ESTIMATION OF ACCURACY IN TESTING

BU-1027-MA

March 1990
Revised October 1990

Jiunn Tzon Hwang[1]
George Casella[2]
Cornell University

Christian Robert[3]
Université Paris VI

Martin T. Wells[4]
Roger H. Farrell
Cornell University

Key Words and Phrases: Hypothesis Testing, Decision Theory, p-values.

AMS 1980 Subject Classification: 62F99, 6CC99, 62A99

[1] Research supported by National Science Foundation Grant No. DMS88-09016

[2] Research supported by National Science Foundation Grant No. DMS89-0039

[3] This research was performed while Professor Robert was Visiting Scientist at the Mathematical Sciences Institute, Cornell University.

[4] Research supported by the Mathematical Sciences Institute, Cornell University.

Summary

The problem of hypothesis testing is approached as an estimation problem rather than a 0–1 decision problem, using a loss function to evaluate estimation rules. The theory developed is quite general, and contains standard (Neyman-Pearson) testing as a special case. Viewing hypothesis testing as estimation allows for formal evaluation of data-dependent frequentist measures of evidence. Particular attention is paid to the viability of the p-value as a measure of evidence, and some interesting conclusions, concerning admissibility in different testing problems, are reached.

1. Introduction

Approaches to hypothesis testing have usually treated the problem of testing as one of decision-making rather than estimation. More precisely, a formal hypothesis test will result in a conclusion as to whether a hypothesis is true, and not provide a measure of evidence to associate with that conclusion. In this paper we consider hypothesis testing as an estimation problem within a decision-theoretic framework, and are able to arrive at some interesting conclusions. In particular, reasonable loss functions result in decision rules that can be regarded as measures of evidence and, under these loss functions, some interesting properties of p-values emerge.

1.1 Standard Approaches

Classical hypothesis testing is built around the Neyman-Pearson Lemma (Lehmann, 1986) and results in decision rules that are 0–1 rules (except for randomized tests). These formal tests, although optimal in a strict frequentist sense, have been criticized from many different directions. Firstly, there have been many Bayesian criticisms (e.g., DeGroot, 1973; Dickey, 1977; Berger, 1985) which point out the drawbacks of the stringent conclusion of the Neyman-Pearson approach. Namely, the experimenter is locked into a two-point action space. Secondly, the assessment of accuracy of the test is typically a pre-data assessment, most often the *size* of the test. This estimate can be quite unreasonable when view post-data, a criticism which has also been leveled at Neyman-Pearson theory by conditionalists (Kiefer, 1977; Robinson, 1979). Alternatives considered by Kiefer include using p-values as an assessment of the likelihood of the null hypothesis. These ideas are in the direction of those proposed here, that the hypothesis test should result in a post-data assessment of evidence. (In fairness to Neyman-Pearson theory, measures of size and power were proposed as pre-data operating characteristics, not post-data assessment of accuracy, of a testing procedure.)

Perhaps the most serious criticism of Neyman-Pearson testing arises from the actions of practitioners. That is, formal Neyman-Pearson theory is not widely used in practice. Subject matter journals are flooded with p-values, but not with the outcomes of an α-level test. Furthermore, the p-value is implicitly used as a measure of evidence for a hypothesis. One of the reasons for undertaking the research presented here was to answer whether there are reasonable scenarios in testing for which the p-value is a reasonable answer. The fact that it is used extensively by experimenters is given; we, as statisticians, should decide whether the p-value has acceptable properties.

1.2 Criticisms of p-values

Most criticisms leveled at p-values have come from the Bayesian school, although there have been others. Even though p-values can be similar to Bayesian posterior probabilities, there are many seeming defects to criticize. Since the p-value is, in many cases, of the form $p(x) = P\Big(T(X) > T(x)\Big)$, where $T(x)$ is the observed value of the random variable $T(X)$, there is the problem of averaging over unlikely sample values (which have not occurred). Moreover, this is in violation of the likelihood principle (Berger and Wolpert, 1984), which states that inference must be based only on the observed data.

