A Significance Paradox

By

Mark J. Schervish[1]

Technical Report No. 598 Revised

Department of Statistics
Carnegie-Mellon University
Pittsburgh, PA 15213

August 2, 1994

[1]Mark J. Schervish is Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213. The author would like to thank Sam Greenhouse for personal communications which led to an in-depth study of interval hypotheses and their associated $P-$values. He would also like to thank Brian Junker for helpful comments on an earlier draft.

Abstract

Significance probabilities (or $P-$values) 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/don't reject decision. Virtually all elementary statistics texts cover the calculation of $P-$values for hypotheses concerning the mean of a sample from a normal distribution. Both one-sided hypotheses and point-null hypotheses are usually discussed. But there is a third case which is intermediate to the one-sided and point-null cases, namely the interval hypothesis, which receives no coverage in elementary texts. We show here that $P-$values for interval hypotheses exhibit a paradoxical sort of behavior which was actually visible in the one-sided and point-null cases, but which has been largely ignored in the past.

Keywords: Measure of support, $P-$value, one-sided hypothesis, point-null hypothesis, interval hypothesis, significance probability.

1 Introduction

To begin, consider an observation X which is thought to have a normal distribution with mean μ and variance 1, denoted $N(\mu, 1)$, conditional on a parameter $\mathrm{M} = \mu$. Most elementary texts on statistics discuss tests of hypotheses of the form $H_1 : \mathrm{M} = \mu_0$ vs. $A_1 : \mu \neq \mu_0$ (called *point-null hypotheses*) and hypotheses of the form $H_2 : \mathrm{M} \leq \mu_0$ vs. $A_2 : \mathrm{M} > \mu_0$ or of the form $H_3 : \mathrm{M} \geq \mu_0$ vs. $A_3 : \mathrm{M} < \mu_0$ (called *one-sided hypotheses*.) These hypotheses can be tested by comparing $X - \mu_0$ to appropriate quantiles of the standard normal distribution. For example, to test H_1 at level $\alpha \in (0, 1)$, one rejects H_1 if $|X - \mu_0| > \Phi^{-1}(1 - .5\alpha)$, where Φ stands for the standard normal CDF and Φ^{-1} is the standard normal quantile function. To test H_3 at level α, one rejects H_3 if $X - \mu_0 < \Phi^{-1}(\alpha)$.

The *significance probability* (or $P-$value) for a hypothesis H and an observed data value $X = x$ can be defined in various equivalent ways. For example, one could define it as the probability $p_H(x)$, in an independent replication of the experiment with data X', that X' is at least as extreme than x given that H is true. What counts as extreme depends on how one would propose to test the hypothesis. For example, with H_1 and the test given above, $p_{H_1}(x) = P_{\mu_0}(|X' - \mu_0| > |x - \mu_0|)$, where P_{μ_0} means conditional probability given that $\mathrm{M} = \mu_0$. For H_3,

$$p_{H_3}(x) = P_{\mu_0}(X' < x) = \Phi(x - \mu_0).$$

"At least as extreme" always means "at least as far into the set of values where we reject the hypothesis." There are certainly more precise ways to say what "at least as extreme" means, but the examples in this paper will be sufficiently simple that we will not need anything more precise.

The $P-$value has the advantage over a level α test that one need not pick a value of α ahead of time. One can calculate $p_H(x)$ and then each person who has their own favorite α value can reject H if $p_H(x) < \alpha$. Since 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 which the observed data $X = x$ lends to H. (See, for example, Lehmann 1975, p. 11 and Berkson 1942.) This suggestion is always informal, and no theory is ever put forward for what properties a measure of support 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, $p_{H_1}(x)$ decreases as x moves away from μ_0, expressing the desired property that the further the data is from the hypothesis, the less support the data lends to the hypothesis. We could also say that $p_{H_1}(x)$ decreases as μ_0 moves away from x, expressing the equally desirable property that the further the hypothesis is from the data, the less support the data lends to the hypothesis.

