARE CONTINUOUS

Assuming that all tests under consideration are uniformly most powerful unbiased (UMPU), then for each hypothesis H and observed data $X = x$ there is associated a significance probability or P value $p_H(x)$. There are several equivalent ways to define P values in well-behaved examples. One way is to define $p_H(x)$ as the probability, in an independent replication of the experiment with data X', that X' is at least as extreme as x given that H is true. Alternatively, one could define $p_H(x)$ as the greatest lower bound on the set of all significance levels α such that we would reject H at level α. For fixed data $X = x$ we can imagine how the P value $p_H(x)$ changes as we focus attention on different hypotheses H. An example is given in Section 3 of a court case in which it was important to consider two different hypotheses concerning the same parameter.

For the remainder of the paper we need to use a slightly more informative notation for P values because we will be considering many different hypotheses of the same form. For $a \in [-\infty, \infty)$ and $b \in [a, \infty]$ let $p_{a,b}(x)$ denote the P value; $a = b$ yields H_1, $a = -\infty$ yields H_2, $b = \infty$ yields H_3, and $a < b$ both finite yields H_4. Let Φ be the standard normal distribution function, and let Φ^{-1} be the standard normal quantile function. Then

$$p_{\mu_0,\mu_0}(x) = 2\Phi(-|x - \mu_0|), \qquad p_{\mu_0,\infty}(x) = \Phi(x - \mu_0),$$

$$p_{-\infty,\mu_0}(x) = \Phi(\mu_0 - x).$$

For interval hypotheses the UMPU level α test is given by Lehmann (1986, sec. 4.2) as rejecting H_4: $\mathrm{M} \in [\mu_1, \mu_2]$ if $|X - .5(\mu_1 + \mu_2)| > c$, where c is chosen so that

$$\Phi(.5[\mu_1 - \mu_2] - c) + \Phi(.5[\mu_2 - \mu_1] - c) = \alpha. \qquad (1)$$

Equation (1) is equal to both P_{μ_1} (reject H_4) and P_{μ_2} (reject H_4), where P_μ means conditional probability given that $\mathrm{M} = \mu$. Notice that (1) is a differentiable strictly decreasing

Mark J. Schervish is Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213. The author thanks Sam Greenhouse for personal communications that led to an in-depth study of interval hypotheses and their associated P values. He also thanks Brian Junker, Peter Imrey, the referees, and an associate editor for helpful comments on earlier drafts. Some of these results were presented at the first Robert Bohrer Memorial Student Workshop in Statistics on November 5, 1994. The author dedicates this work to Bob's memory.

© 1996 American Statistical Association

function of c that equals 1 when $c = 0$ and goes to 0 as $c \to \infty$, so (1) has a unique solution, call it $c(\alpha; \mu_2 - \mu_1)$. Clearly, $c(\alpha; \mu_2 - \mu_1)$ is a continuous strictly decreasing function of α for $\alpha \in (0, 1]$ with $c(1; \mu_2 - \mu_1) = 0$ and $\lim_{\alpha \to 0} c(\alpha; \mu_2 - \mu_1) = \infty$. Using the UMPU test the P value for H_4 and data $X = x$ is

$$p_{\mu_1,\mu_2}(x) = c^{-1}(|x - .5[\mu_1 + \mu_2]|; \mu_2 - \mu_1)$$
$$= \begin{cases} \Phi(x - \mu_1) + \Phi(x - \mu_2) & \text{if } x < .5[\mu_1 + \mu_2] \\ \Phi(\mu_1 - x) + \Phi(\mu_2 - x) & \text{if } x \geq .5[\mu_1 + \mu_2], \end{cases} \quad (2)$$

where $c^{-1}(y; z)$ means that number d such that $c(d; z) = y$. Equation (2) can be verified by setting $c = |x - .5[\mu_1 + \mu_2]|$ in (1) and noting that the resulting α is $p_{\mu_1,\mu_2}(x)$.

