PROBABILITY
AND
MATHEMATICAL STATISTICS
Vol. 15 (1995), pp. 73–83

TESTING A PRECISE NULL HYPOTHESIS BY COMBINING EXPERIMENTS

BY

L. MARK BERLINER (COLUMBUS, OHIO)

Abstract. The problem of combining experimental results to test sharp null hypotheses is considered from a Bayesian viewpoint. Relying on results of Berger and Sellke [8], lower bounds on the posterior probability of the null are obtained based on classes of priors. It is suggested that plots of these lower bounds, as functions of the prior probability, provide a useful summary of results for appraising evidence. An example involving the combination of experiments concerning the value of aspirin usage for heart attack patients is presented. The discussion includes comparison with classical p-values associated with meta-analysis.

1. INTRODUCTION

The general problem of combining information from a variety of sources is both challenging and important. Types of information available are typically categorized as "Data," by which statisticians usually mean observational results of experiments, and "Judgment," which refers to expert scientific opinion concerning the phenomena under study as well as beliefs about the how the data is relevant to inference. A highly recommended review of statistical issues and methods for combining information is available in [16]. For the sake of brevity, I will not offer a review, nor complete references to the literature here, but rather refer the reader to [16].

It is natural to ask how and to what degree of success the recognized approaches to statistical analysis offer solutions for combining information. Consider the usual classification of statistical approaches into three basic schools: Fisherian, Neyman–Pearson(–Wald), and Bayesian. Consider first some general observations before turning to combining information. The notions that statistical inference should be objective and in some sense optimal (Fisher), and, in addition, have frequency based validity (Neyman), appear appealing. However, in the minds of many, the notion of "objectivity," in the sense of Fisher, is not typically achievable (see [5]). Also, assuming frequency properties to be the

primary basis for valid inference requires the creation of an often imaginary universe of replication and can lead to poor "conditional" properties. Such arguments are readily available in the literature. For general discussions and references, see [1].

To focus on combining information, consider a setting in which the results of k experiments are to be combined. Suppose that each experiment produces a natural summary statistic, say X_i. As a potential model, suppose that each $X_i \sim N(\theta, \sigma_i^2)$, and that inference concerning θ is desired. For example, suppose that a hypothesis test with null hypothesis H_0: $\theta = \theta_0$ versus H_a: $\theta \neq \theta_0$ is desired. For $k = 1$ the Fisherian view that the *p-value* provides an objective measure of the evidence in the data against the null is at least implementable as a procedure. To combine experiments in a natural fashion, we could impose an undeniably *subjective* assertion that the experiments are independent. Fisher (see [14]) indeed suggested a method for combining p-values directly for "a number of quite independent tests." This work, along with suggestions by K. Pearson and L. H. C. Tippett, all done in the 1930's (see [16]), has led to an active area of research in combining p-values and *meta-analysis.* However, concern over the interpretation of the p-value, especially in the context of taking an action of acceptance or rejection of the null, was a source of debate between Fisher and Neyman (see [14]). Also, the usefulness of p-values has been questioned and debated from a variety of viewpoints. See [1] for discussion.

A natural extension to the above model involves the consideration of what we might call *biases* for experiments. Namely, each experimental result is modeled as $X_i \sim N(\theta+\eta_i, \sigma_i^2)$, where η_i represents the bias in the i-th experiment. Though this seems a more plausible than assuming all the η_i's are zero in general, it poses foundational problems to both the objective Fisherian and the objective frequentist. The first point is how are the η_i's to be modeled. Of course, the familiar notions of "random" and "fixed" effects models can be brought to bear. Beyond introducing questions of objectivity, such models also introduce a variety of devils in the paradise of frequency. Specifically, where shall we envision our infinite sequence of results? If the frequency inference basis is to be valid as the number of experiments tends to infinity, crucial difficulties arise. At a foundational level, frequency validity will generally require the consideration of an infinite sequence of parameter values, as opposed to just an infinite sequence of experimental results based on a common parameter value. This distinction and a corresponding frequentist philosophy were considered by Neyman; see [2] for critical discussion and further results. Furthermore, some difficulties inherent in inference in the presence of an infinite number of nuisance parameters were originally documented by Neyman and Scott [17].

