BAYESIAN STATISTICS 4, pp. 247–266
J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, (Eds.)
© Oxford University Press, 1992

Who Knows what Alternative Lurks in the Hearts of Significance Tests?

JIM HODGES
Rand Corporation, USA

SUMMARY

Hodges (1990) examined Bayesian and classical significance tests and found them to be more similar than is commonly believed. For a large class of significance tests, that paper constructed an alternative hypothesis and prior probability for the null hypothesis such that the posterior probability of the null is the P-value raised to the power $1-\delta$, for $0 < \delta < 1$; for small δ, the posterior probability of the null is uniformly arbitrarily close to the P-value. Many Bayesian writers have asserted that significance tests implicitly use an alternative, but have compared Bayesian and classical tests without knowing what that alternative is. We now know, and by examining it, this paper throws light on significance testing and on the likelihood principle. Two things are particularly relevant: first, many "pathologies" of P-values are shown to be properties of Bayesian hypothesis tests with proper alternative distributions; and second, the alternative that produces the P-value cannot be written as a parametric alternative but instead is necessarily a distribution specified directly on the observable used in the hypothesis test.

Keywords: SIGNIFICANCE TESTS; HYPOTHESIS TESTS; P-VALUES; COHERENCE; LIKELIHOOD PRINCIPLE.

1. INTRODUCTION

Bayesians have a long tradition of reviling significance tests as inherently unBayesian, going back at least to the first edition of Jeffreys' *Theory of Probability*. But Hodges (1990) showed that to an arbitrarily close approximation, the P-value is the posterior probability of the null hypothesis for a particular alternative, in a large class of cases to be detailed below. One could say that the alternative behind significance tests has been unmasked and can now be used as a club to beat on P-values. On the other hand, it now appears that P-values are not as unBayesian as we have been led to think. In view of the tradition of P-value bashing, it is of some interest to examine the new alternative hypothesis to see whether we must admit P-values to the Bayesian family after all.

It is also of interest to understand the implications this alternative has for the earlier Bayesian literature on the subject. A large portion of that literature could be summarized (with moderate injustice) as follows:

(1) We know significance testers have an alternative lurking somewhere.
(2) They don't tell us what it is, so we'll specify a convenient one for our purpose.
(3) The P-value differs from the posterior probability of the null that we compute from the convenient alternative.
  (3a) Therefore, P-values are defective.

As is now clear, step (2) is cheating, because the P-value does not test against step (2)'s strawman alternative. This violates the principle that you should beat up someone for what he claims to do and fails to deliver, not for what he doesn't claim to do. Also, step (3a) no longer works because it presumes that the Bayesian one is better; but it now appears that both solutions are equally Bayesian. Given the effort devoted to comparisons such as this, it is of interest to understand how the P-value alternative and the strawman differ, that the resulting posterior probabilities could be so different.

To answer these questions, Section 2 restates the construction in Hodges (1990) that yields the P-value as the posterior probability of the null, and explains why it works. Section 3 examines the P-value alternative by comparing it to the strawman alternative for a well-used special case, generalizing when possible. Section 4 admits that the new alternative has an odd property that might provide a new reason to hate significance tests, but argues that such a rejection would require the promulgation of a new principle internal to the Bayesian approach.

The bottom line is that you may continue to hate P-values, but not because they're unBayesian; and if you have been abstaining from exploratory uses of significance tests because you thought they are unBayesian, you need abstain no longer.

2. WHAT IS THE P-VALUE ALTERNATIVE AND WHY DOES IT WORK?

Let $X_1, \ldots, X_n$ be n observables, collectively denoted as x. Suppose you have postulated a null hypothesis to explain them, and that you intend to make a significance test of that null hypothesis with the test statistic $T \equiv T(x)$. For cases in which T has an exact continuous distribution under the null, the P-value $P(T)$ — which can be viewed simply as a function of the observable x — has a uniform distribution under the null, conditional on any unknown parameters. Thus, for these cases, $P(T)$ has a uniform distribution with respect to $F(x)$, the marginal distribution function (with respect to the prior on the unknown parameters) of the observation x under the null. For the remainder of this paper, I will restrict consideration to models and test statistics with this property, although the result can be generalized broadly.

For significance tests in this class, construct the alternative density

$$p_{a,n}(x) = Kf(x)\{P(t)^{\delta-1} - 1\}$$

where $f(x)$ is the density of $F(x)$, $0 < \delta < 1$, and $t = T(x)$. Then $K^{-1} = E(P(T)^{\delta-1}) - 1$, where the expectation is with respect to $F(x)$, i.e., $P(T)$ has a uniform distribution. Set the prior probability of the null hypothesis, π, so that it satisfies $(1-\pi)K/\pi = 1$; this gives $\pi = K/(1+K)$. Because we are only considering cases for which $P(T)$ has a uniform distribution, $K^{-1} = (1-\delta)/\delta$ and $\pi = \delta$. For this alternative distribution and this prior probability on the null hypothesis, the posterior probability of the null is $P(t)^{1-\delta}$, which can be made uniformly arbitrarily close to the P-value $P(t)$ by making δ small. Small δ are of particular interest because of this approximation, but we will see that for any δ, the posterior probability of the null under $p_{a,n}(x)$ displays many of the "pathologies" associated with P-values, so that these pathologies can arise from proper alternative distributions. Note that this alternative can be constructed without cheating: if you supply the null model, a test statistic, and the prior distribution of parameters not specified in the null hypothesis, then the alternative can be constructed. The P-value alternative works for apparently any prior distribution for the parameters not specified in the null hypothesis, so that a P-value is consistent with any such prior.

Three aspects of the construction make it work. First is that the P-value has a uniform distribution under the null hypothesis, so it is possible to find the proportionality constant for $\delta > 0$ and $p_{a,n}(x)$ is a *bona fide* probability density in x for $\delta > 0$.

The second important feature of the construction is that $p_{a,n}(\mathbf{x})$ is a marginal distribution for x, not an alternative specified in terms of the unknown parameters, which are integrated out to yield the marginal distribution for x used in a Bayesian hypothesis test. Section 3 shows that in considerable generality $p_{a,n}(x)$ *cannot* be such a parametric alternative. This is probably why earlier Bayesian writers did not find $p_{a,n}(x)$: to my knowledge, those writers worked exclusively with parametric alternatives.

The third important feature is that π, the prior probability of the null, is δ. Actually, π need not equal δ, but π/δ must be finite and bounded away from zero as δ approaches zero, or else the posterior probability of the null will go to 0 or 1. While this value of π may be distasteful to some, there is nothing illegal about it; you may prefer a different π, but $\pi = \delta$ is legitimate.

3. WHAT DOES THE P-VALUE ALTERNATIVE LOOK LIKE?

The exposition will use a familiar special case and generalize when possible. The special case is that $X_1, \ldots, X_n$ are independent and identically distributed normal random variables, with mean μ and variance 1. The null case is $\mu = 0$, the test statistic is $t = -|\bar{x}|\sqrt{n}$ (this expression simplifies later notation). Berger and Delampady (1987, pp. 317-318) (henceforth BD), to take an earlier instance of this special case, use the strawman alternative hypothesis $\mu \sim N(0, \tau)$, and produce a table in which the P-value — the (limiting) posterior probability of the null under the alternative $p_{a,n}(x)$ — differs from the posterior probability implied by their alternative. Let us compare these two alternatives.