It is a simple matter to notice that $(-\infty, \mu_0] = \cup_{a \leq \mu_0}[a, \mu_0]$. Because the collection of sets $[a, \mu_0]$ is monotone increasing as $a \to -\infty$ for $a \leq \mu_0$, one often writes $(-\infty, \mu_0] = \lim_{a \to -\infty}[a, \mu_0]$. We can now show that the P values obey the same limit. That is, $\lim_{a \to -\infty} p_{a,\mu_0}(x) = p_{-\infty,\mu_0}(x)$. This is easily verified by setting $\mu_2 = \mu_0$ in (2) and letting $\mu_1 \to -\infty$. Eventually, the bottom row is in effect, and $\lim_{a \to -\infty} p_{a,\mu_0}(x) = \Phi(\mu_0 - x) = p_{-\infty,\mu_0}(x)$ for all x. Similarly, $\lim_{b \to \infty} p_{\mu_0,b}(x) = p_{\mu_0,\infty}(x)$ for all x.

At the other extreme consider $\lim_{(\mu_1,\mu_2) \to (\mu_0,\mu_0)} p_{\mu_1,\mu_2}(x)$. If $x < \mu_0$, then the top row of (2) eventually takes effect and the limit is $2\Phi(x - \mu_0) = p_{\mu_0,\mu_0}(x)$. Similarly, if $x > \mu_0$, then the limit is $2\Phi(\mu_0 - x) = p_{\mu_0,\mu_0}(x)$. For $x = \mu_0$, both the top and bottom rows of (2) go to $1 = p_{\mu_0,\mu_0}(\mu_0)$. For intermediate cases, because (2) is continuous as a function of (μ_1, μ_2), it is easy to see that $\lim_{(\mu_1,\mu_2) \to (a,b)} p_{\mu_1,\mu_2}(x) = p_{a,b}(x)$. Finally, if both $\mu_1 \to -\infty$ and $\mu_2 \to \infty$, then both rows of (2) go to $1 = p_{-\infty,\infty}(x)$ for all x.

What we have established is that $p_{\mu_1,\mu_2}(x)$ is continuous as a function of (μ_1, μ_2) even as $\mu_1 \to -\infty$ and/or $\mu_2 \to \infty$, or $|\mu_1 - \mu_2| \to 0$. Just as the point-null and one-sided hypotheses are limits of interval hypotheses, so too are their P values limits of the P values of the interval hypotheses for every data value. This observation allows us to think of point-null hypotheses as approximations to interval hypotheses. If $\mu_2 - \mu_1 > 0$ is sufficiently small and $\mu_0 = .5(\mu_1 + \mu_2)$, then $p_{\mu_0,\mu_0}(x)$ will be close to $p_{\mu_1,\mu_2}(x)$. Also, one-sided hypotheses can be thought of as approximations to very large interval hypotheses.

The continuity of P values as functions of the hypothesis is not restricted to normal distributions. For example, straightforward but tedious calculation shows that continuity also holds when X has exponential, binomial, or uniform distribution.

3. *P* VALUES ARE NOT MEASURES OF SUPPORT

The P value can be used to test hypotheses in the usual fashion. After one calculates and reports $p_H(x)$, a person with a favorite α value can reject H at level α if $p_H(x) < \alpha$. Because larger values of $p_H(x)$ make it harder to reject H, $p_H(x)$ has often been suggested as a measure of the support that the observed data $X = x$ lend to H, or the amount of evidence in favor of H. This suggestion is always informal, and no theory is ever put forward for what properties a measure of support or evidence should have. The suggestion has been very successful in simple problems, perhaps due to several well-known facts. For example, as a function of x, $p_{\mu_0,\mu_0}(x)$ decreases as x moves away from μ_0, expressing the desired property that the further the observation is from the hypothesis, the less support it lends to the hypothesis. We could also say that $p_{\mu_0,\mu_0}(x)$ decreases as μ_0 moves away from x, expressing the equally desirable property that the further the hypothesis is from the observation, the less support the data lend to the hypothesis. Similarly, $p_{-\infty,\mu_0}(x)$ is an increasing function of μ_0 expressing the desirable property that as the hypothesis covers more of the parameter space, the support for the hypothesis increases.

