The book's title & copyright page are at the end

Bayesian and Likelihood Methods in Statistics and Econometrics
S. Geisser, J.S. Hodges, S.J. Press and A. Zellner (Editors)

CAN/MAY BAYESIANS DO PURE TESTS OF SIGNIFICANCE?

James S. Hodges
The RAND Corporation

Abstract

Barnard has often argued that the likelihood can't be used to test hypotheses without alternatives; significance tests must be. This paper reviews the writings of Barnard and Bayesians on the subject; refutes Barnard's argument that Bayesians cannot test against all alternatives, but argues that it is unnecessary; then refutes Bayesian arguments against exploratory use of significance tests by constructing an appealing alternative distribution and prior probabilities on the null and alternative that yield (to arbitrary accuracy) the P–value as the posterior probability of the null. The message is that Bayesians can and should use P–values in an exploratory role.

1. Introduction

1.1 The Role of Significance Tests

One recurring theme in George Barnard's writing[1] is a distinction between situations in which several hypotheses are to be compared in the light of some data, and situations in which the plausibility of a single hypothesis is to be evaluated. For the former case, Barnard argued (1949, 1962, 1972a, 1975, 1985; Barnard, Jenkins, and Winsten 1962) that likelihood ratios or some transform of them should be used to compare the hypotheses. For the latter case, no such comparison is available, and Barnard (1962, 1972a, 1975, 1985) argued that this creates a need for a method that does not rely solely on the likelihood. This need is met by significance tests.

In Barnard's view (1962, 1972a, 1972b, 1975, 1980a, 1985), the quintessential role of significance tests is to assess "whether agreement between this hypothesis and the data is so bad that we must begin to exercise our imaginations to try to think of an alternative that fits the facts better" (1972a, p. 129). "If the data do not fit that [hypothesis], then it is worthwhile going ahead. If [the hypothesis] is consistent with the data let us not waste our time" (1962, p. 85). This implies that the test

[1]References to Barnard's papers will be given by the date of the paper only.

is conducted "*before* seriously considering alternatives" (1980a, p. 305), and thus with alternatives that are, at best, vaguely specified (1975, p. 260). Moreover, the need to conduct the test before considering alternatives apparently implies that any attempt to use a Bayesian approach to this problem would require the specification of *all* possible alternative hypotheses, which Barnard considers an impossible task (e.g. 1985, p. 5; see also Anscombe 1963. Many Bayesians consider this a misperception.). In this respect, among others, Barnard views Bayesian methods as deficient.

The usual multiplicity of differing views can be found among those generally identified as Bayesians. A great deal of Bayesian writing has treated individual significance tests in isolation, as devices for assessing nebulous things like the "evidence" relating to a hypothesis. For example, long articles by Pratt (1965) and Berger and Delampady (1987) make only passing mentions of an exploratory role for significance tests, and Berger and Sellke (1987) makes none at all. The emphasis in these papers is on the relationship between P–values and Bayesian hypothesis tests of a single value of a parameter against a simple or composite alternative stated in terms of that parameter. But the quotes above clearly point to the intrinsically exploratory role of significance tests (see also Anscombe 1963, Box 1980, Andrews 1985), and so this large body of work is largely irrelevant to the issue at hand.

Lindley (1980, 1983) argued "that it does not make any sense to test a hypothesis without alternatives in mind" (1983, p. 435). This approach avoids Barnard's criticism by dismissing Barnard's problem. But any birdwatcher knows that this solution is not acceptable: sometimes the bird you see is plainly not of any species you know, and your methods must be capable of telling you so. Jeffreys is often associated with the position just attributed to Lindley (see, e.g. Hill 1986, p. 226), although his actual position is that it does not make any sense to *reject* a hypothesis in a significance test without one to replace it (Jeffreys, 1961, pp. 383-398).

> Is it of the slightest use to reject a hypothesis until we have some idea of what to put in its place? If there is no clearly stated alternative, and the null hypothesis is rejected, we are simply left without any rule at all, whereas the null hypothesis, though not satisfactory, may at any rate show some sort of correspondence with the facts. [p. 390] ... There has not been a single date in the history of the law of gravitation when a modern significance test would not have rejected all laws and left us with no law. Nevertheless the law did lead to improvement for centuries ... [p. 391]

Nonetheless, Jeffreys (1961) recognized the need to assess deficiencies in a hypothesis even if no alternative were available:

> The fundamental idea [of classical significance tests], and one that I should naturally accept, is that a law should not be accepted on data that themselves show large departures from its predictions. [p. 384] ... If the null hypothesis is not altogether satisfactory we can still point to the apparent discrepancies as possibly needing further attention, and attention to their amount gives an indication of the general magnitude of the errors likely to arise if it is used; and that is the best we can do (p. 391)

that is, without an alternative hypothesis in favor of which the null might be rejected.