Turning to the Bayesian viewpoint, combining information appears to be a natural setting for Bayesian analysis. First, classical statistics is in principle

and by design unable to incorporate the accumulation of scientific thought and judgment. In particular, as suggested in [16], the notion of combining p-values does not appear to lead to satisfactory results in general. Second, as also suggested in [16], the use of hierarchical Bayesian models offers a way of formulating and analyzing a variety of stages associated with combining information. This suggestion is in accord with the view commonly expressed by Bayesians concerning the value of hierarchical analyses in general; see [12]. Third, combining experimental results demands subjective determination of the degree of relevance and "weight" individual experiments should have. In addition to [16], discussion of Bayesian combining of information may be found in [6], [10], and [15].

In this article I will only consider a form of a Bayesian hypothesis test using a hierarchical model to combine results from different experiments. (Though the emphasis in the article is on combining experiments, the analyses presented actually apply to any hierarchical model of the form given in Section 2, including traditional one-way random effects models.) The analysis given relies heavily on the work of Berger and Sellke [8]. The basic idea is to find ranges of Bayesian tests for a collection of prior models. In this sense the ideas and methods are drawn from *Robust Bayesian Analysis*; see [1], [3], and [4] for review. Berger and Sellke [8] concentrated on testing a sharp null hypothesis, and produced lower bounds on the posterior probability of the null. A primary goal was the demonstration of their contention that the p-value tends to overstate the evidence against the null. A counterpoint was given by Casella and Berger [9]; these two papers, along with the accompanying discussions, are highly recommended.

In the next section a simple testing problem involving the combination of experiments will be analyzed based on a "ranges of posterior probability" argument. Section 3 is devoted to a brief illustration of the results in a specific problem involving the combination of studies concerning the effect of aspirin usage on heart disease. The final section of the paper offers additional discussion.

2. FORMULATION AND RESULTS

I begin with a hierarchical Bayesian formulation of the specialized combining experiments problem discussed in Section 1. First, assume that the k experiments each produce a summary statistic X_i. Conditional on a collection of experimental parameters, $\vec{\theta} = (\theta_1, \ldots, \theta_k)^t$, assume that $\vec{X} = (X_1, \ldots, X_k)^t$ has the following multivariate normal distribution:

$$\vec{X} | \vec{\theta} \sim N(\vec{\theta}, \operatorname{diag}[\sigma_i^2]), \tag{1}$$

where the σ_i^2 are known. In typical applications, we would have that each experiment is based on a reasonably large sample size, and that as an approx-

imation the X_i and σ_i^2 are relevant point estimates and squared standard errors. To relate this model to that of Section 1, we simply write $\theta_i = \theta + \eta_i$.

The first stage of the hierarchical prior models a "similarity" or "exchangeability" assumption concerning the experiments. Namely, assume that, given hyperparameters μ and τ^2,

$$\vec{\theta}|\mu, \tau^2 \sim N(\mu\vec{1}, \tau^2 I). \tag{2}$$

The final stage of the modeling requires specification of prior models for μ and τ^2. First, I will assume that μ and τ^2 are independent. In a single prior Bayesian analysis, whose intent is to provide a test of the null and alternative hypotheses suggested in Section 1, a natural specification of the model is then

$$\mu \sim \pi_0\langle\theta_0\rangle + (1-\pi_0)g(\mu) \quad \text{and} \quad \tau^2 \sim h,$$

where π_0 is the prior probability of the null, $\langle\theta_0\rangle$ denotes a degenerate distribution assigning probability one to θ_0, and the distributions g and h are still to be chosen. Rather than specifying g and h, only classes of distributions for these inputs will be chosen. Before proceeding, note that the models in (1) and (2) reduce to

$$\vec{X}|\mu, \tau^2 \sim N(\mu\vec{1}, \operatorname{diag}[\sigma_i^2 + \tau^2]), \tag{3}$$

after integrating out $\vec{\theta}$. Let $f(\vec{x}|\mu, \tau^2)$ denote the corresponding probability density function.