For this setup, the (two-tailed) P-value alternative is

$$p_{a,n}(x) = \frac{\delta}{1-\delta}(2\pi)^{-n/2}\exp\left(-\sum_i x_i^2/2\right)\{2^{\delta-1}\Phi(t)^{\delta-1} - 1\}$$

for Φ the standard normal cumulative distribution function. The BD alternative yields an n-dimensional normal distribution for x with mean zero and covariance

$$I_n + \tau\mathbf{1}\ \mathbf{1}^T$$

for $\mathbf{I}_n$ the n-dimensional identity matrix and $\mathbf{1}$ the n-vector of 1's. Both densities are invariant under orthogonal transformations of x that have $\mathbf{1}$ as an eigenvector, and neither is invariant under other transformations. Thus, it is convenient to represent x in $\mathbf{R}^n$ as a combination of a vector parallel to $\mathbf{1}$ and a vector parallel to $\sum x_i = 0$. I will use the representation $x = \bar{x}\mathbf{1} + wv$, where w is a scalar and $\mathbf{v} = -\sqrt{(n-1)/n}\ \mathbf{1} + \sqrt{n/(n-1)}\ \gamma$, γ being the n-vector $(0, 1, 1, \ldots, 1)^T$. Note that $\mathbf{v}$ and $\mathbf{1}$ are orthogonal and $\mathbf{v}$ has length 1. Thus w is the length of the component of x within the hyperplane $\sum x_i = 0$, and $\bar{x}$ is a convenient index of distance from $\sum x_i = 0$ parallel to $\mathbf{1}$. With x represented this way, the null, the P-value alternative, and the BD alternative are, respectively,

$$(2\pi)^{-n/2}\exp\left\{-\frac{1}{2}(w^2 + n\bar{x}^2)\right\}$$

$$\frac{\delta}{1-\delta}(2\pi)^{-n/2}\exp\left\{-\frac{1}{2}(w^2 + n\bar{x}^2)\right\}\left\{2^{\delta-1}\Phi(-\sqrt{n}|\bar{x}|)^{\delta-1} - 1\right\},$$

and

$$(2\pi)^{-n/2}(1-n\tau)^{-1/2}\exp\left\{-\frac{1}{2}(w^2 + \frac{n}{1+n\tau}\bar{x}^2)\right\}.$$

(The symmetry exploited here is peculiar to the special case. In general, there is no symmetry, a counterexample being the test for the slope in simple linear regression, for which there is symmetry only if the regressors have an exploitable pattern.)

To examine the P-value alternative, note that it has three factors: the proportionality constant $\delta/(1-\delta)$, the null density, and a third factor that is roughly the reciprocal of the P-value. Thus, $p_{a,n}(\boldsymbol{x})$ modifies the null by pushing probability away from $\boldsymbol{x}$ in $\Sigma x_i = 0$ and toward values of x where the P-value is small. In particular, on the hyperplane $\bar{x} = 0$, the P-value is 1, so that $p_{a,n}(\boldsymbol{x}) = 0$ on that hyperplane. This feature of $p_{a,n}(\boldsymbol{x})$ figures prominently below. For fixed $\bar{x}$, the third factor of $p_{a,n}(\boldsymbol{x})$ is constant, so the P-value alternative is proportional to the null and the BD alternative for $\boldsymbol{x}$ varying with $\bar{x}$ fixed. Along lines parallel to $\mathbf{1}$, both the second and third factors of $p_{a,n}(\boldsymbol{x})$ vary, and as $\bar{x}$ increases from 0, $p_{a,n}(\boldsymbol{x})$ rises to a mode and then falls off to zero. This can be seen in Figure 1, which shows $p_{a,n}(\boldsymbol{x})$ along the line $w = 0$, for three values of δ. By contrast, the BD alternative has a single mode at $\bar{x} = 0$ and declines monotonically to the right and left.

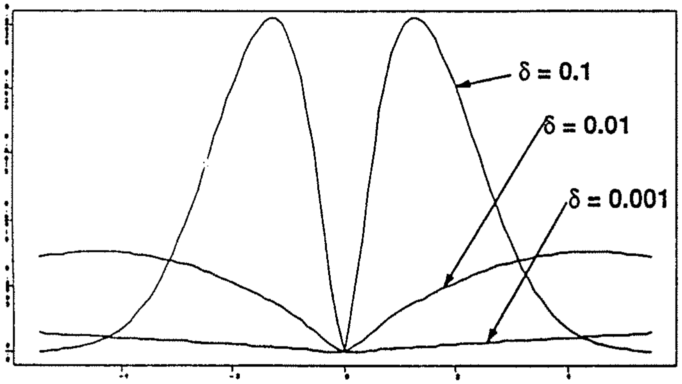

Figure 1. *The P-value alternative $p_{a,n}(\boldsymbol{x})$, viewed along the line $w = 0$, for the special case of the null model being iid $N(0,1)$, for $\delta = 0.1$, 0.01, and 0.001 and $n = 5$.*

(In general, $p_{a,n}(\boldsymbol{x}) = 0$ for values of $\boldsymbol{x}$ in the interior of $\mathbf{R}^n$ that give a P-value of 1. Modes are not necessarily as easily interpreted as in the special case. For example, in the simple linear regression case with a vector of regressors $\mathbf{z}$, $p_{a,n}(\boldsymbol{x})$ is infinite on the subspace $a\mathbf{1}$ and zero on the rest of the subspace $a\mathbf{1} + b\boldsymbol{z}$.)

Figure 1 suggests that as δ goes to zero for fixed n, the P-value alternative becomes an improper uniform distribution. This does not actually happen, for $p_{a,n}(\boldsymbol{x}) = 0$ on $\Sigma x_i = 0$ for all δ and n. However, as Figure 1 suggests, the mode in $\bar{x}$ does increase as δ approaches zero. To see this, consider non-negative $\bar{x}$, compute the derivative of $\log(p_{a,n}(\boldsymbol{x}))$ with respect to $\bar{x}$, and set it equal to zero to obtain this equation in $\bar{x}$ for the mode:

$$\bar{x} = \frac{1-\delta}{2^{1-\delta}\sqrt{n}} \quad \frac{\phi(-\sqrt{n}\bar{x})}{2^{\delta-1}\Phi(-\sqrt{n}\bar{x}) - \Phi^{2-\delta}(-\sqrt{n}\bar{x})}. \tag{1}$$

The right hand side is infinite for $\bar{x} = 0$, so the equation has a solution as long as the right hand side is less than $\bar{x}$ for some value of $\bar{x}$. Taking the limit of the right hand side as $\bar{x}$ approaches infinity and applying L'Hôpital's rule shows this to be the case for large $\bar{x}$ for $\delta > 0$, so that for $\delta > 0$, $p_{a,n}(\boldsymbol{x})$ has a finite non-zero mode in $\bar{x}$. However, the situation changes when $\delta = 0:$ $p_{a,n}(\boldsymbol{x})$ becomes monotonically increasing in $\bar{x}$. This follows simply from noting that $-\exp(-n\bar{x}^2)$ is monotonically increasing, and that

$$\lim_{x\to\infty} \exp(-x^2/2)(\Phi(-x))^{-1}$$

is infinite.

Because the P-value alternative is zero on $\Sigma x_i = 0$ and particularly at the origin, it is easy to show that $p_{a,n}(\boldsymbol{x})$ cannot be derived by specifying a distribution for μ, i.e., as a parametric alternative. If it could, there would be a density $h(\mu)$ such that

$$\int h(\mu)\exp\{-\Sigma(x_i-\mu)^2\}d\mu$$

would be zero at $\boldsymbol{X} = 0$, which is impossible. This confirms what defenders of significance tests, such as Barnard, have been saying for decades: this significance test does not test against the parametric alternative $\mu \neq 0$.

This last property — that the P-value alternative cannot be constructed as a parametric alternative — appears to generalize widely. If $f(\boldsymbol{x}|\theta)$ is any parametric model for the vector $\boldsymbol{x}$ (within the class to which this paper is restricted), T is a test statistic with an exact continuous distribution for some null on θ, some $\boldsymbol{x}^*$ in the interior of $\mathbf{R}^n$ gives a P-value of 1, and $f(\boldsymbol{x}^*|\theta) \neq 0$ off a set of measure zero for any θ, then the only parametric alternatives that can yield $p_{a,n}(\boldsymbol{x})$ have their mass on the sets of measure zero on which $f(\boldsymbol{x}^*|\theta) = 0$; and if there are no such sets, $p_{a,n}(\boldsymbol{x})$ cannot be constructed as a parametric alternative. This is true, for example, of all the familiar two-tailed exact tests associated with regression and ANOVA.

The remaining feature to be considered is the behavior of $p_{a,n}(\boldsymbol{x})$ as n increases. Consider equation (1) above, and let n grow large with δ fixed. By an application of L'Hôpital's rule, the right hand side becomes approximately $(1-\delta)(\bar{x}+n^{-1})$, so that the mode is approximately $\bar{x} = (1-\delta)/\delta n$, which goes to zero as n approaches infinity. Thus, as n grows large for fixed δ, the profile of $p_{a,n}(\boldsymbol{x})$ for any value of w comes to resemble a normal density with a thin pie slice taken out of the density right at its mode. For δ close to zero, the mode is approximately $1/\delta n$, which implies that as n gets large and δ gets small, $p_{a,n}(\boldsymbol{x})$ becomes rather flat.