Lindley (1980) and several others (e.g. Zellner 1980, or Berger and Wolpert 1984, section 4.4) have countered that in any significance test, some alternative is lurking in the background, and the test can only be improved by making the alternative explicit. (Sections 3.2 and 3.3 of this paper develop this point, though to a different end.) Barnard (1965, 1980b) offered counterexamples to this, arising in physics; Dempster (1971), Dawid (1980), Geisser (1980, 1989) and others expressed similar views. Box (1980) offered a particularly forceful rebuttal:

> [S]uppose I have an office that looks onto, say, Oxford Street in London, normally thronged with people. One day I look out of the window at 11 o'clock in the morning and notice that there are only two people in the whole street. My initial reaction surely is that on the null (*status quo*) model this is an unusual event possibly worthy of further investigation. Alternative models that might explain the phenomenon come later. ... [T]he basis of the initial reaction, which requires no alternatives, is surely that I have (or could have) looked out of the window on many previous occasions and rarely have I (would I have) seen as few as two people in the street. The motivation is economy of effort and is employed by all of us hundreds of times in our daily lives—when the null model is plausible I will not worry, but when data make it implausible perhaps I should be concerned. (p. 429)

The notion here is logically prior to and more primitive than the specifics of Bayesian or classical tests. The idea is: if I see something odd—if Oxford Street is empty or if people are packed into it so tightly that they cannot move—*then* I am moved to seek an explanation for this odd observation.

Hill (1986) allowed that the only purpose of a test without alternatives is "to help one decide whether it is better to continue to use the only model one has, ... which may or may not be adequate for the purposes one has in mind, or whether it is better to search for new and improved models or hypotheses" (p. 229). (See also Hill's discussion of Barnard 1980a.) Like Barnard, he found this to be "a very different activity from that involved in comparing two sharply formulated hypotheses, [that] should not be confused with the latter" (p. 229), although he did not accept significance tests as a consequence. Hill (1974) offered a Bayesian competitor to significance tests for linear models with additive errors, although his approach could be applied to other cases; his method specifies a large but not exhaustive class of alternative hypotheses, and thus is apparently vulnerable to Barnard's and Anscombe's criticisms.

Smith (1986) distinguished two forms of model criticism:

> *global criticism*, which basically asks the question "should we abandon the current framework completely, despite the fact that we have nothing at all to propose in its place?"; [and]

local criticism, which asks "should the model be modified or extended in this or that particular direction?" (p. 97)

The latter is plainly susceptible to Bayesian methods; the former "does not appear to make sense in a Bayesian context without some form of assumption about alternative models" (p. 98). In Smith's scheme, these forms of model criticism are applied once a formal framework for a given problem situation has been determined. In the earlier stage in the scientific learning cycle in which the formal framework is constructed, "much of this activity is very definitely 'informal' and not necessarily within the purview of Bayesian thinking, which requires a more or less structured framework" (p. 97). Smith specifically excluded "conventionally sanctioned goodness-of-fit criteria (e.g. deviance)" from the collection of techniques that are legitimate at this stage. One can infer that an acceptable approach to scientific learning would include poking through the data with informal exploratory techniques, such as plots, until some particular formal model is postulated, after which only elaborations and simplifications of this model would be considered, and then only in a Bayesian fashion. Presumably, if doubts begin to arise about the adequacy of the overall framework, the proper response is not a test of significance but a return to the informal exploratory phase. Smith (1986) does not indicate the means by which these doubts might begin to arise.

Box (1980) pointed out that diagnostic plots—a mainstay of Smith's exploratory phase—are animated by the same logic as a significance test. These "plots are designed to make manifest certain 'features' of the data that would rarely be extreme, if the model were true," so that using them to motivate searches for alternative models is just like using a significance test. But this did not bother Box; indeed, he argued (Box 1980) that criticism of models cannot be done within the Bayesian framework ("the difficulty with any attempt to use the Bayesian half of the model alone is that it is eternally conditional. We can move the conditionality around but we cannot lose it," p. 427), so that Bayesian methods must yield to sampling methods for these purposes. (Dempster 1971 made a similar argument.) Box advocated using the predictive distribution of the observed data, conditional only on the model, that is, after integrating out the parameters. In various places in the article he suggested using tail areas calculated under the predictive distribution, or examining its value at the observed data, much as Jeffreys suggested. Geisser (1989) offered several other ways to use predictive distributions for the special case of individual observations that may be discordant with a tentatively entertained model.

Others have taken a pragmatic position: if one's object is to catch certain kinds of deviations from a model, then postulate an alternative that will result in a Bayesian hypothesis test that catches those deviations and don't worry about whether the alternative is a plausible model of anything. Thus West (1986) devised a Bayesian method for monitoring the adequacy of a dynamic linear model. It uses a single alternative at each time t, constructed by increasing the variance of the model entertained at time t. This alternative allows detection of outliers and abrupt shifts in the level of the time series, but it is not motivated by any belief in the truth of the alternative as a description of the world.

Leonard (1975), Zellner (1975), and Chaloner and Brant (1988) went a step farther by in effect choosing capacious alternatives that not only allow the detection of a large number of model failures, but suggest correctives as well. Zellner (1975) treated the unknown actual residuals of a linear regression (*i.e.* actual dependent values minus the unknown correct regression) as the objects of interest. These residuals are linear functions of the data and the unknown parameters, so that once the data have been observed, their joint posterior distribution—conditional on the current regression model—can be obtained easily from the posterior of the regression coefficients. The posterior distribution of the realized residuals can be examined for evidence of model failures, and also to search for alternative models. (Hill 1969 examined other uses of the distribution of the realized residuals.)