2.1. Lower bounds on posterior probabilities. For each g and h, the posterior probability of the null hypothesis, denoted by $\pi_0(\vec{x})$, is given by

$$\pi_0(\vec{x}) = \frac{\pi_0 \int f(\vec{x}|\theta_0, \tau^2)h(\tau^2)d\tau^2}{\pi_0 \int f(\vec{x}|\theta_0, \tau^2)h(\tau^2)d\tau^2 + (1-\pi_0)\int\int f(\vec{x}|\mu, \tau^2)g(\mu)h(\tau^2)d\mu d\tau^2}.$$

It will be convenient to write $\pi_0(\vec{x})$ as

$$\pi_0(\vec{x}) = \left\{1 + b\frac{1-\pi_0}{\pi_0}\right\}^{-1}, \tag{4}$$

where

$$b = \frac{\int\int f(\vec{x}|\mu, \tau^2)g(\mu)h(\tau^2)d\mu d\tau^2}{\int f(\vec{x}|\theta_0, \tau^2)h(\tau^2)d\tau^2}. \tag{5}$$

Note that b^{-1} is typically called the *Bayes factor*.

To obtain lower bounds on the posterior probability of the null hypothesis, we obtain upper bounds on b. To do so first note that

$$\sup_{G,H} b = \sup_{H} \frac{\int \{\sup_G \int f(\vec{x}|\mu, \tau^2)g(\mu)d\mu\}h(\tau^2)d\tau^2}{\int f(\vec{x}|\theta_0, \tau^2)h(\tau^2)d\tau^2}.$$

(Berger and Mortera [7] also make use of this observation.) In the balance of this article, we make the conservative assumption that $H = \{\text{all distributions}\}$. This leads to a further reduction, based on a result of Sivaganesan and Berger ([18], Lemma A.1, p. 887):

$$\sup_{G,H} b = \sup_{\tau^2} \frac{\sup_G \int f(\vec{x}|\mu, \tau^2) g(\mu) d\mu}{f(\vec{x}|\theta_0, \tau^2)}.$$

Let

$$V = \left[\sum 1/(\sigma_i^2 + \tau^2)\right]^{-1}, \quad \hat{\mu} = \left[\sum x_i/(\sigma_i^2 + \tau^2)\right] V, \quad t = |\hat{\mu} - \theta_0|/\sqrt{V}.$$

Note that for fixed τ^2 and under the model in (3), $\hat{\mu}$ is the usual weighted least squares estimate of μ. Furthermore, the distribution of $\hat{\mu}$ is $N(\mu, V)$, and hence t is the *test statistic* used in testing the null hypothesis. With these definitions, simple manipulation yields the representation

$$f(\vec{x}|\mu, \tau^2) = f(\vec{x}|\theta_0, \tau^2) \exp\{0.5t^2\} \exp\{-0.5(\mu - \hat{\mu})^2/V\}.$$

Finally, combining these results, the desired bound on b, denoted by $b(G)$, is given by

$$b(G) = \sup_{\tau^2} \exp\{0.5t^2\} \left\{\sup_G \int \exp\{-0.5(\mu - \hat{\mu})^2/V\} g(\mu) d\mu\right\}. \tag{6}$$

The result from (6) is substituted into (4) to yield the final lower bound on the posterior probability of the null.

2.2. Examples. The first step in computing $b(G)$ defined by (6) involves the calculation of

$$\sup_G \int \exp\{-0.5(\mu - \hat{\mu})^2/V\} g(\mu) d\mu.$$

We next consider the same four specific choices for G as used in [8]. The calculations have been set-up so that their results can be applied directly. The presentation is quite brief; the reader may found the relevant discussion in [8], pp. 116–118.