By considering the data as fixed and the parameter as unknown (the true state of affairs after data are observed) we are led to the recent discussion between Casella and Berger (1987) on the one hand and Berger and Sellke (1987) on the other. Without taking sides in this discussion, I will merely point out that these authors were considering to what extent $p_H(x)$ could behave like a Bayesian posterior probability of an event like $\{H \text{ is true}\}$. For example, without going into details which can be found in virtually any text on Bayesian inference like Box and Tiao (1973, Section 1.31.), with the popular flat improper prior for M, the posterior distribution of M given $X = x$ is $N(x, 1)$. It follows, for example, that the posterior probability that H_3 is true equals $1 - \Phi(\mu_0 - x) = p_{H_3}(x)$. Similarly, the posterior probability that H_2 is true equals $p_{H_2}(x)$. Such probabilities are also available without appeal to any Bayesian argument as part of the fiducial distribution of M as introduced by Fisher (1935). That is, since $X - \mathrm{M} \sim N(0, 1)$ (given M), we have $\Pr(X - \mathrm{M} \leq c) = \Phi(c)$ for all c. This implies $\Pr(\mathrm{M} \leq c - X) = \Phi(c)$, so we can pretend like M has $N(X, 1)$ distribution.

These posterior or fiducial probabilities also have desired properties for a measure of support like

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

The first two properties are easy to understand when the parameter is of the location or scale type. The last property is an analog to the concept of *coherence* used in multiple comparisons. If one hypothesis H implies another H', then tests of H and H' are coherent if rejection of H' always entails rejection of H. In terms of support, any support for H must *a fortiori* be support for H'. As an example, if I feel that there is much support for $H : \mathrm{M} \leq 2$, then surely there is at least as much support for $H' : \mathrm{M} \leq 3$. Similarly, if there is very little support for H', then there can't be any more support for H.

In elementary texts, almost no attention is paid to hypotheses of the form $H_4 : \mathrm{M} \in [\mu_1, \mu_2]$ vs. $A_4 : \mathrm{M} \notin [\mu_1, \mu_2]$. Lehmann (1986, Section 4.2) does cover this case, and he shows that the uniformly most powerful unbiased (UMPU, the traditional optimality criterion) level α test is to reject H_4 if $|X - .5(\mu_1 + \mu_2)| > c$, where c is chosen so that

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

The left-hand side of (1) is equal to both $P_{\mu_1}(\text{reject } H_4)$ and $P_{\mu_2}(\text{reject } H_4)$. Notice that the left-hand side of (1) is a differentiable strictly decreasing function of c which 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 $\lim_{\alpha \to 0} c(\alpha; \mu_2 - \mu_1) = \infty$. Here are some values of $c(\alpha; 1)$ for some of the favorite α values:

α	.05	.025	.01	.005	.001
$c(\alpha; 1)$	2.184	2.485	2.842	3.087	3.596

The $P-$value for H_4 and data $X = x$ is

$$p_{H_4}(x) = c^{-1}(|x - .5[\mu_1 + \mu_2]|; \mu_2 - \mu_1) = \begin{cases} \Phi(x - \mu_1) + \Phi(x - \mu_2) & \text{if } x < .5[\mu_1 + \mu_2] \\ \Phi(\mu_1 - x) + \Phi(\mu_2 - x) & \text{if } x \geq .5[\mu_1 + \mu_2]. \end{cases} \tag{2}$$

This can be verified easily using (1). Hypotheses such as H_4 can be used when interest lies in whether or not M is sufficiently close to some number like $\mu_0 = .5(\mu_1 + \mu_2)$ rather than in whether or not M is exactly equal to μ_0. Perhaps one could argue that the point-null hypothesis H_1 is an approximation to H_4 which frees (prevents) the user from having to decide how far apart to choose μ_1 and μ_2. In a sense, to be demonstrated later, this is correct. However, by neglecting serious consideration of interval hypotheses, we have overlooked a surprising feature in the behavior of $p_H(x)$ as a function of H.

2 Continuity of $P-$values

For fixed data $X = x$, we can imagine how the $P-$value $p_H(x)$ changes as we focus attention on different hypotheses H. One feature which has been well-known for many years, but has caused little consternation, is that the one-sided and point-null $P-$values are incoherent as measures of support in the sense described earlier. That is, for example, if $x > \mu_0$ then $p_{H_2}(x) = .5p_{H_1}(x)$ even though H_1 implies H_2. 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. v. Federal Reserve Bank of Richmond, 698 F.2d 633 (1983) (698 Federal Reporter 2d Series (1983), paragraph [15], p. 652ff). In this lively exchange, a statistical expert tries to explain to a judge how one uses one's opinions to decide between a point-null and a one-sided hypothesis. The fact that the one-sided $P-$value was half the size of the point-null $P-$value was quite clear to the judge, even if nothing else was. Another reason that the incoherence may not be a sore point is that some statisticians believe that the one-sided and point-null hypotheses are of sufficiently different nature that they simply should not be compared. In this section, we will show 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. They really are not such different objects as one might have thought. They are just different versions of the same object.