To explore the limiting behavior of $p_{a,n}(\boldsymbol{x})$ further, consider Lindley's paradox (see, e.g., Berger and Delampady 1987). For our example, Lindley's paradox says: if $\bar{x} = k/\sqrt{n}$ (as n increases), then the P-value stays fixed at α corresponding to k, but the BD posterior probability of the null goes to 1. This has been considered a defect of P-values, but if a Bayesian test is constructed with the alternative $p_{a,n}(\boldsymbol{x})$ with $\delta > 0$, the posterior probability of the null stays fixed at $\alpha^{1-\delta}$ as n increases. Thus, this "paradox" depicts a difference between legitimate alternatives, not a difference between a Bayesian approach and another approach. The "paradox" can be worked in the other direction: if

$$\bar{x}^2 = \xi^2(n) = \{-2\log(B) + \log(1+n\tau)\}(1+n\tau)/n^2\tau,$$

so that $\bar{x}$ goes to zero at the rate $(\log(n)/n)^{1/2}$, then the BD Bayes factor stays fixed at B^{-1}, the BD posterior probability of the null stays fixed at $(1+(1-\pi)/B\pi)^{-1}$, and the P-value goes to zero. These two paths of $\bar{x}$ to the origin provide a means by which to characterize the different shapes of the BD and P-value alternatives as n grows. For non-negative $\bar{x}$, the ratio of $p_{a,n}(\boldsymbol{x})$ to the null and of the BD alternative to the null are, respectively:

$$\frac{\delta}{1-\delta}\left\{2^{\delta-1}\Phi^{\delta-1}(-\sqrt{n}\bar{x}) - 1\right\} \qquad \text{and} \qquad (1+\tau n)^{-1/2}\exp\left\{\frac{1}{2}\;\frac{n^2\tau}{1+n\tau}\bar{x}^2\right\}$$

Reinterpreting Lindley's paradox in these terms, the comparison of the two alternatives to the null can be summarized as follows:

- if $\bar{x} \to 0$ faster than $k/\sqrt{n}$, both alternatives become arbitrarily smaller than the null as n increases, but $p_{a,n}(x)$ goes to zero at the origin while the BD alternative reaches its mode,
- if $\bar{x} \to 0$ as $k/\sqrt{n}$ does, $p_{a,n}(x)$ maintains a ratio of $K(\alpha^{\delta-1}-1)$ compared to the null, while the BD alternative becomes arbitrarily small with respect to the null,
- if $\bar{x} \to 0$ at the rate of, say, $(\log\log(n)/n)^{1/2}$, $p_{a,n}(x)$ gets arbitrarily large with respect to the null and the BD alternative arbitrarily small,
- if $\bar{x} \to 0$ as $\xi(n)$ does, $p_{a,n}(x)$ becomes arbitrarily large relative to the null while the BD alternative maintains the constant ratio of B^{-1},
- if $\bar{x} \to 0$ slower than $(\log(n)/n)^{1/2}$, both alternatives become arbitrarily larger than the null, but for every δ and τ there is an N such that for $n > N$, the BD alternative becomes arbitrarily larger than $p_{a,n}(x)$ (the proof being trivial).

Figure 2 depicts the behavior of the natural logs of the two ratios as functions of $\bar{x}$, for $n = 5$. The vertical lines marked "k/sqrt(n)" and "~sqrt(log(n)/n)" correspond to the two rates of convergence of $\bar{x}$ to zero, described above. At $\bar{x} = k/\sqrt{n}$, the top line maintains a constant height as n grows large, but the lower line becomes arbitrarily negative. At $\bar{x} = \xi(n)$ (roughly $(\log(n)/n)^{1/2}$), the lower line maintains a constant height as n increases, but the upper line becomes arbitrarily higher. To the left of $\bar{x} = k/\sqrt{n}$, the two curves become arbitrarily negative. Between $\bar{x} = k/\sqrt{n}$ and $\bar{x} = \xi(n)$, the top line becomes arbitrarily large and the bottom line arbitrarily small. To the right of $\bar{x} = \xi(n)$, both lines become arbitrarily large, and they eventually cross.

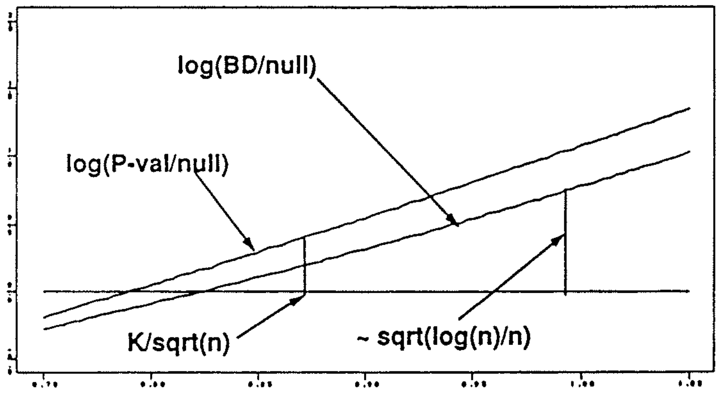

Figure 2. *Logs of the ratios of $p_{a,n}(x)$ to the null and the BD alternative to the null, as a function of $\bar{x}$ for any fixed w, in the special case of the null model being iid $N(0,1)$, for $\delta = 0.1$, $n = 5$.*

What does all this say about the alternative that produces the P-value? Generally, that it's just another alternative: apparent difficulties like Lindley's paradox are merely features of one Bayesian's alternative, which all must respect even if they prefer another. Many of the "pathologies" associated with P-values occur for $\delta > 0$, that is, with a proper alternative.

It is important that the P-value does not test against a parametric alternative, for this provides a way to examine $p_{a,n}(x)$'s implications for the likelihood principle (Berger and Wolpert, 1984). In one sense, $p_{a,n}(x)$ is trivially consistent with the likelihood principle, if the likelihood is the function $f(x|\Psi)$, for Ψ taking the two values H_o and H_a. But this does not address the ways in which the principle has been used to argue against significance tests. Such arguments take the form, "you may not use the P-value to test $\mu = 0$ against $\mu \neq 0$ because the P-value uses aspects of the data not captured in the likelihood for μ." The development in this paper suggests a different argument: "you may not use the P-value to test $\mu = 0$ against $\mu \neq 0$ because the P-value does not test against any alternative specified in terms of μ." The likelihood principle adds nothing to this argument and would appear to be superfluous for this issue.

What *does* the P-value test against? It looks particularly for middling deviations from the values of $\bar{x}$ predicted by the model, and is less interested in the extremes of $\bar{x}$'s distribution than is the more familiar BD alternative. It must be admitted that when rendered as an alternative distribution, the P-value is unfamiliar and difficult to work with. With the backgrounds we have, it is unlikely that someone would, from scratch, directly specify $p_{a,n}(x)$, particularly for more complicated cases. But that does not imply that it should not be used.

4. A STRAW TO GRASP: AN ODD PROPERTY OF THE P-VALUE ALTERNATIVE

Suppose, to continue the special case, that we have $n+1$ observations. Construct $p_{a,n+1}(x)$ and the BD alternative and then, for each one, integrate out the last observation. The result for the BD alternative is the alternative that would have been obtained had we begun with n observations, but the result for $p_{a,n+1}(x)$ is not: in particular, the value at the n-dimensional origin is positive, while $p_{a,n}(0) = 0$. Unfortunately, the integrals involved are intractable and I have not been able to obtain more detailed results.

This result generalizes to cases of $n+1$ independent and identically distributed X_i. Suppose the X_i have common density $f(x_i|\theta)$ conditional on θ, and construct the P-value alternative $p_{a,n+1}(x)$. If there is some n-dimensional x^* in the interior of $\mathbf{R}^n$ such that the P-value is 1 for x^*, then $p_{a,n}(x^*) = 0$ and by a straightforward argument,

$$p_{a,n}(x^*) \neq \int p_{a,n+1}(x^*, x_{n+1})dx_{n+1}$$

unless the product of the individual densities at x^*, $\Pi f(x_i|\theta)$, is zero for all θ. In the following discussion, I will refer to this property by saying that the BD alternative satisfies the integration property but that $p_{a,n}(x)$ does not.

Shall we grasp at this straw to maintain the age-old rejection of significance tests? Is it undesirable that $p_{a,n}(x)$ does not satisfy the integration property? There is apparently nothing unBayesian about it: the P-value alternative for $\delta > 0$ is a *bona fide* distribution, so we cannot object without inventing a new principle which would operate internal to the Bayesian approach. Still, the property does seem odd, for it seems to say that what we learn from the first n observations depends on how many other observations have been taken, even if we don't know their values. A sufficient condition for an alternative to satisfy the integration property is that it arises from a parametric model plus a probability distribution on the parameter values. I do not know if this condition is also necessary. If it is, then if we are to establish the integration property as a new principle, we are enshrining as principle the use of parametric alternatives. On the other hand, if some larger class of alternatives has the integration property, then *that* would be interesting, and defensible as a possible basis for a principle.