Case 1. $G_A = \{\text{All Distributions}\}$. The key result for this case is due to Edwards et al. [11]. It is clear that

$$\sup_{G_A} \int \exp\{-0.5(\mu - \hat{\mu})^2/V\} g(\mu) d\mu$$

results when the prior assigns probability one to $\mu = \hat{\mu}$. Hence

$$b(G_A) = \sup_{\tau^2} \exp\{0.5t^2\}. \tag{7}$$

Case 2. $G_S = \{\text{All Symmetric (about } \theta_0) \text{ Distributions}\}$. Berger and Sellke [8] show that if $t \leqslant 1$, then $b(G_S)$ occurs when the distribution on μ assigns

probability one to the point θ_0. For $t > 1$, they approximate the optimizer by a two-point distribution assigning probabilities of 0.5 to each of the points $\hat{\mu}$ and $2\theta_0 - \hat{\mu}$. Combining these results, we find that

$$b(G_S) = \sup_{\tau^2} [1I_{(-\infty,1]}(t),\ 0.5\exp\{0.5t^2\}(1+\exp\{-2t^2\})I_{(1,\infty)}(t)], \tag{8}$$

where $I_M(t)$ denotes the usual indicator function on a set M.

Case 3. G_{US} = {All Unimodal, Symmetric (about θ_0) Distributions}. As in Case 2, if $t \leqslant 1$, then $b(G_S)$ occurs when the distribution on μ assigns probability one to the point θ_0. For $t > 1$, Berger and Sellke [8] show that the optimizing distribution is uniform on an interval $(\theta_0 - \kappa\sqrt{V},\ \theta_0 + \kappa\sqrt{V})$. The problem is then to find the optimal κ. Their analysis shows that the desired value of κ is the solution to

$$\kappa\{\varphi(\kappa+t)+\varphi(\kappa-t)\} = \Phi(\kappa-t)-\Phi\big(-(\kappa+t)\big),$$

where φ and Φ denote the density and distribution function, respectively, of an $N(0, 1)$ random variable. As a result we have

$$b(G_{US}) = \sup_{\tau^2} [1I_{(-\infty,1]}(t),\ \exp\{-0.5\kappa^2\}\cosh(\kappa t)I_{(1,\infty)}(t)]. \tag{9}$$

Case 4. G_{NOR} = {All Normal Distributions}. The underlying result in this case is again due to Edwards et al. [11]:

$$b(G_{NOR}) = \sup_{\tau^2}\left[1I_{(-\infty,1]}(t),\ \frac{\exp\{0.5t^2\}}{t\sqrt{e}}I_{(1,\infty)}(t)\right]. \tag{10}$$

To complete the analyses, we are to perform the indicated optimizations with respect to τ^2 in each of equations (7)–(10). This step is rather formidable since t is a nontrivial function of τ^2. Case 3 is particularly tedious since κ is the solution to a transcendental equation which must be solved for each τ^2. Hence, closed form general solutions will not be sought. However, numerical, and, in particular, graphical procedures are conceptually straightforward. The results of the next section were obtained in this fashion. Regarding Case 3, a simple recursive formula for κ, as given in formula (4.112) of [1], p. 234, was used.

3. EXAMPLE: ASPIRIN AND HEART DISEASE

To illustrate the lower bound calculations discussed in Section 2, I present analyses for an important example used often in [16]. The example involves the use of aspirin for heart patients. Quoting from Chapter 1 of [16]: "From 1970 to 1979 six major multicenter randomized trials of the use of aspirin and placebo by patients following a heart attack... were conducted in the United States and Europe." The following table summarizes results in the notation of Section 2.

The sample sizes for the first five studies ranged between 300 and 850 on each arm; the sample sizes on each arm of the last study were both over 2200. The X values represent the difference (placebo–aspirin) between mortality rates, and the σ values are the corresponding standard errors. Note that the last, and by far, largest study yielded results at odds, with respect to the indicated usefulness of aspirin, with the other five studies. I have also included classical p-values for each experiment, based on a normal approximation. The row of 2-sided p-values are based on the alternative hypotheses that $\theta_i \neq 0$, while the 1-sided p-values are for the alternatives $\theta_i > 0$ ("aspirin reduces mortality rate").