Even though p-values can fall within the range of Bayes solutions (Casella and Berger, 1987), they are fundamentally different. This, in itself, is not a cause for concern, as good frequentist and Bayesian procedures may be different, but there have been many criticisms involving paradoxes (e.g., Lindley, 1957; Berger and Sellke, 1987; Berger and Delampady, 1988). These paradoxes are all based on the fact that, at the tails, the p-value may be much smaller than Bayesian posterior probabilities in the two-sided testing problem. In the one-sided problem, however, this paradox does not appear (Casella and Berger, 1987), as the p-value is a limit of Bayes rules (see also Schaarfsma, et al., 1989). This observation may seem to be at odds with the previous paragraph, where we noted that the p-value violates the likelihood principle, something not done by a Bayes rule. The agreement of p-values and Bayesian posterior probabilities, however, is a mathematical identity specifying the agreement of two different integrals. Foundationally, the calculations are different.

The different behavior of the p-value in the one-sided and two-sided problem is one reason for the present investigation. This different behavior suggests that the formulation of the problems themselves may be to blame. For example, difficulties arise in the Bayesian formulation of the two-sided point null testing problem, or the classical two-sided composite null problem. A decision-theoretic formulation of testing may answer our concerns in these cases, and possibly clear doubts about p-values.

Other than Bayesian criticisms, problems with p-values can arise within classical statistics. (A strict Neyman-Pearson frequentist despises p-values with even more fervor than a Bayesian, as p-values have no real roots in frequentist theory. However, through their widespread use they are closely associated with classical, rather than Bayesian, statistics. It is thus the job of the frequentist to deal conclusively with p-values.) In some cases, p-values may be difficult to define (see the binomial example in Berger and Delampady, 1988, p.324), but if the p-value is defined in a straightforward way, it is usually from some Neyman-Pearson optimal test. The problem here is that many users implicitly and wrongly assume that any optimality derived from the Neyman-Pearson Lemma can be transferred to data-

dependent measures of accuracy. Although there has been some investigations about the behavior of p-values using loss functions (Gutmann, 1984; Schaarfsma, et al., 1989; Thompson, 1989), there has not been any systematic evaluation of post-data frequentist measures. There is a need for such an evaluation, and decision theory provides a natural mechanism for this task.

1.3 A Decision-Theoretic Approach

In a hypothesis testing problem we observe a value x of a random vector X with density (for convenience) $f(x|\theta)$, and desire a conclusion about the hypotheses

$$H_0: \theta \in \Theta_0 \quad \text{versus} \quad H_1: \theta \in \Theta_0^c , \tag{1.1}$$

where Θ_0 is a specified subset of the parameter space Θ. (We will not directly deal with the case of H_1 of the form $H_1: \theta \in \Theta_1$, where $\Theta_1 \neq \Theta_0^c$, although many results can be extended to this case.) We view our task as that of estimating the viability of the set specified by H_0, that is, of estimating the function $I_{\Theta_0}(\theta)$ (where $I_A(\cdot)$ denotes the indicator of a set A). The performance of a decision rule, $\phi(x)$, is evaluated with respect to a loss function

$$L(\theta,\phi) = d\Big(I_{\Theta_0}(\theta) - \phi\Big) , \tag{1.2}$$

where the function d(t) is minimum at $t = 0$, nondecreasing for $t > 0$, and nonincreasing for $t < 0$.

An important point to note is that we are considering this problem as one of estimation, not of deciding between H_0 and H_1. Thus, we are making an assessment of H_0, rather than drawing a conclusion about H_0. To assess H_0, we try to estimate $I_{\Theta_0}(\theta)$ with $\phi(x)$, where we consider the parameter $I_{\Theta_0}(\theta)$ to measure the accuracy of the test (hence the title of the paper). The rule $\phi(x)$ has the interpretation that large values support H_0 and small values support H_1, much like a p-value or a posterior probability of H_0, and thus $\phi(x)$ can be used by an experimenter in a similar way. Note however, that $\phi(x)$ does not measure "evidence" in a formal sense, as that can only be done through the likelihood ratio (Birnbaum, 1962; Royall, 1986). Thus, we make the important distinction of referring to $\phi(x)$ as a measure of accuracy, not evidence.