5. CONCLUSION

The motivation for Hodges (1990) was that Bayesians, like others, need tools for doing exploratory data analysis, but that Bayesians, unlike others, are mostly uncomfortable with standard EDA tools because they rely on the logic of significance tests (Box 1980). More candid Bayesians like Smith (1986) argue that the Bayesian approach cannot be used until a formal framework has been set up for a problem, but reject significance tests even in the informal work that precedes establishment of the formal framework. The point of the P-value alternative in Hodges (1990) was that Bayesians need not be so fastidious about making exploratory use of P-values, because they are a Bayesian construction to a close enough approximation that it isn't worth arguing over. I reiterate that point here: Bayesians, use P-values in exploration and feel good about it!

On a less polemical note, Bayesians have been saying for about a half-century that an alternative must lurk behind a significance test, so it is of some interest that the P-value alternative can be constructed at all. It is satisfying that $p_{a,n}(x)$ can not, in general, be expressed as an alternative in terms of the parameter specified in the null, just as defenders of significance tests have said all along. Instead, $p_{a,n}(x)$ tests the null model more generally, as Fisher's disjunction (Fisher 1973, p. 42) implies, although even after the examination in this paper, it is not obvious just what features of the model are being tested.

ACKNOWLEDGMENTS

In preparing this paper I had the helpful comments of John Adams, Bob Bell, and Mike Mattock of RAND, and of Paul von Batenburg of Touche Ross Nederlands. Sharon Koga prepared this paper in TEX.

REFERENCES

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

Berger, J. O. and Wolpert, R. L. (1984). *The Likelihood Principle.* Hayward, CA: IMS.

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

Fisher, R. A. (1973). *Statistical Methods and Scientific Inference.* New York: Hafner.

Hodges, J. S. (1990). Can/May Bayesians use pure tests of significance?. *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard* (S. Geisser, J. S. Hodges, S. J. Press and A. Zellner, eds.), Amsterdam, North Holland, 75–90.

Smith, A. F. M. (1986). Some Bayesian thoughts on modelling and model choice. *The Statistician* **35**, 97–102.

DISCUSSION

D. J. POIRIER (*University of Toronto, Canada*)

Jim Hodges has tackled the daunting task of putting significance testing on a firm Bayesian foundation. In discussing this undertaking I am reminded of comments at the first Valencia meeting by Dawid (1980, p. 311):

> "I have learned to be wary of those who claim that they would like to reconcile the various opposing views on statistical inference. In my experience, the invariable consequence is, rather, a polarisation of attitudes and a great deal of fruitless apoplexy".

At the same meeting Kadane (1980, p. 317) pulled no punches regarding the desirability of Hodges' goal "... as a general matter, I believe that significance testing threatens the respectability of statistics more than any other single factor". Also, our Conference President has offered his assessment of significance tests: "They are widely used, yet are logicall indefensible". (Lindley (1986, p 502)).

In an attempt to understand why Hodges tackles a topic in the face of skepticism of such Bayesian titans, I will discuss three essential ingredients of pure significance tests: (1) the use of sampling theory, (2) the use of unspecified alternatives, and (3) ambiguity over wha is to be done if the null is rejected. Contrary to Hodges' claim, I will argue that his analysi is "unBayesian".

The issue regarding (1) was eloquently expressed by Jeffreys (1961, p. 385) in thi assessment of the underlying logic of significance tests: "... a hypothesis that may be tru may be rejected because it has not predicted observable results that have not occurred.' The issue revolves around a litmus test of a statistician's pedigree, namely, the Likelihoo Principle. I do not fully understand Hodges' remarks regarding the Likelihood Principle There is an extensive literature on the Likelihood Principle versus significance test (e.g. Berger and Wolpert (1988, pp. 104–110)). In particular, Hodges fails to explain his implici acceptance of a host of embarrassments (e.g., susceptibility to noninformative stopping rules that go along with use of P-values. Honest Bayesians acknowledge their Achilles's heel (the prior) and address it through prior sensitivity analysis. It is time proponents of frequentis reasoning do the same, address their vulnerability to ambiguity over the sampling distribution and acknowledge their willingness to consider data that could have been observed but wer not.

As for (2), I have no trouble with the eloquent observation of Box (1980, p. 387) that: "... it seems a matter of ordinary human experience that an appreciation that a situation is unusual does not necessarily depend on the immediate availability of an alternative" Significance tests are intended to aid in that appreciation, i.e., to help assess, in the words of Barnard (1972, p. 129), "whether agreement between this hypothesis and the data is sc bad that we must begin to exercise our imaginations to try to think of an alternative tha fits the facts better". Indeed, Cromwell's Rule warns against the arrogance of dogmaticall assuming all is well and cannot be made better, and I have adopted it as one of my Pragmatic Principles of Model Building (Poirier, (1988, p. 140)).

Such acknowledgements, however, are different than advocating statistical testing of nul hypotheses without explicit alternatives. As Cornfield (1970, p. 28) noted long ago:

> "... the development of new hypotheses is ... no different from mathematics itself, which is concerned with methods of proving theorems, but has no advice on how to formulate new ones. Perhaps not the least of the advantages of the Bayesian outlook therefore is that it provides a clear-cut distinction between creative activity such as hypothesis formulation, which can be performed only by trained and imaginative people, and formal analysis which is in principle capable of reduction to routine performance by robots".

Hodges's attempt to provide a Bayesian foundation for testing without well-specified alternatives blurs Cornfield's distinction and obfuscates the difference between Bayesian techniques and the Bayesian outlook. The manner in which the alternative data density varies with x depends crucially on the choice of discrepancy function (tantamount to the choice of an alternative hypothesis) used to define the P-value, a matter about which the significance testing literature has little to say. Unless Hodges can argue that this density has a consistent and sensible interpretation across problems, however, these "foundations" seem to undermine, as much as support, significance tests.

In econometrics, where I usually hang my hat, pure significance tests have become popular in the rush by researchers to "out-test" their competitors. An example is the information test of White (1982), and the historical response to this test was typical: researchers sought alternative hypotheses for which the test had non-trivial power (e.g., Hall (1987)) in an attempt to understand how to use the test. Significance tests do not require alternatives for their derivation, but their conscientious use requires alternative hypotheses.

As for (3), proponents of significance tests often disarm their opponents by claiming such tests are only intended as "quick and dirty" methods for signalling the need to think more and reconsider the current window on the observable world. Who can argue with the advice to think more? Another sidestep is to argue that significance tests are intended for "assessing deficiencies" rather than testing. Such vagueness makes it impossible to define a "procedure" whose long-run performance is to be evaluated. This leaves frequentists without advice to give, but also conveniently "off the hook" since pretesting concerns cannot be directly raised. Hodges (1990) emphasizes the exploratory role of significance tests, but what are the rules of this game? Isn't this playing tennis without a net?

From the Bayesian perspective, I believe this vagueness is *less* devastating, although not without its annoyances. Once the researcher's mind has been jogged, by whatever manner, the troublesome issue is how to pick priors for a hypothesis suggested by the data. In memory of our dearly beloved and departed friend Morrie DeGroot, let me recall his characteristic wit and perception into the double-counting conundrum facing the Bayesian:

> "We open the newspaper in the morning and read some data on a topic we had not previously thought about. In order to process the data, we try to think about what our prior distribution would have been before we saw the data so we can calculate our posterior distribution. But we are too late. Who can say what our prior distribution would have been before we saw the data? We lost our virginity when we read the newspaper."
>
> DeGroot (1980, pp. 311–312)

How does one restore virginity lost? The importance of the question depends on the purity of the researcher. Once the researcher gives up the ideal state of the "single-prior Bayesian" and admits the need for sensitivity analysis in public research, one is left with the usual task of presenting a variety of mappings from "interesting" priors to posteriors and leaving it to the reader to decide whether the priors are sufficiently plausible to warrant serious consideration. Those who prefer "virgin priors" are likely "virgin data analysts".

Many Bayesian researchers have noted the possibility of assigning only $1-\varepsilon$ prior probability to the hypotheses H_j $(j = 1, 2, \ldots, J)$ and reserving ε $(0 < \varepsilon < 1)$ probability for the hypothesis H_{J+1}: "something else". Provided the researcher specifies priors for the unknown parameters given H_j $(j = 1, 2, , \ldots, J)$ and the relative prior probabilities of these J hypotheses, standard posterior odds analyses permit comparison of the *relative* posterior probabilities of the hypotheses without specifying ε. If in the process the researcher's creative mind has a new insight leading to specification of "something else", then analysis can proceed straightforwardly given interesting sets of values for ε and priors for the parameters given H_{J+1}. The Bayesian moral is simple: only make *relative* probability statements about specifications explicitly entertained. Be suspicious of anyone promising more!