Because posterior probabilities are intended for measuring support for hypotheses when the data are fixed (the true state of affairs after the data are observed), many authors have considered the extent to which P values can be interpreted as posterior probabilities. Notable among this group are DeGroot (1973), Casella and Berger (1987), Berger and Sellke (1987), and Hodges (1992). Fisher (1935) introduced fiducial distributions as a means of producing probabilities on the parameter space without using Bayesian reasoning. Our goal in this section is much more modest. We borrow a simple logical condition from the theory of multiple comparisons, and show why a measure of support should satisfy this condition. We then demonstrate that P values do not satisfy the condition. Nevertheless, it is useful to recall one result concerning the connection between P values and posterior and fiducial distributions. If one uses the improper prior distribution for M with constant density, then the posterior distribution of M given $X = x$ is $N(x, 1)$ (Box and Tiao 1973, sec. 1.31). Similarly, the fiducial distribution of M after observing $X = x$ is $N(x, 1)$. It follows, for example, that the posterior and fiducial probabilities that M $\geq \mu_0$ equal $1 - \Phi(\mu_0 - x) = p_{\mu_0,\infty}(x)$. Similarly, the posterior and fiducial probabilities that M $\leq \mu_0$ equal $p_{-\infty,\mu_0}(x)$. That is, the posterior and fiducial probabilities of one-sided hypotheses are equal to the corresponding P values.

These posterior and fiducial probabilities for one-sided hypotheses also have desirable properties for a measure of support, such as

- The farther the hypothesis is from the data, the less support there is;
- The farther into the hypothesis the data are, the more support there is;
- The larger the hypothesis is, the more support there is.

The third property above is an analog to the concept of coherence used in multiple comparisons as introduced by Gabriel (1969). Suppose that one hypothesis H implies another H'. Then tests of H and H' are coherent if rejection of H' always entails rejection of H. We can say that a measure of support for hypotheses is coherent if, whenever H implies H', the measure of support for H' is at least as large as the measure of support for H. That is, any support for

Downloaded by [Yale University Library] at 07:40 24 February 2013

H must a fortiori be support for H'. As an example there must be at least as much support for H: $M \leq 3$ as there is for H': $M \leq 2$. For these one-sided hypotheses P values behave coherently.

In general, however, P values are incoherent as measures of support in the sense just described. For example, suppose that $x > \mu_0$. Then $p_{-\infty,\mu_0}(x) = .5p_{\mu_0,\mu_0}(x)$ even though H_1: $M = \mu_0$ implies H_2: $M \leq \mu_0$. There may be several reasons why this incoherence has not been very troubling. First, people rarely consider point-null and one-sided hypotheses in the same problem. A notable exception appears in the case of E.E.O.C. vs. Federal Reserve Bank of Richmond. (See Russell 1983, para. [15], pp. 652–654.) In the lively exchange the plaintiff's statistical expert tries to explain to a judge why one should use a one-sided test (with P value .037 in this example) rather than a two-sided test (with P value .074). The significance of the choice of hypothesis was quite apparent to the judge.

Another reason that the incoherence may not be a sore point is that some statisticians believe that one-sided and point-null hypotheses are of a sufficiently different nature that they simply should not be compared. In Section 2 we showed how these hypotheses lie at opposite ends of a continuum of hypotheses bridged by the interval hypotheses. In a sense the interval hypotheses form the missing link between the one-sided and point-null cases. The two extremes really are not such different objects as one might have thought.

In addition to bridging the gap between one-sided and point-null hypotheses, interval hypotheses help to cast even more doubt on the ability of P values to measure support for hypotheses. For $\mu_1, \mu_2 < x$ it follows easily from (2) that $p_{\mu_1,\mu_2}(x)$ is a strictly increasing function of both μ_1 and μ_2. Pick arbitrary $\mu_1 < \mu_2 < x$. Let $\mu_1' < \mu_1$. Then $p_{\mu_1',\mu_2}(x) < p_{\mu_1,\mu_2}(x)$. Because $p_{\mu_1',b}(x)$ is continuous in b, there exists $\mu_2' \in (\mu_2, x)$ such that

$$p_{\mu_1',\mu_2'}(x) - p_{\mu_1',\mu_2}(x) < p_{\mu_1,\mu_2}(x) - p_{\mu_1',\mu_2}(x).$$

Simplifying this equation yields $p_{\mu_1',\mu_2'}(x) < p_{\mu_1,\mu_2}(x)$, which means that the P values are incoherent as measures of support because $[\mu_1, \mu_2]$ is a proper subset of $[\mu_1', \mu_2']$.

As a numerical example let $x = 2.18, \mu_1 = -.5$, and $\mu_2 = .5$. From (2) we calculate

$$p_{-.5,.5}(2.18) = \Phi(-2.68) + \Phi(-1.68) = .0502.$$

If we let $\mu_1' = -.82$ and $\mu_2' = .52$, then (2) gives

$$p_{-.82,.52}(2.18) = \Phi(-3) + \Phi(-1.66) = .0498.$$

If we use the P value as a measure of support for the hypothesis, we are saying that there is more support for the hypothesis H: $M \in [-.5, .5]$ than there is for H': $M \in [-.82, .52]$ even though H implies H'.