Although (1.1) and (1.2) define the general problem of estimation in testing hypotheses, we will only consider some special cases in what follows, using losses of the form

$$L_k(\theta,\phi) = \Big|I_{\Theta_0}(\theta) - \phi(X)\Big|^k, \quad k = 1,2, \tag{1.3}$$

with associated risk functions

$$R_k(\theta,\phi) = E_\theta \left| I_{\Theta_0}(\theta) - \phi(X) \right|^k , \qquad k = 1,2. \tag{1.4}$$

Note that standard Neyman-Pearson type results may be viewed as decision-theoretic results using a loss of the form (1.3) with k = 1. In particular, the Bayes rules with respect to (1.3), with k=1, are Neyman-Pearson-type solutions. (Be mindful that our estimators $\phi(x)$ estimate $I_{\Theta_0}(\theta)$, and are not a rejection probability. Thus, a Neyman-Pearson *critical function* would be equivalent to $1-\phi(x)$.)

Although there is a technical connection between the decision-theoretic approach with absolute error loss and Neyman-Pearson theory, the answers are not the same. In Neyman-Pearson theory the goal is to maximize power for a fixed α-level, while here the goal is to estimate $I_{\Theta_0}(\theta)$ using the loss (1.3), with no concern for a pre-experimental α-level. An example of the difference is the Neyman-Pearson need to consider randomized tests in discrete distributions. The estimator $\phi(x)$ is only equivalent to a randomized test if k = 1 in (1.3). Furthermore, there is no correspondence between decision-theoretic testing/estimation and confidence set estimation unless an artificial α-level is reintroduced. (This is as it should be, as the two problems address different questions.)

In hypothesis testing we are assessing $\phi(x)$ as an estimator of $I_{\Theta_0}(\theta)$, while in set estimation we are concerned with the coverage of a set C(x). This can be expressed as assessing an estimator $\gamma(x)$ of $I_{C(x)}(\theta)$. Decision theoretic approaches to set estimation are the major concern of Casella, Hwang, and Robert (1989,1990), and Bayesian solutions are treated by Berger (1985a,b; 1986). Other papers dealing with estimation of accuracy in set estimation include Robinson (1979a,b), Brown and Hwang (1989), Hwang and Brown (1989), George and Casella (1989), Lu and Berger (1989), and Robert and Casella (1990).

1.4 Summary of Results

The two hypothesis testing problems we will be concerned with are the one-sided testing problem

$$H_0: \theta \le \theta_0 \qquad \text{versus} \qquad H_1: \theta > \theta_0 , \tag{1.5}$$

where θ_0 is specified, and the two-sided testing problem

$$H_0: \theta \in [\theta_0,\theta_1] \qquad \text{versus} \qquad H_1: \theta \in [\theta_0,\theta_1]^c , \tag{1.6}$$

where θ_0 and θ_1 are specified. In either case we observe X = x, where X is a random vector with density $f(x|\theta)$, and we base our inference on the statistic T(X) with density $f_T(t\,|\,\theta)$.

In some cases, particularly in the two-sided testing problem, there are difficulties in defining a p-value. To eliminate these difficulties, we follow Lehmann (1986) and define it as follows. If R_α is the rejection region of an α-level test (most often UMPU) on which the p-value, p(x), is to be based, we define

$$p(x) = \inf\{\alpha : x \in R_\alpha\}, \tag{1.7}$$

which eliminates ambiguities (as long as R_α is specified for each α). This also implicitly defines the p-value in terms of the random variable T(X).

In Section 2 we examine the loss functions of (1.3) in more detail. We argue that absolute error loss (on which Neyman-Pearson testing is based) may not be the most reasonable loss function, and show that squared error loss emerges as an attractive alternative. Throughout the remainder of the paper we concentrate on squared error loss. In Section 3 we develop the decision theory using squared error loss. We provide an example then investigate minimaxity and admissibility, and are able to characterize the admissible rules in both one-sided and two-sided testing. Application of these results is in Section 4, which also contains a rather startling set of conclusions about p-values. Under certain assumptions, the p-value is admissible in the one-sided problem and inadmissible in the two-sided problem. However, it cannot be uniformly dominated by a proper Bayes rule in the two-sided problem. Section 5 contains a discussion, and there is an Appendix containing the proofs of the theorems in Section 3.