For the remainder of the paper, we need to use a slightly more informative notation for $P-$values, since we will be considering many different hypotheses of the same form. For the one-sided case $H_2 : \mathrm{M} \leq \mu_0$, let $p_{-\infty,\mu_0}(x)$ denote the $P-$value. For $H_1 : \mathrm{M} = \mu_0$, use $p_{\mu_0,\mu_0}(x)$. For $H_3 : \mathrm{M} \geq \mu_0$, use $p_{\mu_0,\infty}(x)$, and for $H_4 : \mu_1 \leq \mathrm{M} \leq \mu_2$, use $p_{\mu_1,\mu_2}(x)$. It is a simple matter to notice that $(-\infty, \mu_0] = \bigcup_{a \leq \mu_0} [a, \mu_0]$. Since the collection of sets $[a, \mu_0]$ is monotone increasing as $a \to -\infty$ for $a \leq \mu_0$, one often writes $(-\infty, \mu_0] = \lim_{a \to -\infty} [a, \mu_0]$. Perhaps surprisingly, 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, since (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 $\mu_2 \to \infty$. 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 is what we meant earlier, when we said that point-null hypotheses of the form H_1 were approximations to interval hypotheses of the form H_4. If $\mu_2 - \mu_1$ is sufficiently small and $\mu_0 = .5(\mu_1 + \mu_2)$, then $p_{\mu_0,\mu_0}(x)$ will be close to $p_{\mu_1,\mu_2}(x)$.

3 More Incoherence

Those who were not troubled by the incoherence of one-sided and point-null $P-$values as measures of support might find troubling the same kind of incoherence for interval 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 $p_{\mu_1',\mu_2}(x) < p_{\mu_1,\mu_2}(x)$. Since $p_{a,b}(x)$ 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 is incoherent as measures of support since $[\mu_1, \mu_2]$ is a proper subset of $[\mu_1', \mu_2']$.

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

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

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 $H : \mathrm{M} \in [-.5, .5]$ than there is for $H' : \mathrm{M} \in [-.82, .52]$ even though H implies H'. Put another way, if person A tries to test H and person B tries to test H', both at level .05, then person B will reject H', but person A cannot reject H when they both use the same data. Persons A and B now both report their findings to person C. Person C is now confident that M is outside of the interval $[-.82, .52]$ but still must act as if M is inside of $[-.5, .5]$ (at least at level .05). Person C can now empathize with the judge in E.E.O.C. v. Federal Reserve Bank of Richmond. One might argue that, in testing two hypotheses with the same data, one should use a multiple comparisons procedure. If that is the case, then set $\alpha = .1$ in the previous example, and use the Bonferroni method to divide the type I error equally between the two hypotheses.

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 $p_{.5,.5}(2.18) = 0.0930$, 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 which the data lends to the hypothesis without further restrictions.

4 Other Distributions

Although normal distributions are used widely in introductory texts and computer packages, there are other distributions for data. The incoherence of $P-$values (as measures of support) illustrated with normal distributions extends to other distributions as well. In this section, we will briefly sketch the arguments for three other distributions, the exponential, binomial, and uniform.

4.1 Exponential Distributions

Suppose that X has exponential distribution with mean θ conditional on $\Theta = \theta$, denoted $Exp(1/\theta)$. That is, the density of X is $\exp(-x/\theta)/\theta$, for $x > 0$. We will use the same notation as before to denote $P-$values, namely $p_{a,b}(x)$ to stand for the $P-$value for hypothesis $H : \Theta \in [a, b]$ vs. $A : \Theta \notin [a, b]$ and observed data $X = x$. Now, however, $a \geq 0$ is required. Since $\Theta = 0$ is not allowed, we will take intervals of the form $[0, b]$ to stand for one-sided hypotheses of the form $H : \Theta \leq b$. If $a < b$, the UMPU level α test of $H : \Theta \in [a, b]$ vs. $A : \Theta \notin [a, b]$ is to reject H if $X > c_2$ or $X < c_1$ where $c_1 < c_2$ are the solutions to the equations

$$\alpha = 1 - \exp(-c_1/a) + \exp(-c_2/a) = 1 - \exp(-c_1/b) + \exp(-c_2/b). \tag{3}$$