TABLE 1. Aspirin experiments data

Study	UK-1	CDPA	GAMS	UK-2	PARIS	AMIS
X	2.77	2.50	1.84	2.56	2.31	−1.15
σ	1.65	1.31	2.34	1.67	1.98	0.90
2-sided p-value	0.094	0.056	0.432	0.124	0.258	0.204
1-sided p-value	0.047	0.028	0.216	0.062	0.129	0.898

Consider combining these experiments to test the null hypothesis that the mean difference between placebo and aspirin effects is zero. Setting $\theta_0 = 0$, all numerical inputs to the analyses of Section 2 are specified.

The numerical analysis of the four cases considered in Section 1 leads to the following values of $b(G)$:

$$5.37,\ 2.69,\ 2.02,\ 1.78,$$

respectively. These results are of some interest themselves. Their inverses correspond to the lowest value, for each class considered, of the Bayes factor for the null versus alternative hypotheses. To inspect the combined effects of the Bayes factors and the prior probabilites, π_0, of the null, Figure 1 is presented. In this figure each of the resulting lower bounds of the posterior probabilities of the null are graphed as functions of π_0 for $0 < \pi_0 \leqslant 0.5$. This figure summarizes the analysis for the scientists inspection.

I now turn to a comparison of these results with more classical approaches. In [16] a variety of p-value analyses for this example is reviewed. First, the combined p-value approach due to Fisher involves computation of the quantity $-2\log(\prod_{i=1}^{k} P_i)$, where each P_i is the p-value from the i-th experiment. Under the null, this quantity has a χ^2 distribution with $2k$ degrees of freedom. The final overall p-value, denoted by P_F, for the aspirin example is 0.035. By comparison to Figure 1, P_F seems to greatly exaggerate the evidence against the null. Even for the G_{NOR} class, the π_0 would have to be quite small (less than 0.08) to match the evidence against H_0. (This comparison is based on the notion that such comparisons of p-values and posterior probabilities are

reasonable.) In this sense, the conclusion is in accord with those of Berger and Sellke [8]. Also, two natural competitors to Fisher's p-value due to Tippett, $P_T = 1-(1-P_{[1]})^k$, where $P_{[1]}$ is the smallest of the original k p-values, and Pearson, $P_P = \prod_{i=1}^{k}(1-P_i)$, lead to 0.29 and 0.25, respectively. These values do not seem particularly unreasonable, in the sense of comparison fo Figure 1, for this example.

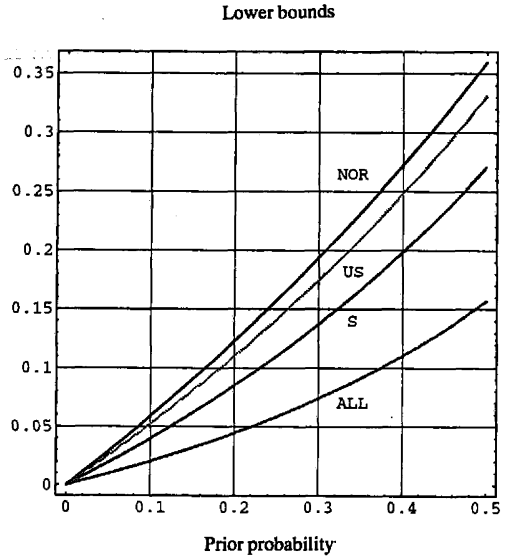

Fig. 1

For the aspirin data, it seems quite natural to consider 1-sided, rather than 2-sided, alternatives. Specifically, we might hope to find evidence that aspirin actually is beneficial in reducing mortality rates. This corresponds to an alternative hypothesis that $\mu > 0$ for our set-up. We can quickly provide a comparison for the assumption that $G = G_A$. For this case it is clear that the bounds reflected in Figure 1 remain unchanged. The corresponding p-values are given in Table 1. These result in $P_F = 0.01$, $P_T = 0.157$, and $P_P = 0.06$. Note that comparison to Figure 1 suggests that both Fisher's and Pearson's rules grossly overestimate the evidence against H_0.