In summary, the thrust of my comments are echoes of other Bayesians who have come before. Criticism (like discovery) lies outside any current statistical paradigm. Perhaps that is the way it should be since a theory of criticism would amount to a theory of creativity. Exploratory data analysis may be best left as an art. I am willing to be convinced otherwise, but if anything, this paper has strengthened my conviction to leave art outside the protection of the Bayesian umbrella.

J. O. BERGER (*Purdue University, USA*)

I find the notion of specifying an 'alternative' distribution as a marginal distribution of x interesting, though from a practical perspective I agree with the author that the P-value alternative is "unfamiliar and difficult to work with." I presume others will address the sensibility of this alternative, and hence will confine my comments to three fundamental issues.

Issue 1. What does it mean to match P–values and posterior probabilities? The author has constructed an 'alternative' for which the posterior probability of the null is $P(t)^{1-\delta}$, which is approximately equal to the P–value, $P(t)$, for small δ. To be a bit more precise, suppose we agree that 'approximately equal' means $P(t)^{1-\delta} = P(t)(1+\varepsilon)$ for some specified small ε, which then implies that $\delta \cong -\varepsilon/\log P(t)$. Since this was to be set equal to the prior probability, π, of the null hypothesis, we have

$$\pi \cong -\varepsilon/\log P(t). \tag{1}$$

Now consider the usual Bayesian approach of specifying a parametric alternative to H_0 such as $H_1{:}\ \mu \neq \mu_0$, along with a conditional prior densit/, $g(\mu)$, on H_1. Supposing that $f(x|\mu)$ is the density (with $\mu = \mu_0$ specifying the null distribution and π again standing for its prior probability), the Bayes factor is

$$B_g(x) = f(x|\mu_0) \Big/ \int_{\{\mu\neq\mu_0\}} f(x|\mu)g(\mu)d\mu.$$

It is then easy to see that the posterior probability of H_0 and the P–value are equal if π is chosen to be

$$\pi = \left(1 + \frac{1}{B_g(x)}\left(\frac{1}{P(t)} - 1\right)^{-1}\right)^{-1}. \tag{2}$$

Frequently, $B_g(x)$ can be written as a function $\psi(P(t))$, in which case it seems that (1) says exactly the same thing as (2): *by appropriate choice of the prior probability π, one can always force agreement between the posterior probability and the P–value.* And note that (2) is usually a monotonically increasing function of $P(t)$, as is (1). It would be helpful for the author to clarify the distinction between his conclusion and the standard (2).

Issue 2. Do Bayesians denigrate a strawman? I don't think so. The author refers to the "strawman" $g_1(\mu) = \mathcal{N}(0, \tau)$ discussed in Berger and Delampady (1987). Although $g_1(\mu)$ was discussed in BD for certain purposes, the thrust of the article was to draw conclusions that are valid simultaneously for any 'reasonable' prior. The reason that Bayesians do not like P–values is that, *for any reasonable $g(\mu)$, the P–value will be very different from the Bayes factor $B_g(x)$.*

This can be said in many different ways. My current favorite is the "frequentist interpretation" in Example 6 of BD. It is very hard to read that example and afterwards take the familiar interpretation of P–values seriously, at least in situations with a parametric alternative. Of course, a feature of the reasoning here is that allowing freely varying π, when making comparisons between P–values and posterior probabilities, is senseless because of (2), so that comparisons require either fixing π (as in Example 6 of BD) or considering the Bayes factor. This also seems obvious from a pragmatic perspective: if one is trying to have the data say something about a null model, one must make sure to remove the influence of the (subjective) prior probability of the null model.

Issue 3. Are P–values useful in exploratory analysis? I do not think it is necessary to work hard to argue that P–values are useful in exploratory analysis. Virtually by definition, exploratory analysis cannot be done in a formal Bayesian way, meaning that adhoc indicators must be used. I am not convinced that P–values are particularly useful adhoc indicators and they may well do more harm than good in actual use, but 'adhoc' is not automatically bad if Bayes cannot be done.

Note that my argument here is not based on a feeling that parametric alternatives and nonparametric alternatives (typical in exploratory analysis) are fundamentally different; indeed, in BD several arguments are given to the effect that P–values are also misleading for nonparametric alternatives. Rather, the argument is that, for nonparametric alternatives, it may simply be impossible to do a believable Bayesian analysis.

L. M. BERLINER (*Ohio State University, USA*) and
C. ROBERT (*Université Pierre et Marie Curie, France*)

One might think that the evidence accumulated about the questionable value of P-values would outweigh the appeal of a formal similarity with Bayesian answers. However, since the debate is not quite closed, we take the opportunity of discussing this paper "to pound another nail in the coffin," to borrow an expression from Berger and Sellke (1987). First, note that the phenomenon exhibited by Hodges, namely the fact that $p(x)^{1-\delta}$ is the answer associated with the "marginal" distribution

$$m(x) = (\delta/(1-\delta))f(x)[p(x)^{\delta-1} - 1], \qquad (1)$$

is in fact, a criticism in itself, since the "Bayesian" answer is $p(x)^{1-\delta}$, instead of $p(x)$. Hence, the P-value is not obtained as a Bayesian answer by Hodges, but only an approximation for small δ. To assess the approximation, consider the following values of $p(x)^{1-\delta}$ for various "interesting" values of $p(x)$ and various δ.

P-value	.01	.05	.10
δ			
.005	0.0102	0.0508	0.1012
.025	0.0112	0.0539	0.1059
.050	0.0126	0.0581	0.1122
.100	0.0158	0.0675	0.1259
.200	0.0251	0.0910	0.1585
.300	0.0398	0.1228	0.1995
.400	0.0631	0.1657	0.2512
.500	0.1000	0.2236	0.3162

Values of $p(x)^{1-\delta}$.

Note that for prior probabilities of the null as small as $\delta = .2$, the P-value is roughly 1/2 Hodges' answer of $p(x)^{1-\delta}$. On the other hand, for δ smaller than the P-value, yet in the range where Hodges' approximation is good, one notes that the P-value approximation actually can lend evidence *in favor* of the null. For intermediate values like $\delta = .05$ or .1 and when δ and the P-value are close, one must not only decide if the approximation is good, but also, whether or not one should pretend anything has actually been learned.

Our next point is that a procedure which corresponds "formally" to a Bayesian answer is not necessarily a good procedure. (Of course, Hodges only shows the P-value to be an approximation.) The literature is full of assessments of the behavior of the P-value from various points of view. Pertinent recent results of Hwang *et al.* (1991) are of interest. They demonstrate inadmissibility results concerning the P-value. In the case of continuous exponential families and two-sided hypotheses, one can view the P-value as an estimator of the indicator function at the null hypothesis, $I_{H_0}(\theta)$, and then evaluate an estimator of

$I_{H_0}(\theta)$ under quadratic loss. Hwang *et al.* (1991) show that the P-value is inadmissible in this case, and is thus suboptimal even from a frequentist perspective. Hwang and Pemantle (1990) have generalized this result to a large class of proper loss functions.

Next, we turn to some other operational and Bayesian criticisms of Hodges' suggestion:

(i) Note the following discrepancy in one-sided testing. Consider the normal distribution $X \sim N(\theta, 1)$ and the null hypothesis $H_0 : \theta \leq 0$. The P-value can then be written as the Bayesian posterior probability of the null corresponding to the uniform, improper prior. (See Casella and Berger, 1987, for generalizations.) Such results do not agree with Hodges' suggestion.

(ii) The P-value is not always uniquely defined in complicated settings. Different definitions of the P-value may lead to different forms for (1), leading to potential obvious contradictions to the likelihood principle.

(iii) (1) appears to be of no use at all in the presence of nuisance parameters.

(iv) Consider the independence of (1) with respect to the alternate hypothesis. For instance, hypothesis tests of $H_0 : X \sim N(0,1)$ against either $H_a : X \sim N(\theta, 1), \theta > 0$, or $H_{a'} : X \sim N(\theta, (1+\theta)^{-1}), \theta > 0$, produce the same P-value, $p(x) = P_0(X > x)$, since both alternatives are stochastically larger than H_0, and, hence, the same "marginal" (1). However, it seems to us that the marginal should be more skewed to the right in the second case than in the first. Problems also appear in nonparametric settings even when the P-value is naturally defined. For example, the following case points out a fundamental flaw in Hodges' suggestion. Consider $H_0 : X \sim f_0$ and $H_a : X \sim f_1$, where f_1 is stochastically larger than f_0. The P-value is $p(x) = P_0(X > x)$. Following Hodges literally, we should be testing H_0 against (1): We should replace one simple alternative with another! Such behavior can obviously be extended to composite alternatives in a fashion so that the marginal (1) does not even belong to the original collection of alternatives. This makes Hodges' suggestion questionable even in the context of exploratory work.