To find the $P-$value, say for $X = x > b$, set $c_2 = x$ in (3) and discover that

$$\exp(-c_1/a) - \exp(-c_1/b) = \exp(-x/a) - \exp(-x/b), \tag{4}$$

which can be solved for c_1. Call the solution $c_1(x; a, b)$. Then

$$p_{a,b}(x) = 1 - \exp(-c_1(x; a, b)/a) + \exp(-x/a).$$

For the case $a = b$, there is only one equation in (3). The second equation to be solved is

$$c_2 \exp(-c_2/b) = c_1 \exp(-c_1/b). \tag{5}$$

Notice that as $a \uparrow b$ in (4), both sides of (4) approach 0, but taking derivatives of both sides with respect to a and evaluating at $a = b$ gives (5). So, the point-null case is still the limit of the interval cases. If we let $a \to 0$ in (4), then clearly $c_1(x; a, b) \to 0$ as well. In fact, $\exp(-x/a) - \exp(-c_1(x; a, b)) \to 1 - \exp(-x/b)$, so that $\lim_{a\to 0} p_{a,b}(x) = \exp(-x/b) = p_{0,b}(x)$. In addition, one can show that $p_{a,b}(x)$ is continuous in (a, b), but we will not demonstrate all of the cases. Now, note that $p_{b,b}(x) > p_{0,b}(x)$ for $x > b$. Once again, one-sided and point-null $P-$values are incoherent as measures of support.

As a numerical example, let $x = 7$, $a = 1$, and $b = 2$. Then $c_1(7; 1, 2) = 0.0613$ and $p_{1,2}(7) = 0.0604$. Now, let $a' = .5$ and $b' = 2.1$ so that $c_1(7, .5, 2.1) = .02411$ and $p_{.5,2.1}(7) = 0.0471$. Once again, we would make conflicting decisions at level .05. It is informative to note that the one-sided $P-$values ($p_{0,b}(x)$ and $p_{a,\infty}(x)$) both equal the fiducial probabilities of their corresponding hypotheses derived from the fact that $X/\Theta \sim Exp(1)$. This means that $\Pr(X/\Theta \geq c) = \exp(-c)$ for all $c > 0$, so $\Pr(\Theta \leq x/c) = \exp(-c)$. It follows that the fiducial CDF of Θ is $F(\theta) = \exp(-x/\theta)$, for $\theta > 0$. As in the case of normal distributions, the one-sided $P-$values will have all the desired properties of a measure of support, but the interval $P-$values will not.

4.2 Binomial Distributions

If X has binomial distribution $Bin(n, \theta)$ given $\Theta = \theta$, the UMPU level α test might be a randomized test. It is common, in calculation of $P-$values, to consider only non-randomized tests. In this case, it is easy to verify that $p_{a,a}(x) \geq p_{0,a}(x)$ with strict inequality in many cases, so that one-sided and point-null $P-$values are incohrent as measures of support. For example, if $n = 10$, $a = .5$, and $x = 8$, we have $p_{0,.5}(8) = 0.0547$, while $p_{.5,.5}(8) = 0.1094$. Incoherence occurs with interval hypotheses as well. Without going into details, suppose that $n = 10$, $x = 8$, $a = .45$, and $b = .55$. Then $p_{.45,.55}(8) = .1270$, the sum of $\binom{10}{y}\theta^y(1-\theta)^{n-y}$ for $y \in \{0, 1, 2, 8, 9, 10\}$ with $\theta = .45$ or $\theta = .55$. If we consider $a' = .3268$ and $b' = .56$, then $p_{.3268,.56}(8) = .1149$, the sum of $\binom{10}{y}\theta^y(1-\theta)^{n-y}$ for $y \in \{0, 1, 8, 9, 10\}$ with $\theta = .3268$ or $\theta = .56$. Once again, the measures of support are incoherent.