Leonard (1975) and Chaloner and Brant (1988) treated realized residuals as unknown parameters in similar ways. Chaloner and Brant said that their approach—which is essentially Zellner's, but with a different emphasis—"is a general exploratory method of investigating whether there is a problem when no alternative model is immediately apparent" (p. 2). Zellner used very similar words to describe his method, although he noted that once his procedure had been used for exploration, one could "compute posterior odds ratios relating to variants of a standard model ... that are helpful in model choice" (p. 130; also Dempster 1971, p. 61).

In sum, significance tests fill an exploratory role, and Bayesian researchers have generally agreed that there is a role to be filled. Some have accepted ordinary significance tests or variants on them, while others have attempted Bayesian surrogates in the apparent absence of any possibility of finding a palatable direct analog. However, little systematic scrutiny has been given to Barnard's assertion that no analog is possible, and no-one has been able to explain significance tests in satisfactory Bayesian terms.

1.2 Answering Barnard Directly

Though their other merits are considerable, none of these approaches addresses Barnard's criticism head on, leaving the impression that Barnard has indeed found a problem that Bayesian methods cannot handle. In section 2.1, I show that for a large class of problems, it is possible to specify all of the alternatives to a given

model; in fact it is quite simple. However, it is also quite unnecessary, ever, and the simpler construction—the Bayesian analog to significance tests—is given in Section 2.2. In section 3.1, this Bayesian construction is shown to have the same structure as a significance test a la Barnard. Many Bayesian writers (e.g. the contributions of Dickey, Good, Hill, Leonard, and Zellner to the discussion of Barnard 1980a) have attempted to interpret P-values from significance tests as approximations to Bayesian quantities. For pure significance tests, though—as Barnard (1980a, p. 317) pointed out—such approximations require a specification of an alternative, which the authors listed above did not give. In sections 3.2, I show that classical significance tests are equivalent to any of a very large class of Bayesian significance tests in the sense that they produce the same ranking of possible data outcomes according to their discrepancy from the model. In section 3.3, the P-value, viewed as a function of the data, is shown to be, to arbitrary accuracy, the posterior probability of the null model, for a specific alternative and a specific prior probability of the null. Section 4 discusses these results and their implications for Bayesian arguments against the use of significance tests as exploratory tools.

The main point here is intended mostly for Bayesian readers. Bayesian analogs to "alternative-free" tests are straightforward to define, and it turns out that in a large class of cases, classical significance tests are essentially the same as these Bayesian tests. Given the exploratory—and thus rough-and-ready—role that significance tests are intended to fill, it follows that little can be gained by a finicky concern over logical niceties and an attendant refusal to make exploratory use of the abundance of non-Bayesian computer software, especially when competing Bayesian software is non-existent or much more difficult to use. To anticipate the results further, West's (1986) method, cited above, is an archetype of the Bayesian analog to significance tests. Like a significance test, West's test tells its user that something is wrong, but it doesn't say what; the user must figure that out some other way. We have some Bayesian methods that *do* suggest what is wrong (also cited above), so perhaps the message to be drawn from all this is that it is time we stopped worrying about significance tests and spent our energies developing more methods like those of Leonard, Zellner, and Chaloner and Brant.

2. *Can* Bayesians Do Pure Tests of Significance?

2.1 A Bayesian Can Specify All Possible Alternatives for Some Problems

Let $X = (x_1, x_2, \ldots, x_n)$ be an n-vector of observables, each of which can take a finite number of values. For ease of exposition, let each x_j be a binary observable,

taking the values 0 and 1; extension to the more general case is straightforward. (Berger and Wolpert 1984 allow countable sample spaces, but restrict themselves to independent and identically distributed random variables.) Any probability distribution for X, with arbitrary dependencies, can be specified as a 2^n-vector p as follows. Label the coordinates of p with the 2^n patterns of 0's and 1's, so that $p = (p_{00\ldots0}, p_{10\ldots0}, \ldots, p_{11\ldots0}, p_{11\ldots1})$. The coordinates of p are in the closed interval from zero and one, they sum to one, and each coordinate gives the probability of observing the value of X that is its subscript. Call the collection of such p's S; it is the space of all probability distributions for X. Any alternative to a given hypothesis regarding the distribution of X can be represented as a subset of S, and so all alternatives are captured in it. (Diaconis 1977, Meeden 1986, Draper 1988, and others have used this construction for different purposes.) Thus, existing Bayesian hypothesis testing ideas can (not to say should) be used in this manner to construct an explicit test of a model against all alternatives.