2. Consideration of Loss Functions

For the hypothesis testing problem (1.1), we now investigate reasonable forms for a loss function $L(\theta,\phi)$, to assess the worth of the estimator $\gamma(x)$ of $\mathbf{I}_{\Theta_0}(\theta)$. Since our parameter of interest has only two values, the loss function is of the form

$$L(\theta,\phi(x)) = \begin{cases} L(1,\phi(x)) & \text{if } \theta \in \Theta_0 \\ L(0,\phi(x)) & \text{if } \theta \notin \Theta_0 \end{cases}. \tag{2.1}$$

A minimal property for a loss function to have is that it be *proper* (Lindley, 1985). A proper loss function is one for which a Bayesian's best strategy is to tell the truth. (Whether one is a Bayesian, such a property is reasonable.) Thus, consider a prior distribution of $\pi(\theta)$ on Θ. The posterior expected loss, of the loss function (2.1), given X=x, is

$$\begin{aligned} E(L(\theta,\phi(x))|X=x) &= \int_\Theta L(\theta,\phi(x))\, \pi(\theta|x)d\theta. \\ &= L(1,\phi(x))\, P(\theta\in\Theta_0|x) + L(0,\phi(x))\, P(\theta\in\Theta_0{}^c|x), \end{aligned} \tag{2.2}$$

where $\pi(\theta|x) = f(x|\theta)\pi(\theta)/\int_\Theta f(x|\theta)\pi(\theta)d\theta$ is the posterior distribution, and

$$P(\theta\in\Theta_0|x) = \int_{\Theta_0} \pi(\theta|x)d\theta. \tag{2.3}$$

To say that the Bayesian's best strategy is to tell the truth is to say that the best estimator of $I_{\Theta_0}(\theta)$ is the Bayesian's best assessment of the probability of its occurrence. Thus, $L(\theta,\phi(x))$ is proper if

$$\min_{\phi(x)} E(L(\theta,\phi(x))|X=x) = E\Big(L(\theta,P(\theta\in\Theta_0|x))|X=x\Big). \tag{2.4}$$

Many common loss functions are proper, perhaps the most notable being squared error loss, corresponding to k=2 in (1.3), that is

$$L_2(\theta,\phi) = (I_{\Theta_0}(\theta) - \phi(x))^2. \tag{2.5}$$

Another, less common, proper loss is logarithmic loss, given by

$$L(\theta,\phi) = \log\Big|I_{\Theta_0}(\theta) + \phi(x)\text{-}1\Big|. \tag{2.6}$$

This loss also has the interesting property of yielding an infinite penalty if $\phi(x)$ is as wrong as possible. Surprisingly, absolute error loss, corresponding to k=1 in (1.3), or

$$L_1(\theta,\phi) = \Big|I_\Theta(\theta) - \phi(x)\Big| \tag{2.7}$$

is not a proper loss. Thus, if consideration is restricted to proper losses, absolute error loss, which corresponds to classical Neyman-Pearson theory, would be eliminated.

The loss L_1 thus suffers from a foundational view, but its shortcomings have been known (perhaps informally) to many. For example, there is risk equivalence between (dreaded) randomized tests and estimators of $I_{\Theta_0}(\theta)$. This fact partially explains why L_1 leads to 0–1 Bayes solutions. This equivalence is easy to see if we write the risk of the decision rule $\phi(x)$ as

$$R(\theta,\phi) = \int_{-\infty}^{\infty} \Big|I_{\Theta_0}(\theta) - \phi(x)\Big|f(x|\theta)dx = I_{\Theta_0^c}(\theta)\int_{-\infty}^{\infty} \phi(x)f(x|\theta)dx + I_{\Theta_0}(\theta)\int_{-\infty}^{\infty} \Big(1-\phi(x)\Big)f(x|\theta)dx$$

which is the risk of the randomized test $\phi(x)$ (or, in Neyman-Pearson terms, the risk of the critical function $1\text{-}\phi(X)$ under 0-1 loss). It is also possible, under suitable regularity conditions, to establish a converse. That is, the loss L_1 is the only loss under which there is a direct correspondence between estimators of $I_{\Theta_0}(\theta)$ and randomized tests.