It is not the case that all one-sided $P-$values, in the binomial situation, are probabilities of their respective hypotheses calculated from a single distribution over the parameter space. This is easily seen by example. If $n = 10$ and $x = 8$ and we try to test $H : \Theta \geq .5$, we get a $P-$value $p_{.5,1}(8) = 0.9893$. If we add this to $p_{0,.5}(8) = 0.1094$, calculated earlier, we get 1.0439. The excess, 0.0439, is the probability that $X' = 8$ given $\Theta = .5$, which is included in both $P-$values. If it were the case that $p_{.5,1}(8) = P([.5, 1])$ and $p_{0,.5}(8) = P([0, .5])$ for some probability P, then it would have to be the case that $P(\{.5\}) = .0439$.

This is not itself a contradiction. However, we can now repeat these calculations for $p_{0,.4}(8)$ and $p_{.4,1}(8)$, etc. Eventually, we would have a finite set of θ values whose probabilities added up to more than 1. On the other hand, all of the one-sided $P-$values of the form $p_{0,a}(x)$ can be computed using the distribution $Beta(x, n-x+1)$ for Θ, and all of the ones of the form $p_{a,1}(x)$ can be computed using the distribution $Beta(x+1, n-x)$ for Θ.

4.3 Uniform Distributions

For the case in which X has uniform distribution on the interval $[0, \theta]$ given $\Theta = \theta$, denoted $U(0, \theta)$, the situation is much different. Uniformly most powerful (UMP) tests exist for all of the hypotheses considered in this paper. For $H : \Theta \in [a, b]$ vs. $A : \Theta \notin [a, b]$, the UMP level α test is to reject H if $X < \alpha a$ or if $X > \alpha a + (1-\alpha)b$. This is true even if $a = b$. It is easy to compute

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

which is true even if $a = b$ (since the middle condition never holds if $a = b$.) The one-sided UMP tests are to reject $H : \Theta \le b$ if $X > (1-\alpha)b$ and to reject $H : \Theta \ge a$ if $X < a\alpha$. (Actually, there are many UMP level α tests of $H : \Theta \le b$, but the one given above has the smallest power function for $\theta \le b$.) The corresponding $P-$values are

$$\begin{aligned} p_{0,b}(x) &= \begin{cases} 1 - x/b & \text{if } x \le b \\ 0 & \text{if } x > b \end{cases} \\ p_{a,\infty}(x) &= \begin{cases} x/a & \text{if } x \le a \\ 1 & \text{if } x > a. \end{cases} \end{aligned}$$

Note that $p_{0,b}(x)$ and $p_{a,\infty}(x)$ are the appropriate limiting values of $p_{a,b}(x)$. The one-sided $P-$values are also the same as the fiducial probabilities derived from the fact that $X/\Theta \sim U(0,1)$. This means that $\Pr(X/\Theta \ge c) = 1 - c$ for all $c \in (0,1)$, so $\Pr(\Theta \le X/c) = 1 - c$. The fiducial CDF of Θ is $F(\theta) = 1 - x/\theta$, for $\theta > x$.

If $a < x \le b$, we have that $p_{a,b}(x)$ is an increasing function of both a and b, so we will have the type of incoherence noted earlier. For example, if $a = 1$, $b = 3$, and $x = 2.88$, then $p_{1,3}(2.88) = .06$. If $a' = .015$ and $b' = 3.015$, then $p_{.015,3.015}(2.88) = .045$.

5 What Does the $P-$value Measure?

The $P-$value was defined earlier (loosely) as a probability that, in an independent replication of the experiment, the observation would be at least as extreme as the one we observed (given that the hypothesis is true.) Equivalently, it is the greatest lower bound on the set of values α such that the hypothesis could be rejected at level α. From a purely mathematical point of view, this is all we can say about the $P-$value, and one should never try to interpret $P-$values as anything more.

For those who demand (or at least expect) more from the $P-$value, the result of this paper should be cause for reflection. Is there 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 this question 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 at least not be completely incoherent, although it would, in the normal case at least, give the same support, namely 1, to all hypotheses $H : \mathrm{M} \in [a, b]$ 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 $p_H(x)$ for different values of x as the different degrees to which different data values support a single hypothesis H. This is fine so long as we do not acknowledge the possibility of other hypotheses. As soon as we acknowledge that there might be other hypotheses, we are open to the same incoherence problems which we discovered earlier. Those who are willing to hide their heads in the sand and refuse to recognize that other hypotheses are possible now have a solution to the incoherence problem.