For example, consider testing whether, conditional on θ, the x_j are independent and identically distributed Bernoulli random variables with $Pr(x_j = 1 \mid \theta) = \theta$. In the manner of Bayesian hypothesis tests, let H_o be this Bernoulli model with a beta distribution as the prior for θ:

$$f(\theta|a,b) = B(a,b)^{-1}\theta^{a-1}(1-\theta)^{b-1} \quad \text{for} \quad \theta \in (0,1) \tag{1}$$

where $B(a,b)$ is the beta function. (In practice, of course, the prior would depend on the situation.) For the alternative, let H_1 be S, the space of all models, with the flat prior distribution $f(p) = \Gamma(2^n)$ on the simplex S. (In practice one would almost certainly use another prior, as will be seen below.) Let the prior probabilities for H_o and H_1 be π and $1-\pi$, respectively.

By a straightforward application of familiar Bayesian theory,

$$Pr(H_o|X) = Q\ Pr(X|H_o)Pr(H_o) = Q\pi B(a,b)^{-1}B(a+Y, b+n-Y) \tag{2}$$

and

$$Pr(H_1|X) = Q\ Pr(X|H_1)Pr(H_1) = Q(1-\pi)2^{-n} \tag{3}$$

where Q is a proportionality constant, $Y = \Sigma x_j$, and

$$\int_S p_X f(p)dp = \Gamma(2^n)/\Gamma(2^n+1) = 2^{-n} \tag{4}$$

(This last is easy to derive directly, without computing the integral.) Thus, for $a = b = 1$, the posterior odds ratio $Pr(H_o|X)/Pr(H_1|X)$ is

$$\frac{\pi}{1-\pi}\ \frac{2^n\Gamma(Y+1)\Gamma(n-Y+1)}{\Gamma(n+2)} \tag{5}$$

which, as it turns out, has a number of unpalatable properties. The most important—which holds for all choices of a and b—is that it depends on the data only through Y, so that according to this test one's posterior beliefs about the tenability of the binomial model depend only on the sufficient statistic under that model.

Apart from exemplifying the point that Bayesians can indeed test against all alternatives, this example illustrates once again that the outcome of a Bayesian hypothesis test must depend on which prior is used conditional on the null and the alternative hypotheses. Parenthetically, it also illustrates that the naive flat prior on S is a piece of information with less than obvious implications, which should be used with great caution.

Contrary to Barnard's assertion, then, a Bayesian can sometimes specify all alternatives and, in the usual manner, test a precisely specified hypothesis against them. But a reading of Barnard's papers on this subject suggests that this concern might not be satisfied by S and the constructed hypothesis test, and that he might raise several objections at this point. The first might be that this construction fails as a prospective way of encompassing all possibilities, because it will not capture unanticipated outcomes like the flipped coin that is lost in the flipping (1949, p. 136). But Barnard's (1947) discussion of a Fisherian significance test regarding seeds clearly permits the *post hoc* construction of the sample space, so this objection will not hold.

A more interesting objection is that the combination of S and a prior distribution on it cannot properly represent the nature of uncertainty about hypotheses as yet unconceived or inconceivable (1985, p. 5). But the construction permits positive chunks of prior probability to be assigned to subspaces corresponding to specific parametric models; and is it unreasonable that point hypotheses elsewhere are given prior (and thus posterior) probability zero? Posterior probability in S can still aggregate in the neighborhood of such hypotheses, and besides, the failure to assign positive probability to a specific model *a posteriori* does not seem to be a disadvantage; it may even be preferable.

For a third possible objection, Barnard's papers show a strong disinclination to place prior probabilities on hypotheses as a matter of course, so that he might object to the whole idea of placing a prior distribution on S, judging it to be incapable of sufficiently precise specification and thus arbitrary and unworthy of the name "probability" (1985, p. 5). But this would hold the Bayesian approach to a higher standard than that applied to significance tests by Anscombe (1963), which Barnard (1972a) endorsed. This point is discussed further in section 3.1.

2.2 But It Is Not Necessary to Specify All Possible Alternatives

It is possible to specify all alternatives for some problems, but the derivation in the last section shows that there is no need to do so, for this type of problem or for any other. Recall that

$$Pr(H_1|X) \propto Pr(X|H_1)Pr(H_1) = (1-\pi)2^{-n}, \tag{6}$$

so that the posterior probability of H_1 depends only on the marginal probability of the data given H_1, that is, on the *a priori* predictive probability of the data given H_1. Thus, the only new thing that a Bayesian needs to make an analog to the pure significance test is an alternative predictive distribution for the data—not a parametric alternative or even a class of parametric alternatives. (This fact was alluded to by Pratt 1965 in his rejoinder to the discussion.) In this sense, West's (1986) "model monitoring" method is the archetype of the Bayesian significance test.

3. The Similarity of Bayesian and Conventional Significance Tests

3.1 Both Tests Rank the Possible Data Outcomes by Their "Discrepancy"

A conventional significance test requires two ingredients: a hypothesized model and a function that assigns a "discrepancy" to the possible data outcomes (1962, 1980a; Cox 1977; Anscombe 1963 uses the term "test criterion"). The discrepancy function sorts all possible data outcomes from the least to the most discrepant; when the actual data are observed, the P-value is computed as the probability, under the hypothesized model, of observing an outcome as discrepant as or more discrepant than that which was actually observed. A conventional significance test, then, has two products: a ranking of the possible outcomes and a method of attaching a probability to the sorted outcomes. The choice of this discrepancy function is quite arbitrary; it depends on the model failures one considers likely or interesting (Anscombe 1963). To Jeffreys (1961), this was a disadvantage, because for any set of data, some function is likely to look odd. But as will become clear shortly, Bayesian methods are no more immune to this criticism than classical tests.