In general, "reconciliation" or approximation of P-values with Bayesian posterior probabilities of the null is hardly sufficient justification for suggesting the use of P-values. The analyst must ask whether or not the structure of priors leading, even approximately, to the P-value are sensible. Regarding Hodges' derivation, the fact that the marginal distribution (1) is only a "pseudo-marginal distribution" (there is no nondegenerate prior-likelihood pair leading to (1)), is extremely disturbing to the Bayesian who takes the Bayesian view of statistics seriously. Specifically, what sort of bona fide subjective reasoning could consistently lead to (1) as a meaningful distribution? Indeed, we take the view that Hodges' analysis actually suggests another reas6n *not* to use the P-value: It often approximates a Bayesian analysis corresponding to a ridiculous "prior" specification.

G. CASELLA (*Cornell University, USA*)

This article represents a valiant attempt to reconcile evidence in two-sided point-null hypothesis testing, an attempt for which Dr. Hodges is to be congratulated. Unfortunately, it appears to be impossible to reconcile frequentist and Bayesian evidence in this case. A cause of this is the fundamentally different approach to two-sided testing taken by frequentists and Bayesians. Much of the strange behavior displayed by the author's P-value alternative is directly attributable to the prior placing a point mass on the null hypothesis.

To obtain reconciliation between P-values and posterior probabilities means that a prior can be specified for which the P-value equals the posterior probability for all data values. That is, when testing $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_0^c$ based on observing $X = x$, where

$X \sim f(x|\theta)$ and $\theta \sim \pi(\theta)$, evidence can be reconciled if

$$P(H_0|x) = \frac{\int_{\theta_0} \pi(\theta|x)d\theta}{\int_{\theta_0} \pi(\theta|x)d\theta + \int_{\theta_0^c} \pi(\theta|x)d\theta} = P\text{-value} = P(x) \qquad \forall x,$$

where $\pi(\theta|x)$ is the posterior obtained from Bayes rule.

In a pair of papers, J. Berger and Sellke (1987) and Casella and R. Berger (1987a) examined reconciliation issues in both one-sided and two-sided testing. Berger and Sellke (1987) (and later Berger and Delampady, 1987) argued that reconciliation is impossible in two-sided testing, and saw this as a shortcoming of the P-values. Since, in two-sided testing, posterior probabilities are usually greater than P-values, it was concluded that P-values overstate evidence against H_0. However, Casella and Berger (1987a), in the spirit of the work of DeGroot (1973), showed reconciliation is possible in the one-sided case. They blamed the two-sided discrepancies on the Bayesian treatment of the point-null, and argued that point-mass priors place too much weight on H_0, causing the Bayesian posterior probability to be large (see also Casella and Berger, 1987b).

The different behavior in the one- and two-sided case was partially explained by Hwang *et al.* (1991). They showed that the P-value is generalized Bayes (and admissible) in the one-sided problem, but is not generalized Bayes in the two-sided problem. This latter result, which formalizes some of the author's ideas, shows that true reconciliation is impossible in point-null testing. A different approach to reconciliation was taken by Casella and Wells (1990). Using group structures, they established necessary and sufficient conditions for reconciliation, conditions that essentially eliminate the point-null case.

Since we know that, in point null testing, the P-value cannot be a posterior probability arising from a prior, even an improper prior, how do we interpret Hodges' result? Does the P-value alternative, which specifies a posterior unrelated to any prior, help us in understanding the discrepancies between P-values and posterior probabilities. Unfortunately, the answer seems to be no.

The solution proposed in the paper is a mathematical one: equate the P-value with the posterior probability and solve for the function $p_{a,n}(x)$ that must be the P-value alternative. This mathematical solution, however, is not a statistical solution, and arguing that $p_{a,n}(x)$ is an alternative density (or posterior probability) will anger frequentists as much as (I'm sure) it angers Bayesians. The function $p_{a,n}(x)$ cannot be updated (since it fails the "integration property") and doesn't allow a prior-based interpretation. Although Hodges says that "... the P-value alternative works for apparently any prior distribution ...", what is really the truth is that the P-value alternative works for no prior distribution.

To a frequentist, the P-value alternative is distasteful simply because of its strange behavior. In effect, to accept the P-value alternative means to defend it. Although I am not against using P-values for significance testing, I do not want to defend the P-value alternative.

The strange behavior of the P-value alternative is mainly a consequence of the fact that the P-value equals 1 at H_0. This behavior does not occur in one-sided testing, where the P-value only approaches 1 as the data go deeper in to H_0. That is, for $H_0 : \theta \leq 0$ vs. $H_1 : \theta > 0$, $\lim_{x \to -\infty} P(x) = 1$, but $P(x) < 1$, for $x > -\infty$. Since the P-value only equals 1 in the limit, reconciliation is possible. On the other hand, point-mass priors result in posterior probabilities that place too much weight on H_0. Until these particular features are addressed, true reconciliation can never obtain. I applaud the author's efforts at a mathematical reconciliation, and hope he will turn next to a more statistical reconciliation.

A. P. DAWID (*University College London, UK*)

The behaviour described in Section 4 of this paper is more than a mere curiosity — it displays a fundamental incoherence in the author's method of constructing a P-value alternative. The argument for the Bayesian position, as developed by de Finetti for example, starts from the idea that the decisions we take in different circumstances should, in a certain technical but intuitively forceful sense, *cohere*. When we consider the implications of this for a single experiment, we find that it is necessary and sufficient that the decision maker act as though he or she had a joint probability distribution for all the variables involved. Hodges claims that use of his P-value alternative is Bayesian because it has this property. However, coherence goes further than this, and also applies to the comparison of decisions made in different experiments — for example, with differing sample sizes. The importance of this aspect of coherence has been emphasised by Lindley (1978): some implications have been explored by Brown (1980) and Dawid (1988). As a simple example, consider the quadratic-loss estimation of a probability parameter θ on the basis of r successes in n trials. The minimax estimate is $\frac{r+\sqrt{n}/2}{n+\sqrt{n}}$, which is also the Bayes estimate if we use a $\beta(\sqrt{n}/2, \sqrt{n}/2)$ prior for θ. However, the dependence of this prior on n means that, whilst use of minimax for any one value of n could be defended as Bayesian, willingness to use it for all n is incoherent, and hence open to successful counterbetting. In exactly parallel fashion, the failure of the "integration property" means that willingness to use the P-value alternative for all n is incoherent and unBayesian.

Hodges speculates that the integration property requires putting a prior distribution on the model, but this, while sufficient, is clearly not necessary: the property is nothing other than Kolmogorov's consistency requirement for the existence of a joint distribution for an unlimited number of variables, and as such is satisfied for a collection $\{p_n(\cdot)\}$ if and only if there exist a fixed joint distribution for the infinite sequence of observations whose marginal density for the first n is $p_n(\cdot)$. Were this to be the case for $\{p_{a,n}(\cdot)\}$, the would-be coherent decision maker might have been able to take the P-value alternative more seriously.

E. I. GEORGE (*University of Chicago, USA*)

At this point in my career, I think of myself as both a Bayesian and a Frequentist. That is, I find both points of view enlightening, and neither without shortcomings. Having accepted ambiguity as a fact of life, I am willing to suffer any of the contradictions inherent in maintaining both perspectives. From my vantage point, I applaud this paper because it helps me understand significance testing in the light of Bayesian machinery. For example, I previously found it perplexing that Bayesian critics of P-value were able to find such large discrepancies between posterior probabilities and P-values. The paper resolves this dilemma for me by showing how the P-value alternatives can be quite far from the parametric alternatives which are so often chosen by Bayesians. I also find it enlightening to understand Lindley's paradox as a phenomenon resulting from a particular sequence of legitimate alternatives. It is curious that as opposed to a fence sitter like myself, an exclusive Bayesian and an exclusive Frequentist would each be inclined to reject this paper as irrelevant. The exclusive Bayesian would argue that statistical procedures should depend on the choice of prior and not vice-versa, whereas the exclusive Frequentist would argue that statistical procedures need no Bayesian justification.