Because hypothesis tests are so closely related to P values, it is not surprising to learn that the incoherence of P values as measures of support reflects the incoherence of level α tests for multiple hypotheses in the sense of Gabriel (1969). Suppose that someone tries to test H and H' both at level .05. (Perhaps they are using a Bonferroni multiple comparison procedure with $\alpha = .1$ split equally between the two hypotheses.) This person will reject H', but cannot reject H after observing $X = 2.18$. They are now confident that M is outside the interval $[-.82, .52]$, but still must act as if M is inside $[-.5, .5]$.

We have shown that interval P values are incoherent with each other as measures of support for their respective hypotheses. It is also possible to show that they are incoherent when considered simultaneously with one-sided and/or point-null P values. For example, $p_{.5,.5}(2.18) = .0930$, which is larger than both $p_{-.5,.5}(2.18)$ and $p_{-.82,.52}(2.18)$ calculated earlier. Hence it is wrong to claim that P values give a measure of the support that the data lend to the hypothesis without further restrictions. Just as the continuity of P values extends to distributions other than normal, so too does the incoherence of P values as measures of support.

4. WHAT CAN *P* VALUES MEASURE?

Two (rather loose) definitions of the P value were given in Section 2. When made precise these interpretations of P values are valid. What we have shown to be invalid is the use of P values to measure support or evidence for hypotheses.

However, one could ask if there is a way to modify or reinterpret the P value so that it can stand for something related to a measure of support. One simple-minded answer is to define a measure of support as $p'_{a,b}(x) = \sup_{\theta \in [a,b]} p_{\theta,\theta}(x)$. Such a measure of support would be coherent, although it would give the same support, namely 1, to all hypotheses H: $M \in [a, b]$ such that $a \leq x \leq b$. This would not be particularly useful.

An approach based on the reasoning behind significance tests is to try to interpret the P value for a fixed hypothesis as a function of the data (presumably prior to observing the data). In this way one can try to think of the P values for different values of x as the different degrees to which different data values would support a single hypothesis H. This might work as long as we do not acknowledge the possibility of other hypotheses. For example, this approach would not be available in the case of E.E.O.C. vs. Federal Reserve Bank of Richmond because the court was confronted with two different hypotheses and only one data set. Another serious drawback to this approach is that the scale on which support is measured is not absolute, but rather depends on the hypothesis. (Also see Berry 1990, sec. 4.3.) For example, suppose that one person tries to measure the support for H: $M = .9$, and another person tries to measure the support for H': $M \leq 1$. After they both observe the same data $X = 2.7$, the first person calculates $p_{.9,.9}(2.7) = .0718$, and the second person calculates $p_{-\infty,1}(2.7) = .0446$. Surely the data offer more support for H' than for H, so .0718 must reflect a lower level of support for a point-null hypothesis than .0446 reflects for a one-sided hypothesis. In the numerical example in Section 3 each of the interval hypotheses needs its own scale of support, even though they are both of the interval type. In short, the interpretation of a

Downloaded by [Yale University Library] at 07:40 24 February 2013

particular value on the scale of support, such as the popular .05, must vary with the hypothesis.

Another interesting approach is suggested by the observation that the examples of incoherence cited in this paper involve situations in which the larger hypothesis contains parameter values that are farther away (in a sense to be made precise momentarily) from the data than the smaller hypothesis. For example, let H: $\mathrm{M} \in [-.5, .5]$ and H': $\mathrm{M} \in [-.82, .52]$ with data $x = 2.18$. The values in the interval $[-.82, -.5)$ contained in H' but not in H are certainly farther from the observed data than the values in H. One might try to argue that H' should be "penalized" for containing extra values that are farther away from the data. This idea sounds appealing at first, but one then notices that one-sided P values do not penalize larger hypotheses when they contain farther away values. For example, $p_{-\infty,\mu_0}(x)$ increases as a function of μ_0 even when μ_0 is quite far from x. But in this case the added values (near μ_0) are not as far away from x as some values (near $-\infty$) already in the hypothesis. Suppose that we define "farther away" as follows.

Definition 1. Let H be a hypothesis, and let θ' not be part of H. Suppose that for every θ in $H, p_{\theta',\theta'}(x) < p_{\theta,\theta}(x)$. Then say that θ' is farther away from x than H is.