A Bayesian significance test, as defined in section 2, requires four ingredients: the same hypothesized model as in the conventional significance test, plus a prior distribution on the parameters of the hypothesized model, a predictive distribution that specifies the alternative hypothesis and a prior probability of the null hypothesis. (Actually, three ingredients suffice: predictive distributions for the null and alternative, and the prior probability of the null.) The prior distribution on the parameters and the alternative predictive distribution produce, as in the conventional case, a criterion for ranking the possible data outcomes from least to most discrepant—where in this case increasing discrepancy corresponds to decreasing posterior probability of

the hypothesized model—but altogether these ingredients yield a different method for tacking probabilities onto the ranked outcomes.

This structural similarity removes the force of the third objection to the construction of section 2.1, that priors on S (and thus alternative predictive distributions) are arbitrary. The arbitrariness of the alternative predictive distribution corresponds precisely to the arbitrariness of the discrepancy function of a conventional significance test. If the latter is tolerable for the limited legitimate purpose of significance testing, the former must be tolerable as well. This structural similarity also removes the force of Jeffreys' criticism (noted above) of classical tests based on the arbitrariness of the discrepancy function. It is possible to argue that Jeffreys' criticism still holds as long as the alternative models are chosen before looking at the data, but this requirement is incompatible with the exploratory function; if any kind of test is to be useful in exploration, it must be used to consider things that were not envisioned before the data were analyzed.

The foregoing shows that both Bayesian and conventional significance tests produce rankings of the possible data outcomes and formulae for assigning a measure of discrepancy calibrated as a probability. This prompts two questions: (i) can sorting algorithms produced by significance tests be duplicated by Bayesian tests, and (ii) if so, are the P-values posterior probabilities for any interesting constructed alternative and prior for the null model? The answers will be shown in sections 3.2 and 3.3 to be (i) yes, always, and (ii) to an arbitrary degree of accuracy, for significance tests in which the P-value is distributed as a uniform random variable under the null hypothesis. These results extend those in Cox (1962, p. 84-85) and DeGroot (1973). Many Bayesian authors show similar relations between classical significance tests of a null hypothesis and Bayesian hypothesis tests of the same null against a specific alternative. For example, Zellner (1984) gives an example of a Bayesian test of a hypothesis about a regression parameter in which the P-value is identical to the posterior probability that the hypothesis is true. The correspondence between Bayesian and classical methods in this literature is different from the one to be developed now.

3.2 The Significance Test's Ranking Can Always Be Reproduced Exactly

Let X be a random variable (scalar, vector, matrix, or whatever) and let x be a realization of X. Let $T = T(x)$ be a test statistic or test criterion that can be calculated from x, and suppose without loss of generality that it is to be used in a test of significance in which larger values will be treated as less favorable to the null model than smaller values. Let $P(x)$ be the P-value computed for the observation x from the null model; P is actually a function of T alone, but in the sequel the argument of P will be given as T or x as convenience and clarity dictate. For now,

this need not be an exact P-value (although in section 3.3, it will), but it must be a function of the data and not of any unknown parameters. Let $F(x)$ be the marginal distribution function of these data x under the hypothesized model, and for convenience let it have a density $f(x)$ with respect to Lebesgue measure. (This does not affect the generality of the result.) This distribution is marginal with respect to the prior for any unknown parameters of the null model; no restrictions are required on that prior except that $F(x)$ must exist.

Let $\phi(T)$ be any positive monotone increasing function that can take T as an argument, and define an alternative density for X by $p_a(x) = K_\phi f(x)\phi(T(x))$, where K_ϕ is the constant of proportionality that makes p_a a probability distribution. This alternative places more mass in those regions where the test statistic is large, which mimics the intuition behind selecting the test statistic. (DeGroot 1973 used a very similar construction for a somewhat different purpose; Cox 1962 used a special case.) The requirement that K_ϕ be non-zero places a further constraint on $\phi(T)$, which I will henceforth assume to be satisfied. This is not generally an onerous restriction; for statistics T having moment generating functions under the null model, for example, the exponential function will be permissible. If π is the prior probability of the null model, then the posterior probability given x is

$$\left[1 + \frac{1-\pi}{\pi}\frac{p_a(x)}{f(x)}\right]^{-1} = \left[1 + \frac{1-\pi}{\pi}K_\phi\phi(T)\right]^{-1}, \tag{7}$$

so that the ranking produced by this alternative is equivalent to the ranking given by $\phi(T)$, and thus, by the monotonicity of ϕ, equivalent to the ranking given by T itself.

Thus the ranking given by the significance test can be reproduced by a Bayesian significance test, indeed, by any Bayesian significance test in a large class.

3.3. Usually the Significance Test's P-Value Is, To Arbitrary Accuracy, a Posterior Probability

For cases in which an exact continuous distribution can be given for the test statistic T, the P-value $P(T)$ 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 of the observation X under the null. For the remainder of this section, I will restrict consideration to test statistics and P-values having this property. This includes all of the familiar tests associated with the analysis of variance, for example. It excludes approximate P-values and P-values for discrete distributions; results similar to those below can be produced for at least some of these cases, but I will not do so here.