I would like to comment on perhaps the most provocative conclusion of the paper, that the P-value alternatives does not satisfy the "integration property", namely that $p_{a,n}$ is not obtained from $p_{a,n+1}$ by integrating out the extra dimension. To me, this aspect of $p_{a,n}$ is not a defect but rather a property arising from a reasonable lack of invariance with respect to the defining construction. More precisely, $p_{a,n}$ is obtained by weighting those values of x which re "more extreme" according to the P-value. It is surprising that the more extreme values in ?^{n}cannot be obtained by integrating or averaging over values in R^{N+1}. In fact, the argument ould be made that alternative distributions which do satisfy the integration property are less xtreme, and so make it more difficult to discriminate against the null. Thus, this feature rovides further insight into the often cited discrepancies between posterior probabilities and -values.

. MORTERA (*University of Rome, Italy*)

The author states that P-values are a Bayesian construction and that we should use them nd feel good about it. He reaches this conclusion by introducing a particular alternative denity so that the P-value almost coincides with the posterior probability of the null hypothesis. 3y this same reasoning one can reach the completely opposite conclusion. For example, by ubstituting $P(t)$ with $1 - P(t)$ (which also has a uniform distribution) in the alternative lensity one has that the P-value, for small δ, coincides with the posterior probability of the *ilternative* hypothesis! The point, of course, is that virtually anything is possible with an *id hoc* alternative, and merely being Bayes with respect to some alternative should not be a ource of much comfort.

DeGroot (1973) considers alternative densities constructed so as to "justify" P-values, out questions whether this class of alternative distributions can be derived from some "naturral" assumptions. Are there any "natural" assumptions that justify the use of the bimodal alternative density $p_{a,n}(x)$ for the normal model in Section 2?

As Hodges states, taking a small δ (prior probability on the null) the P-value is close to the posterior probability of the null. If one is interested in testing a null hypothesis (versus a simple alternative) one surely has *some* prior belief in it being true. For $\delta \simeq 0$, there would hardly be any need to test!

There is a vast literature (see Berger's work) showing that the conclusions reached using P-values are often in disagreement with those reached even using a robust Bayesian approach. I don't think that P-values can be resuscitated for practical use.

G. PARMIGIANI (*Duke University, USA*)

"Who knows what alternative lurks in the heart of significance tests?" In a paper entitled "Doing what comes naturally: Interpreting a Tail Probability as a Posterior Probability or as a Likelihood Ratio" (1973), Morrie DeGroot constructed a family of alternatives that yields, as the title promises, the P-value as the posterior probablity of the null hypothesis. Dr. Hodges develops a different solution, and it is perhaps worthwhile to compare the two briefly.

Hodges proposes a simple alternative distribution on the sample space, given by:

$$p_{a,n}(x) = \frac{\delta}{1-\delta}[(1 - G(t(x)))^{\delta - 1} - 1]f(x),$$

where f is the joint density of the data x and G is the cdf of the test statistic $t(x)$. The posterior probability of the null hypothesis can be made arbitrarily close to the P-value as δ becomes small. On the other hand, DeGroot assigned the alternative directly on the space of outcomes of the test statistic; he used the family of alternatives:

$$p^{\theta}_{a,n}(t) = (1 + \theta)G^{\theta}(t)g(t),$$

indexed by a parameter θ. If θ takes any nonnegative integer value with improper prior distribution $1/(1 + \theta)$, the posterior probability of the null hypothesis coincides with the P-value.

In both cases, the prior probability of the null hypothesis has to be, in some sense, very small. However, DeGroot's solution has the attractive feature that the alternative never depends on the prior probability of the null hypothesis. Also, DeGroot's family leaves the joint distribution unspecified, assuming that the statistic is all that is observed. So it is not possible to directly pose the question of the incoherence of probability assignments for varying n. A natural modification of DeGroot's family, that still leads to an exact P-value, is given by:

$$q^{\theta}_{a,n}(x) = (1 + \theta)G^{\theta}(t(x))f(x).$$

This can be shown to suffer from the same marginalization problems that arise with Hodges' alternative. It would perhaps be interesting to investigate whether there exist a family of joint distributions that yields consistent marginalizations and generates DeGroot's alternative.

K. PÖTZELBERGER (*University of Economics, Vienna, Austria*)

In this talk Hodges presented a Bayesian interpretation of frequentist testing procedures based on the P-value. Formally, this can be done by defining a distribution on the alternative so that the corresponding Bayes factor is a function of the frequentist test statistic. However, this distribution on the alternative does not obey coherency. The distribution is, in many cases, not consistent in the sense that integrating out a part of the data will change the distribution, i.e., the distribution of the alternative depends on what might have been observed. Thus the Bayesian interpretation of the test turn out to exhibit the deficiencies of the test, rather than reconciling classical significance tests with Bayesian ideas.

We discuss the approach by showing that in certain situations using the marginal distribution of the observed data and computing the Bayes factor, can lead to an equivalent procedure. However, the level of the test may not be chosen freely any more. In the following example it has to be $\alpha = 0.32$.

Let $T(x)$ be a frequentist test statistic for a simple null hypothesis H_0 versus a composite alternative H_1. Let $f_n(x)$ denote the distribution of the observation $x = (x_1, \ldots, x_n)$ under the null hypothesis. Furthermore, we assume that under $H_0, ET = 0$ and $\sigma(T) = 1$ and that the null hypothesis is rejected, when $|T(x)| > c$. Define $h_n(x) = T^2(x)f_n(x)$ as the distribution of x under the alternative hypothesis. Then, formally, the corresponding Bayes factor is $B = 1/T^2(x)$, which leads hence to an acceptance of H_0 if $B \geq c_0 := 1/c^2$.

$h_n(x)$ is an exchangeable distribution. Denote by $h_{k,n}$ the marginal of $x^k = (x_1, \ldots, x_k)$ $(k \leq n)$, computed from h_n. Usually, $h_{k,n}$ differs from h_k, indicating that the interpretation of the test based on T is not coherent. One might, however, try to modify the model by replacing h_k by $h_{k,n}$, where $n > k$. This distribution depends on n. As an example we consider $x_i \sim^{iid} N(\theta, 1)$, with $H_0 = \{0\}$ and $H_1 = \{\theta \neq 0\}$. Let $T = \sqrt{n}\overline{x}_n$, so that $h_n \propto \overline{x}^2_n f_n(x)$ and $h_{k,n}(x^k) = \epsilon h_k(x^k) + (1 - \epsilon)f_k(x^k)$ with $\epsilon = k/n$. The corresponding Bayes factor is again a function of f_k/h_k. Precisely,

$$\frac{f_k}{h_{k,n}} = \frac{f_k}{\epsilon h_k + (1-\epsilon)f_k} = \frac{1}{\epsilon h_k/f_k + (1-\epsilon)},$$

so that $f_k/h_k \leq c_0$ if and only if

$$\frac{f_k}{h_{k,n}} \leq \frac{c_0}{\epsilon + (1-\epsilon)c_0}.$$

Again, $c_0/(\epsilon + (1 - \epsilon)c_0)$ should be independent of n, which is the case only for $c_0 = 1$. We conclude that only for $c_0 = 1$, is a Bayesian interpretation of the frequentist test based on $\overline{x}_k$ possible. $c_0 = 1$ leads to $P(|\sqrt{k}\overline{x}_k| > c) \approx 0.32$.

REPLY TO THE DISCUSSION

My objective was to construct an alternative that yields the P-value as the posterior probability of the null, to examine it to see if it clarifies issues in hypothesis testing, and, in conjunction with Hodges (1990), to see whether it might make some Bayesians more comfortable about using P-values for exploratory purposes. This rejoinder clears up some areas of confusion and then considers new material raised in the discussion.

The point of the paper was not to defend P-values for testing a null hypothesis about a parameter θ against an alternative about θ (Berliner, Casella) or against an unstated alternative (Poirier). The point was to see what alternative the P-value tests against. As it turns out, in some generality that alternative cannot be expressed in terms of θ, which makes it plain why P-values appear to do poorly at testing against parametric alternatives: they're set up (implicitly) to test against something else (as George noted).

(In this connection, Mike Lavine of Duke University has pointed out that $p_{a,n}(x)$ is an exchangeable density for cases in which x is exchangeable under the null and T is symmetric in the x_i. This suggests that, in the limit as n approaches infinity at least, $p_{a,n}(x)$ can be expressed as some sort of mixture, although it is beyond me to derive it.)