4. DISCUSSION

4.1. Choosing hypotheses in testing. The most immediate question involves my choice of this example in illustrating approaches for testing a precise null. In the context of the aspirin example, it may seem natural to consider the "broad" null hypothesis H_0: $\mu \leqslant 0$. I elected a sharp null analysis for two reasons. First, lower bound arguments with large classes of priors and normal likelihoods tend to be uninteresting in the case of broad nulls [9].

Secondly, I believe that testing a broad null often has little justification in practical settings. For the aspirin example, as soon as we start to consider negative versus positive effects, we should take more care in the form of analysis taken. The losses for various actions (e.g., what happens if we recommend aspirin and it turns out to not just be ineffective, but in fact harmful) for various actions become paramount. Such concerns raise decision theoretic issues. Though Neyman–Pearson testing is a decision theoretic approach, my perhaps limited imagination cannot come up with many examples in which constant losses on the alternatives are reasonable. In a discussion of p-values, Jeffreys ([13], pp. 387–388) wrote "... the total area of the tail represents the probability, given the data, that the estimated difference has the wrong sign – provided that there is no question whether the difference is zero." He continued "... These are all problems of pure estimation. But their use as significance tests covers a looseness of statement of what question is being asked." We should then ask how much "looseness of statement" should be tolerated in a given problem. Counterarguments concerning the validity of "point nulls" are also often made [9]. However, Jeffreys again wrote "Some feeling of discomfort seems to attach itself to the assertion of the special value as *right*, since it may be slightly wrong but not sufficiently to be revealed by a test on the data available; but no significance test asserts it as certainly right. We are aiming at the best way of progress, not at the unattainable ideal of immediate certainty." Berger and Sellke [8], offer results along these lines.

4.2. The value of bounds. The analyses presented here are intended to offer a low level, and perhaps much needed, replacement for the use of p-values in meta-analysis; at the least, they can be used in assessing classical results. In defense of this suggestion, it can be argued that results such as those depicted in Figure 1 may be interpreted as "objective" summaries of the data. This version of objectivity is achieved by displaying the range of Bayesian answers that could occur from a wide variety of prior beliefs. But lower bounds cannot generally replace the need for consideration of prior information. Their chief weakness involves their one-directional interpretation: If, based on a large class of priors, the lower bound on the posterior probability of the null is small, we cannot generally claim to have strong evidence against the null. However, such bounds can be used as a supporting tool for Bayesian sensitivity analyses, in which results corresponding to a few plausible candidate priors are inspected. (A natural, and recommended when feasible, suggestion is to also compute upper bounds. For the formulation and the first three large classes considered in Section 2, upper bounds unfortunately offer no information. More refined classes of priors are required.) Perhaps the most important use of lower bounds is the identification of what prior information is crucial in judging the evidence available in data. I think many shun Bayesian methods when the prior matters, but are willing to use Bayesian approaches if the prior is not relevant. Perhaps

this view should be questioned. Indeed, von Mises reversed this logic. He argued that one must be a Bayesian when the prior matters *because* the prior matters; see [19], pp. 158–159.