For significance tests in this restricted class, construct an alternative distribution by specifying $\phi(T) = P(T)^{\delta-1} - 1$ in the notation of section 3.2, where $0 < \delta < 1$. Then $K_\phi^{-1} = E(P(T)^{\delta-1}) - 1$, where the expectation is with respect to the marginal distribution $F(x)$, that is, where $P(T)$ has a uniform distribution. Set the prior probability of the null hypothesis, π_ϕ, so that it satisfies $(1 - \pi_\phi)K_\phi/\pi_\phi = 1$; this gives $\pi_\phi = K_\phi/(1+K_\phi)$. Because we are only considering cases for which $P(T)$ has a uniform distribution, $K_\phi^{-1} = (1-\delta)/\delta$ and $\pi_\phi = \delta$. For this alternative distribution and this prior probability on the null hypothesis, the posterior probability of the null is $P(T)^{1-\delta}$. That is, the P-value can be made arbitrarily close to the posterior probability of the null model, whatever x is, for a specific alternative and prior probability of the null that do not depend on the data. In the limit as δ goes to zero, however, the constructed alternative becomes improper and π_ϕ goes to zero.

4. So What? (*May* Bayesians Do Pure Tests of Significance?)

As section 1.1 made clear, some people associated with the Bayesian viewpoint have no difficulty with the use of significance tests for exploratory purposes. Others continue to object because significance tests appear to be in violation of the likelihood principle (Lindley 1980, Berger and Wolpert 1984), or because they are apparently otherwise incapable of a satisfactory Bayesian explanation. It appears that the basis of these continuing objections has been weakened considerably.

Pratt (1965) argued that significance tests suppress uncertainty by hiding the uncertainty about the implicit alternative hypothesis. But section 3.2 showed that for a very broad class of possible alternative hypotheses, the ranking produced by the resulting Bayesian significance test is identical to that of the Barnardian significance test. To insist that this class is not large enough is to insist on a logical point beyond any practical implication; certainly this class of alternatives captures the vague notion that moves a statistician to consider a particular test statistic. Continuing objection can only be based on the assignment of the specific posterior probability, described in section 3.3.

The first point here, shown in section 3.3, is that for a large class of commonly used significance tests, the P-values *are*, to arbitrary accuracy, posterior probabilities for a specific constructed alternative hypothesis and prior probability of the null. So the only objection that seems to remain is that the P-value is only *a* posterior probability, not *my* posterior probability in some given instance. But what would one's posterior probability be in this situation? To quote Dempster (1971, p. 60):

> In principle ... the concept of a prior probability of a null hypothesis clashes with the very concept of a null hypothesis as a tentative stopping place on the way to devising a model in accord with the known facts. One knows that no null hypothesis could ever be true, but hopes to find one that fits and illumines the facts. Even to conceive of a numerical degree of certainty attached to such an entity is awkward, letting alone the practical question of specifying the number. In practice, when the data are such that judgments of significance are borderline, the dependence of the posterior probability of the null hypothesis on its prior probability will be crucial, and the user will be led to agonize over a largely fictitious aspect of his prior knowledge at the expense of constructively examining his available data.

This is, ultimately, the point: why waste any energy over the remaining difference? Berger and Delampady (1987), among others, argue that one can always construct an alternative in cases apparently without one, and that one should do so and then conduct the Bayesian test. But the P-value is precisely the result of such a construction; so why not use it for exploration — especially when we have so many readily available tools for computing P-values, and almost no tools for the competing Bayesian quantities? It is a counsel of inefficiency to insist that we should not use what is available and should use what is not available, particularly when the latter offers no practical advantage and the most ethereal of logical advantages.

4.1 A Footnote

Many Bayesians (e.g., Pratt 1965) have found it objectionable that significance tests will, for a large enough sample, find any deficiency in a model, no matter how small. It is odd that this should be considered a disadvantage; what are we to make of a tool for finding problems that fails to detect them under advantageous circumstances? To quote again from Dempster (1971, p. 62):

> Some may hold that the beauty of science lies in a conjunction of truth and simplicity, and certainly there are way stations along the path of science which give that impression, but I would hold ... that the real world, even the real world of some quite restricted scientific phenomenon, is endlessly complex. ... The real life process of constructing a model may be roughly conceived as operating between a floor and a ceiling where increasing altitude means increasing *complexity* of the model. The floor pushing complexity upward consists in the need for the model adequately to fit the known facts of the phenomenon under study. The ceiling, however, is a much less well understood logical construct which I should like to label *confusion*. Confusion arises because too many dimensions of an overly complex model are insufficiently determined by the available facts, so that predictions and insights from the model are muffled. I believe, in other words, that there are inherent limitations on the process of estimation which imply that restricted fixed data simply cannot determine a broadly acceptable posterior distribution on the many parameters of a highly complex model. Thus, the processes of model searching and estimation interact through the determination of a comfortable altitude of complexity at which the available supply of factual data can be reasonably absorbed and interpreted.