The construction does not require cheating in the specification of δ (Berger): δ is fixed. Furthermore, $0 < P(x)^{1-\delta} - P(x) < \delta(1-\delta)^{(1-\delta)/\delta}$, with the maximum occurring at $P(x) = (1-\delta)^{1/\delta}$, so the difference between the P-value and the posterior probability is uniformly bounded and becomes arbitrarily small as δ does. Large values of δ are of no interest, although some P-value "pathologies" (such as Lindley's paradox) occur for all values of δ. For the two discrete cases I have worked through, one can set $\delta = 0$ without making the P-value alternative improper.

The final matter of confusion is whether nuisance parameters are a problem (Berliner, Casella). They are not: $f(x)$ is the marginal density of x under the null, with nuisance parameters integrated out against their prior. The construction works for any such prior because $f(x)$ is a factor of the alternative, so it does not appear in the Bayes factor.

Parmigiani pointed out DeGroot (1973), which I was remiss in not discussing. In DeGroot's construction, the null case obtains when his parameter $\theta = 0$, with the positive integral θ making up the alternative. If DeGroot's alternative is expressed as a single density on the test statistic t — by summing out θ from 1 to infinity — it is

$$Q^{-1}f(t)\{P(t)^{-1}-1\} \qquad \text{for} \qquad Q=\sum_{k=0}^{\infty}(k+1)^{-1},$$

where the sum does not converge. If $p_{a,n}(x)$ is re-expressed in terms of the test statistic, then DeGroot's alternative is identical to $p_{a,n}(x)$ with $\delta = 0$. Note that the prior on DeGroot's parameter θ is improper, so that $\theta = 0$ — the null — has probability zero under his construction as it does in the limiting case of $p_{a,n}(x)$.

If the sample space of DeGroot's θ is restricted to the integers $0, 1, \ldots, K$, his distribution across θ is proper, and his alternative becomes

$$Q^{-1}f(t)\left\{\frac{1-F^{K+1}(t)}{1-F(t)}-1\right\} \qquad \text{for} \qquad Q=\sum_{k=0}^{K}(k+1)^{-1},$$

where $F(t)$ is the cdf of $f(t)$. The prior probability of the null hypothesis is $Q^{-1} > 0$; thus if DeGroot's limiting case is approached by letting K become large, his construction, like mine, makes the prior probability of the null depend on the alternative. I believe it is impossible to

re-express this proper-prior version of DeGroot's alternative to make it identical to $p_{a,n}(x$ so it and $p_{a,n}(x)$ are distinct approaches to the P-value.

Given these similarities, DeGroot's motivation was intriguing:

> The purpose of this article is to present a few simple ideas which indicate how the calculation of tail areas can be made compatible with the principles of Bayesian statistics. These ideas, if successful, will serve the dual purpose of putting χ^2 tests, F tests, and other such procedures back into the repertory of the Bayesian statistician and of giving all statisticians the freedom that comes from being able to interpret the evidence exhibited in a tail area simply as a likelihood ratio or as a posterior probability. (p. 967)

If a Bayesian as stalwart as DeGroot could utter such a sentiment, one might wonder at th vehemence with which some discussants denounced it. Some had no problem with DeGroot' aim, and accept "*ad hoc* indicators" even without such a construction (Berger). Others accep the need for adhockery but reject P-values (Smith 1986). Still others balked at adhocker inadmissibility, or incoherence (Berliner, Dawid, Mortera, or Poirier).

What shall we make of such objections? Dawid notes that $p_{a,n}(x)$ is coherent for an given n; if it weren't, then when I'm handed a sample of size n, my beliefs about it would b constrained by samples I could have gotten but didn't. (Sounds familiar?) For other case: the response to $p_{a,n}(x)$'s incoherence makes an interesting contrast with the blasé respons to Bernardo's incoherent reference priors. Berger addressed this issue in his talk here b noting that in any given analysis, he has only so much time to spend and might be willing t use an incoherent reference prior – risking a loss of utility – so he can concentrate his effo on aspects of the problem that could cost him more utility. (See also Bernardo, 1984.) Th same reasoning applies to P-values used in exploration: You have only so much time, Yo may choose to spend less of it by using something simple, incoherent, *and available*, an thereby have more time for other aspects of the analysis.

Mortera and Poirier objected to adhockery without mentioning inadmissibility or incoherence. Mortera considered it absurd that replacing $P(x)$ by $1 - P(x)$ in $p_{a,n}(x)$ yields proper density and a posterior probability of the null of $(1 - P(x))^{1-\delta}$. To the contrary, sh has constructed the usual test of whether the observations agree with the model "too well," as might be used to detect that Mendel's assistant fudged his data. Presumably Mortera i interested in such tests, but she could not make one in the iid normal example simply b changing τ in the BD alternative. Instead, she would need another alternative, one using ii normals with variance less than 1, or a t on 3 *df* scaled to have variance 1. Which is less *a hoc*, the normal or the t? Plainly neither, and neither is less *ad hoc* than the alternative tha Mortera constructed and scorned. Perhaps physicists derive models from first principles, bu few if any others can; when a Bayesian casts about for a model and throws her line into th pool of handy specifications, how is this not *ad hoc*? Thus, I see no force to the objection of Mortera and Poirier regarding adhockery.

The conference President, Dennis Lindley, asked if our view of Bayesian statistics i too narrow. I would say yes, and the discussion illustrates why: our theory is written as i likelihoods are found in the cabbage patch, when in fact they are often created as part of the analysis. As the foregoing indicates, when we consider data analysis in its fullness, some comfortable old verities may not seem so comfortable anymore. At a minimum, it is clear that Bayesians are far from a consensus on how to think about real data analyses, and we should not hesitate to re-examine cherished beliefs in the search for a consensus.

ADDITIONAL REFERENCES IN THE DISCUSSION

Barnard, G. A. (1972). Review of *The Logic of Statistical Inference* by I. Hacking. *British J. Philosophy of Science* **23**, 123–190.

Barnard, G. A. (1980). Pivotal inference and the Bayesian controversy. *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Valencia: University Press, 293–318, (with discussion).

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of *P* values and evidence. *J. Amer. Statist. Assoc.* **82**, 112–122.

Bernardo, J. M. (1984) Discussion of Geisser (1984). *Ann. Statist.* **38**, 247–248.

Brown, P. J. (1980). Coherence and complexity in classification problems. *Scandinavian J. Statist.* **7**, 95–98.

Casella, G. and Berger, R. L. (1987a). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J. Amer. Statist. Assoc.* **82**, 106–111.

Casella, G. and Berger, R. L. (1987b). Comment on "Testing precise hypotheses" by J. O. Berger and M. Delampady. *Statist. Sci.* **2**, 344–417.

Casella, G. and Wells, M. T. (1990). Reconciliation, coherence, and *P*-values. *Tech. Rep.* 1100, Cornell University.

Cornfield, J. (1970). The frequency theory of probability, Bayes theorem, and sequential clinical trials. *Bayesian Statistics* (D. L. Meyer and R. O. Collier, Jr., eds.). Itasca, IL: Peacock, 1–28.

Dawid, A. P. (1988). The infinite regress and its conjugate analysis. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Oxford: University Press, 96–110, (with discussion).

Dawid, A. P. (1980). Discussion of Barnard (1980). *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Valencia: University Press, 311.

DeGroot, M. H. (1973). Doing what comes naturally: interpreting a tail area as a posterior probablity or as a likelihood ratio. *J. Amer. Statist. Assoc.* **68**, 966–969.

DeGroot, M. H. (1980). Discussion of Barnard (1980). *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Valencia: University Press, 311–312.

Geisser, S. (1984). On prior distributions for binary trials. *Ann. Statist.* **38**, 244–247.

Hall, A. (1987). The information matrix test for the linear model. *Review of Economic Studies* **54**, 257–263.

Hwang, J. T., Casella, G., Robert, C., Wells, M. and Farrell, R. (1991). Estimation of accuracy in testing. *Ann. Statist.* **19**.

Hwang, J. T. and Pemantle, R. (1990). Evaluation of estimators of statistical significance under a class of proper loss functions. *Tech. Rep.* Cornell University.

Jeffreys, H. (1961). *Theory of Probability*. Oxford: Clarendon Press.

Kadane, J. B. (1980). Discussion of Barnard (1980). *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Valencia: University Press 315–316.

Lindley, D. V. (1978). The Bayesian approach. *Scand. J. Statist.* **5**, 1–26, (with discussion).

Lindley, D. V. (1986). Discussion of "Test of significance in theory and practice". *The Statistician* **35**, 502–504.

Poirier, D. J. (1988). Frequentist and subjectivist perspectives on the problems of model building in economics. *Journal of Economic Perspectives* **2**, 121–170, (with discussion).

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.