The fact that the loss L_1 is so closely related to Neyman-Pearson 0–1 loss leads to estimators that will not be smooth and, as such, may have problems (especially conditional ones). For example, suppose we have one observation from a $n(\theta,1)$ density. A Bayes rule is $\phi^\pi(x) = I_{[0,c]}(|x|)$, which corresponds to the Neyman-Pearson UMPU test. The problem with this rule is that the same inference is made whether $x = 0$ or $|x| = c$.

If we turn to a straightforward decision-theoretic evaluation, similar answers would be found. Consider the following theorem, which is easy to establish.

Theorem 2.1: *a. The decision-theoretic Bayes rule with respect to the loss* L_1 *minimizes* $E^{\pi}(L_1(\theta,\phi)|x)$, *and is given by*

$$\phi_1^{\pi}(x) = \begin{cases} 0 & \text{if } P(\theta \in \Theta_0|x) < \frac{1}{2} \\ 1 & \text{otherwise} \end{cases} . \tag{2.8}$$

b. The decision-theoretic Bayes rule with respect to the loss L_2 *minimizes* $E^{\pi}(L_2(\theta,\phi)|x)$, *and is given by*

$$\phi_2^{\pi}(x) = P(\theta \in \Theta_0|x). \tag{2.9}$$

We therefore see that the absolute error loss, L_1, leads to a 0-1 Bayes solution, and can yield the usual Neyman-Pearson test in some cases (using point-mass priors). For the smoother loss L_2 we get a smoother Bayes rule, which is, of course, the Bayesian estimate of the indicator function $I_{\Theta_0}(\theta)$. As we shall see, in some cases the p-value is a limit of rules of the form (2.9), an impossibility in other cases.

The decision-theoretic derivation leads to the same place that the proper loss argument led. If we consider choosing between L_1 and L_2, the fact that for this loss function the Bayes rules are posterior probabilities is overwhelming. Since our goal is to estimate a probability, it is reassuring that the class of Bayes rules are exactly what we want. This observation is also of interest to non-Bayesians, since the class of Bayes rules is a subset of all admissible rules. Thus the loss L_2 not only provides a smooth alternative to L_1, it provides an alternative that produces sensible rules.

Whether we argue based on decision theory or proper loss functions, L_2 emerges as an extremely reasonable alternative to L_1. Since classical testing theory is equivalent to decision theory based on L_1, examination of decision theory based on L_2 is in order.

Of course, most of our arguments for preferring L_2 loss over L_1 loss could just as well support the use of any proper loss over L_1. (Any proper loss will result in the Bayes estimator of $I_{\Theta_0}(\theta)$ being $P(\theta \in \Theta_0|x)$.) Seen in this light, it might be argued that we should investigate the decision theory of other proper losses.

There are a number of reasons for not doing this. Firstly, the fact that all proper losses lead to the same Bayes estimator will result in similar decision-theoretic answers. Secondly, Hwang and Pemantle (1990) have found that L_2 plays a special role among proper losses. In investigating admissibility for a class of proper losses, they found that admissibility with respect to L_2 implied admissibility in the entire class.

Thirdly, there is a correspondence between admissibility with respect to L_2 and the nonexistence of relevant betting procedures, as described by Robinson (1979a). This means that admissibility with respect to L_2 will guarantee acceptable conditional performance.

Combining all of the arguments in this section, we arrive at two conclusions. First, the loss L_1 has inherent problems, and thus alternative losses should be considered. Second, among reasonable alternative losses, L_2 emerges as an eminent choice. Thus, for the remainder of this paper, we concentrate on decision-theoretic hypothesis testing using L_2 loss.

3. Decision Theoretic Results.

Under the loss L_2, we now examine some decision-theoretic consequences. To get a better idea of the situation, we first look at an example showing the behavior of some common rules. We then examine the criterion of minimaxity which, surprisingly, turns out to be a dead end. Then, admissibility is considered, and we are able to describe complete classes for both the one-sided and two-sided testing problem.

3.1 An Example.

To illustrate the risk behavior of some typical rules, we consider the simple situation of testing H_0: $\theta \leq 0$ versus H_1: $\theta > 0$, based on one observation, x, from a normal distribution with mean θ and variance 1. Two obvious estimators of $I(\Theta \leq 0)$ are the p-value, $P(Z>x)$, where Z is a standard normal random variable, and the Neyman-Pearson rule $\phi_c(x)=I(x<c)$, where c is a constant chosen according to the size of the test.