As a body of data gets larger, should we not expect, indeed, desire that it will reveal deficiencies in any currently held model? For various reasons we may choose

to ignore the revealed deficiencies, but that does not mean that we are better off ignorant of them.

Acknowledgements

While writing this paper I received the thoughtful comments of Seymour Geisser, Bruce Hill, John Pratt, and Arnold Zellner. They do not necessarily share and are not responsible for the judgments and interpretations expressed in this paper.

References

Andrews, D.F. (1985). Discussion of Barnard (1985).

Anscombe, F.J. (1963). "Tests of Goodness of Fit" (with discussion), *Journal of the Royal Statistical Society, Series B*, Vol. 25, 81-94.

Barnard, G.A. (1947). "The Meaning of a Significance Level," *Biometrika*, Vol. 34, 179-182.

Barnard, G.A. (1949). "Statistical Inference" (with discussion), *Journal of the Royal Statistical Society, Series B*, Vol. 11, 115-149.

Barnard, G.A. (1962). Prepared contribution and discussion in *Foundations of Statistical Inference*, L.J. Savage, ed., London: Methuen.

Barnard, G.A. (1965). Comment on Pratt (1965).

Barnard, G.A. (1972a). Review of I. Hacking, "The Logic of Statistical Inference," *British Journal for the Philosophy of Science*, Vol. 23, 123-190.

Barnard, G.A. (1972b). "The Unity of Statistics" (with discussion), *Journal of the Royal Statistical Society, Series A*, Vol. 135, 1-14.

Barnard, G.A. (1975). Comment on J.D. Kalbfleisch, "Sufficiency and Conditionality," *Biometrika*, Vol. 62, 251-268.

Barnard, G.A. (1980a). "Pivotal Inference and the Bayesian Controversy" (with discussion), in *Bayesian Statistics*, J.M. Bernardo, M.H. DeGroot, D.V. Lindley, A.F.M. Smith, eds., Valencia, Spain: University Press.

Barnard, G.A. (1980b). Comment on Box (1980).

Barnard, G.A. (1985). "A Coherent View of Statistical Inference" (with discussion), Technical Report, University of Waterloo, Department of Statistics and Actuarial Science.

Barnard, G.A., G.M. Jenkins, and C.B. Winsten (1962). "Likelihood Inference and Time Series" (with discussion), *Journal of the Royal Statistical Society, Series A*, Vol. 125, 321-372.

Berger, J.O. and M. Delampady (1987). "Testing Precise Hypotheses" (with discussion), *Statistical Science*, Vol. 2, 317-352.

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

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

Box, G.E.P. (1980). "Sampling and Bayes Inference in Scientific Modelling and Robustness" (with discussion), *Journal of the Royal Statistical Society, Series A*, Vol. 143, 383-430.

Chaloner, C. and R. Brant (1988). "A Bayesian Approach to Outlier Detection and Residual Analysis," *Biometrika*, Vol. 75, No. 4, 651-659.

Cox, D.R. (1962). *Discussion in Foundations of Statistical Inference*, L.J. Savage, ed., London: Methuen.

Cox, D.R. (1977). "The Role of Significance Tests" (with discussion), *Scandinavian Journal of Statistics*, Vol. 4, 49-70.

Dawid, A.P. (1980). Comment on Box (1980).

Dempster, A.P. (1971). "Model Searching and Estimation in the Logic of Inference" (with discussion) in *Foundations of Statistical Inference*, V.P. Godambe and D.A. Sprott, eds., Toronto: Holt, Rinehart, and Winston.

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

Diaconis, P. (1977). "Finite Forms of de Finetti's Theorem on Exchangeability," *Synthese*, Vol. 36, 271-281.

Draper, D.C. (1988). "Rejoinder," *Statistical Science*, Vol. 3, No. 2, May, 266-271.

Geisser, S. (1980). Comment on Box (1980).

Geisser, S. (1989). "On Predictive Tests of Discordancy," in *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George Barnard*, S. Geisser, J.S. Hodges, S.J. Press, A. Zellner, eds., Amsterdam: North-Holland.

Hill, B.M. (1969). "Foundations for the Theory of Least Squares," *Journal of the Royal Statistical Society, Series B*, Vol. 31, No. 1, 89-97.

Hill, B.M. (1974). "On Coherence, Inadmissibility, and Inference About Many Parameters in the Theory of Least Squares," in *Studies in Bayesian Econometrics and Statistics: In Honor of Leonard J. Savage*, S.E. Fienberg, A. Zellner, eds., New York: North-Holland/American Elsevier.

Hill, B.M. (1986). "Some Subjective Bayesian Considerations in the Selection of Models" (with discussion), *Econometric Reviews*, Vol. 4, 191-246.

Jeffreys, H. (1961). *Theory of Probability*, third edition. London: Oxford University Press.

Leonard, T. (1975). "Bayesian Estimation Methods for Two-Way Contingency Tables," *Journal of the Royal Statistical Society, Series B*, Vol. 37, 23-37.

Lindley, D.V. (1980). Comment on Box (1980).