I do not mean to suggest that the sort of analysis suggested here should be viewed as a replacement for a careful and complete Bayesian and robust Bayesian approach. On the other hand and especially in the context of meta-analysis, it may well be preferable for scientists to argue and debate over a result like that in Figure 1 than expend the same effort interpreting a variety of combined p-values. We Bayesians base our philosophy on the belief that a fully Bayesian approach to inference is the best and most efficient way to do induction based on data. However, optimal information processing may not actually be the goal in all scientific investigations. Indeed, this is a possible explanation for why Bayesian methods do not dominate practical statistics. Many scientists interpret the "scientific method" as requiring external verification of hypotheses, rather than arriving at the "best" answer based on all information available. (If this view has a role in science, I would hope that engineers, physicians, etc., take a different view when making decisions.) That is, a scientist may wish to deliberately avoid use of well-founded judgment in an attempt to defend such judgment. Much like a legal court room setting in which jurors are often sheltered from relevant information, the scientist may want to use the data to bear fair witness on the hypotheses. However, it should be emphasized that Bayesian statistics can be used in both "trial" and "optimal information processing" modes. (I. J. Good's ideas concerning Bayes/non-Bayes compromise are highly relevant; see [12].) Bayesian "trial" methods include both the notions of "noninformative priors" and ranges of posterior results. In doing such analyses, it is important that the evidence-seeking nature be remembered. In the same way that many non-Bayesians treat p-values as indicators of interest, rather than a basis for decision making, we can provide ranges arguments that are suggestive and useful. In the end, however, real decisions of real import usually require real Bayesian analyses.

REFERENCES

[1] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd edition, Springer, New York 1985.

[2] — *The frequentist viewpoint and conditioning*, in: *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, Vol. I, L. M. LeCam and R. A. Olshen (Eds.), Wadsworth, Belmont, California, 1985, pp. 15–44.

[3] — *Robust Bayesian inference: Sensitivity to the prior*, J. Statist. Plann. Inference 25 (1990), pp. 303–328.

[4] — *An overview of robust Bayesian analysis*, Technical Report # 93-53C, Department of Statistics, Purdue University, W. Lafayette, Indiana, 1993.

[5] — and D. Berry, *Statistical analysis and the illusion of objectivity*, Amer. Sci. 76 (1988), pp. 159–165.

[6] J. O. Berger and J. Mortera, *Interpreting the stars in precise hypothesis testing*, Internat. Statist. Rev. 59 (1991), pp. 337–353.

[7] — *Robust Bayesian hypothesis testing in the presence of nuisance parameters*, Technical Report # 91-69C, Department of Statistics, Purdue University, W. Lafayette, Indiana, 1992.

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

[9] G. Casella and R. L. Berger, *Reconciling Bayesian and frequentist evidence in the one-sided testing problem*, ibidem 82 (1987), pp. 106–111.

[10] W. H. DuMouchel and J. E. Harris, *Bayes methods for combining the results of cancer studies in human and other species (with discussion)*, ibidem 78 (1983), pp. 293–315.

[11] W. Edwards, H. Lindman and L. J. Savage, *Bayesian statistical inference for psychological research*, Psych. Rev. 70 (1963), pp. 193–242.

[12] I. J. Good, *The interface between statistics and philosophy of science*, Statist. Sci. 3 (1988), pp. 386–412.

[13] H. Jeffreys, *Theory of Probability*, 3rd edition, Clarendon Press, Oxford 1983.

[14] E. L. Lehmann, *The Fisher, Neyman–Pearson theories of testing hypotheses: One theory or two?* J. Amer. Statist. Assoc. 88 (1993), pp. 1242–1249.

[15] C. N. Morris and S. L. Normand, *Hierarchical models for combining information and meta-analysis*, in: *Bayesian Statistics 4*, J. Bernardo, J. Berger, A. Dawid, and A. F. M. Smith (Eds.), Oxford University Press, London 1992.

[16] *National Research Council. Combining Information: Statistical Issues and Opportunities for Research*, National Academy Press, Washington, D. C., 1992.

[17] J. Neyman and E. L. Scott, *Consistent estimates based on partially consistent observations*, Econometrica 16 (1948), pp. 1–32.

[18] S. Sivaganesan and J. O. Berger, *Ranges of posterior measures for priors with unimodal contaminations*, Ann. Statist. 17 (1989), pp. 868–889.

[19] R. von Mises, *Probability, Statistics, and Truth*, Dover, New York 1957.

Ohio State University
Columbus, Ohio 43210, U.S.A.

Received on 15.2.1994