Another interesting approach is suggested by the observation that all of the cases of incoherence cited in this paper involve situations in which the larger hypothesis contains parameter values which are farther away (in a sense to be made precise momentarily) from the data than the smaller hypothesis. For example, in the normal case we had $H : \mathrm{M} \in [-.5, .5]$ 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 which 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 μ' not be part of H. Suppose that for every μ in H, $p_{\mu',\mu'}(x) < p_{\mu,\mu}(x)$. Then we will say that μ' is *farther away from x than H is.*

At least for the case of normal distributions, we have that, 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 of our examples. In the case of data with $U(0, \theta)$ distribution with $x < a < b < b'$, let $H : \Theta \in [a, b]$ and $H' : \Theta \in [a, b']$. 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 be remember to interpret it as such. That is, low measure of support would not mean that there is evidence that the hypothesis is false, but rather that the hypothesis contained at least some parameter values which are not supported by the data. It might still contain some parameter values which are highly supported by the data. Nevertheless, if one feels compelled to penalize hypotheses which are too big, the $P-$value still does not do the job. Until there is a consistent interpretation of the $P-$value as something other than a cutoff in the space of possible levels for a test, it would be safest to interpret it as nothing more than that.

6 Discussion

I have tried to argue that the apparent incoherence of one-sided and point-null $P-$values (when interpreted as measures of support for their respective hypotheses) actually extends to interval hypotheses. Also, I have argued that the two extreme types of hypotheses are not two different objects which should never be compared, but rather they are just different versions of the same object of which interval hypotheses are versions as well. Although one-sided $P-$values usually behave well as the hypothesis changes, we found that it was difficult, if not impossible, to interpret one-sided, point-null, and interval $P-$values simultaneously. One could try to argue, with normal distributions at least, that the $P-$value penalizes hypotheses which contain additional parameter values that are far away from the data, but even this argument fails in the case of uniform distributions.

In a sense I have come to agree with both sides in the discussion between Casella and Berger (1987) and Berger and Sellke (1987). Casella and Berger argued that one-sided $P-$values usually can be justified as measures of support for their hypotheses, while Berger and Sellke argued that point-null $P-$values are difficult to justify as measures of support. I do not mean to suggest that the $P-$value is the definitive measure of support for one-sided hypotheses, however. For example, in the binomial situation (as with other discrete data) there are definite problems which we saw. I would merely assert that one runs into fewer inconsistencies trying to interpret one-sided $P-$values as measures of support, than would be required in order to interpret all $P-$values.

All of the examples used to support the above claims came from one-parameter families in which UMPU or UMP tests exist for all of the hypotheses under consideration. If one tried to extend the results to hypotheses concerning the mean of a normal distribution with unknown mean and variance, one would quickly discover that the theory of UMPU tests does not cover hypotheses of the form $H : \mathrm{M} \in [\mu_1, \mu_2]$ vs. $A : \mathrm{M} \notin [\mu_1, \mu_2]$. There are reasonable tests of such hypotheses, but they tend to be biased. One could change the problem to testing $H' : (\mathrm{M} - \mu_0)/\Sigma \in [a, b]$ or $H'' : (\mathrm{M} - \mu_0)/\Sigma^2 \in [a, b]$. A test of H' could be based on the non-central t distribution of the usual $t-$statistic for testing $H : \mathrm{M} = \mu_0$. The case of H'' turns out to be handled in Lehmann (1986, Section 4.4), and a UMPU test exists. The construction of the $P-$value requires numerical integration, however.

The use of $P-$values for point-null hypotheses is very common in applied statistics. For example, virtually every statistical package which does regression and/or analysis of variance prints $P-$values for hypotheses that regression coefficients equal 0 or for the $F-$test of equal cell means. These are cases of multiparameter families which were not considered in this paper. However, given that $P-$values are so difficult to interpret in the simple problems of this paper, one would suspect that their interpretation as measures of support in more complicated cases would be suspect as well.

References

Berger, J. O., and Sellke, T. (1987). Testing a Point Null Hypothesis: The Irreconcilability of $P-$values and Evidence (with discussion). *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.

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

Casella, G., and Berger, R. (1987). Reconcilling Bayesian and Frequentist Evidence in the One-sided Testing Problem (with discussion). *Journal of the American Statistical Association*, 82, 106–111.

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

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

Lehmann, E. L. (1986). *Testing Statistical Hypotheses* (2^{nd} ed.). New York: Wiley.