For normal distributions, when H implies H' and the additional points in H' not in H are farther away from x than H is, then $p_{H'}(x) < p_H(x)$. Unfortunately, even this does not happen in all cases. If X has $U(0, \theta)$ distribution, then one can easily check that

$$p_{a,b}(x) = \begin{cases} x/a & \text{if } x \le a \\ (b-x)/(b-a) & \text{if } a < x \le b \\ 0 & \text{if } x > b, \end{cases}$$

even for $a = b, a = 0$, and/or $b = \infty$. Suppose that H: $\mathrm{M} \in [a, b]$ and H': $\mathrm{M} \in [a, b']$ with $b < b'$. If $X = x < a$ is observed (i.e., x has positive density given every θ in both H and H'), then $p_{a,b}(x) = p_{a,b'}(x)$ even though every $\theta' > b$ is farther away from x than H is. (Note that for $\theta' > b, p_{\theta',\theta'}(x) = x/\theta' < x/\theta = p_{\theta,\theta}(x)$, if $\theta \in [a, b]$.)

Even if one were able to construct a consistent measure of penalized support for hypotheses, one would have to remember that a low measure of penalized support would not mean that there is evidence against the hypothesis, but rather that the hypothesis contained at least some parameter values that are not supported by the data. It might still contain some parameter values that are highly supported by the data. Even so, the P value does not provide a measure of penalized support. In summary, we have been unable to construct a consistent interpretation of the P value as anything similar to a measure of support for its hypothesis.

5. DISCUSSION

We have shown that one-sided and point-null hypotheses are not two different objects that should never be compared, but rather they are just different versions of the same object of which interval hypotheses are versions as well. For nice data distributions the P value is continuous as a function of the hypothesis. We have also seen that P values cannot be interpreted as measures of support for their respective hypotheses. One could try to argue, with normal distributions at least, that the P value penalizes hypotheses that contain additional parameter values that are far away from the data, but even this argument fails in the case of uniform distributions.

The reporting of P values is very common in applied statistics, usually in multiparameter problems that were not considered in this paper. However, given that P values cannot be interpreted as measures of support in the simple problems of this article, one would suspect that their interpretation as measures of support in more complicated cases would be suspect as well.

[Received August 1994. Revised September 1995.]

REFERENCES

Berger, J. O., and Sellke, T. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of P-Values and Evidence" (with comments), *Journal of the American Statistical Association*, 82, 112–122.

Berkson, J. (1942), "Tests of Significance Considered as Evidence," *Journal of the American Statistical Association*, 37, 325–335.

Berry, D. A. (1990), "Basic Principles in Designing and Analyzing Clinical Studies," in *Statistical Methodology in the Pharmaceutical Sciences*, ed. D. A. Berry, New York: Marcel Dekker, pp. 1–55.

Blyth, C. R., and Staudte, R. G. (1995), "Estimating Statistical Hypotheses," *Statistics and Probability Letters*, 23, 45–52.

Box, G. E. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, Reading, MA: Addison-Wesley.

Casella, G., and Berger, R. L. (1987), "Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem" (with comments), *Journal of the American Statistical Association*, 82, 106–111.

DeGroot, M. H. (1973), "Doing What Comes Naturally: Interpreting a Tail Area as a Posterior Probability or as a Likelihood Ratio," *Journal of the American Statistical Association*, 68, 966–969.

Fisher, R. A. (1935), "The Fiducial Argument in Statistical Inference," *Annals of Eugenics*, 6, 391–398.

Gabriel, K. R. (1969), "Simultaneous Test Procedures—Some Theory of Multiple Comparisons," *The Annals of Mathematical Statistics*, 40, 224–250.

Hodges, J. (1992), "Who Knows What Alternative Lurks in the Hearts of Significance Tests?" (with comments), in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Clarendon Press, pp. 247–266.

Lehmann, E. L. (1975), *Nonparametrics: Statistical Methods Based on Ranks*, San Francisco: Holden-Day.

——— (1986), *Testing Statistical Hypotheses* (2nd ed.), New York: John Wiley.

Russell, D. (1983), "Equal Employment Opportunity Commission v. Federal Reserve Bank of Richmond," *698 Federal Reporter 2d Series*, 633–675, United States Court of Appeals, Fourth Circuit.

Downloaded by [Yale University Library] at 07:40 24 February 2013