Lindley, D.V. (1983). "The Role of Randomization in Inference," *Philosophy of Science Association 1982*, Vol. 2, 431-446 (Philosophy of Science Association).

Meeden, G. (1986). "Sufficiency and Partitions of the Class of All Possible Discrete Distributions," *The American Statistician*, Vol. 40, 42-44.

Pratt, J.W. (1965). "Bayesian Interpretation of Standard Inference Statements" (with discussion), *Journal of the Royal Statistical Society, Series B*, Vol. 27, 169-203.

Smith, A.F.M. (1986). "Some Bayesian Thoughts on Modelling and Model Choice," *The Statistician*, Vol. 35, 97-102.

West, M. (1986). "Bayesian Model Monitoring," *Journal of the Royal Statistical Society, Series B*, Vol. 48, 70-78.

Zellner, A. (1975). "Bayesian Analysis of Regression Error Terms," *Journal of the American Statistical Association*, Vol. 70, 138-144.

Zellner, A. (1980). Comment on Barnard (1980a).

Zellner, A. (1984). "Posterior Odds Ratios for Regression Hypotheses: General Considerations and Some Specific Results," in *Basic Issues in Econometrics*, Chicago: University of Chicago Press.

Bayesian and Likelihood Methods in Statistics and Econometrics
S. Geisser, J.S. Hodges, S.J. Press and A. Zellner (Editors)

CONGLOMERABILITY, COHERENCE AND COUNTABLE ADDIVITY

David A. Lane
University of Minnesota

Abstract

This paper offers answers to two foundational questions relating to the Law of Total Probability:

- Is the Law always available? That is, for a given partition of the sample space and unconditional probability distribution, is there necessarily a set of conditional probability distributions that yield the unconditional distribution when the Law is applied?
- When the Law is available, what are the consequences of violating it?

1. Introduction

In this paper, I will offer answers to two foundational questions relating to the Law of Total Probability. Statisticians and probabilists of all foundational persuasions use the Law as a fundamental tool for probability assessment. That is, to determine the probability of a proposition A, one often finds it convenient to introduce a set $\mathcal{H}$ of mutually exclusive, exhaustive propositions; then to evaluate a probability distribution μ over $\mathcal{H}$ and, for each h in $\mathcal{H}$, $P(A \mid h)$; and finally to invoke the Law to assess the probability for A by the formula

$$P(A) = \int P(A \mid h) d\mu(h). \tag{1}$$

The questions relating to the Law that I will consider are the following:

- Is the Law always available? That is, for a given $\mathcal{H}$ and unconditional probability P, is there necessarily a set of conditional probabilities $\{P(\cdot \mid h)\}$, such that P can be obtained according to equation (1)?
- When the Law is available, must it be respected? And if so, why? That is, what are the consequences of violating it?

TUDIES IN
)METRICS
TATISTICS

Editors

ARNOLD ZELLNER
JOSEPH B. KADANE

Volume 7

'ORK • OXFORD • TOKYO

BAYESIAN AND LIKELIHOOD METHODS IN STATISTICS AND ECONOMETRICS

Essays in Honor of George A. Barnard

Edited by

SEYMOUR GEISSER
University of Minnesota

JAMES S. HODGES
The RAND Corporation

S. JAMES PRESS
University of California, Riverside

ARNOLD ZELLNER
University of Chicago

1990

NORTH-HOLLAND – AMSTERDAM • NEW YORK • OXFORD • TOKYO

ELSEVIER SCIENCE PUBLISHERS B.V.
Sara Burgerhartstraat 25
P.O. Box 211, 1000 AE Amsterdam, The Netherlands

Distributors for the U.S.A. and Canada:
ELSEVIER SCIENCE PUBLISHING COMPANY, INC.
655 Avenue of the Americas
New York, N.Y. 10010, U.S.A.

Library of Congress Cataloging-in-Publication Data

Bayesian and likelihood methods in statistics and econometrics :
essays in honor of George A. Barnard / edited by Seymour Geisser ...
[et al.].
p. cm. -- (Studies in Bayesian econometrics and statistics ;
v. 7)
Includes bibliographical references.
ISBN 0-444-88376-2 (Elsevier Science)
1. Econometrics. 2. Mathematical statistics. 3. Probabilities.
4. Bayesian statistical decision theory. 5. Barnard, George A.
(George Alfred) I. Barnard, George A. (George Alfred)
II. Geisser, Seymour. III. Series.
HB139.B394 1990
330'.01'5195--dc20 89-28088
CIP

ISBN: 0 444 88376 2

Printed in The Netherlands

Acknowledgments

I. INTRODUCT

The Editors, Introd

II. A SURVEY
STATISTIC

D. V. Lindley, A S

Publications of Ge

III. FOUNDAT

A. P. Dempster, B

Bruce M. Hill, A T

James S. Hodges, C

David A. Lane, Co

IV. SELECTED

Siddhartha Chib, S.
Smoothing in a Tin

L. Denby, R. Gnan
An Analysis of Que
A Case Study in th

Stephen E. Fienber
What Goes Where?

Yoel Haitovsky, See
Hierarchical Structu

Hiroki Tsurumi, C
Information Estima