The risk of these rules is shown in Figure 1 along with the risk of two proper Bayes rules, using a $n(0,\tau^2)$ prior, and the minimax rule $\phi_0=\frac{1}{2}$. The Bayes rules dominate the p-value for θ near zero, since the Bayes estimator, $\phi_\tau(x)$, is

$$\phi_\tau(x) = P\left(Z>\left(\frac{\tau^2}{\tau^2+1}\right)^{\frac{1}{2}}x\right) > P(Z>x) = \text{p-value}.$$

However, as θ moves away from zero, the p-value becomes dominant. (For τ bigger than 1, the risk of the Bayes rule is extremely close to that of the p-value.) The risk of the Neyman-Pearson rules, however, is quite high, and is dominated by the p-value. (A more complete comparison of p-values versus Neyman-Pearson rules is given in Casella and Wells, 1990.) Finally, the risk of the constant risk minimax estimator $\phi_0=\frac{1}{2}$ is shown. We see that this rule is easily dominated for moderate values of θ, but performs well for θ near H_0. The next section shows that ϕ_0 is admissible.

3.2 Minimaxity

Deriving a minimax rule is similar under either L_2 or L_1 loss, so we state the result in one theorem.

Theorem 3.1: *For the hypothesis testing problem of* (1.1) *with density* $f(x|\theta)$ *continuous in* θ, *and loss* $L_k(\theta,\phi)$ *of* (1.3), *suppose that* Θ_0 *and* Θ_1 *have a common limit point. If* $0 \leq \phi \leq 1$ *is any decision rule, then*

$$\max_{\theta} E_\theta L_k(\theta,\phi) \geq \max_{\theta} E_\theta L_k(\theta,\phi_0) ,$$

where $\phi_0(x) = \frac{1}{2}$. *If* $k>1$ *then* ϕ_0 *is unique minimax, hence admissible.*

Proof: The proof uses standard applications of the Bounded Convergence Theorem and Liaponov's inequality, and is valid for all $k \geq 1$. The uniqueness of ϕ_0 follows from the strict convexity of the loss if $k>1$.

It is interesting to note that the p-value is a minimax rule under L_1. Thus, although minimaxity does not prove an interesting property for L_2, it does provide an optimality property for the p-value under L_1.

3.3 Complete Class Theorems Under L_2 Loss.

We next characterize complete classes of decision rules for both the one-sided (1.5) and two-sided (1.6) testing problem. The proofs of the main theorems become quite technical, and are placed in an appendix.

For the complete class theorems we only consider the exponential family. We observe X=x, where X has a density in the one-parameter exponential family. Since estimators which are functions of the sufficient statistic T(x) are complete class, we confine attention to density functions defined on R, the range of T,

$$f(t|\theta) = e^{\theta t - \Psi(\theta)}, \ \theta \in \text{interior of } \Theta, \tag{3.1}$$

where $\log \Psi(\theta) = \int_R e^{\theta t} f(t|\theta) d\mu(t)$. Because the results are valid in discrete as well as continuous problems, the integrating measure has been left vague.

The rules in the complete class are essentially generalized estimators, after modification of the parameter space (see the Appendix) and allowance for *truncation*, which we now define. In the one-sided problem (1.5) an interval $[t_1, t_2]$ is a *truncation set* for the estimator ϕ if $t < t_1$ implies $\phi(t) = 1$ and $t > t_2$ implies $\phi(t) = 0$. In the two-sided problem (1.6) an interval $[t_1, t_2]$ is a truncation set for the estimator ϕ if $t \in [t_1, t_2]^c$ implies $\phi(t) = 0$. (The idea of a truncation set originated in Farrell (1968).)

Theorem 3.2: *In the two-sided problem* (1.6) *the estimator* ϕ *with truncation set* $[t_1, t_2]$ *is admissible if there exists a probability measure* π_0 *supported on* $[\theta_0, \theta_1]$ *and a* σ*-finite measure* π_1 *supported on* $(-\infty, \theta_0] \cup [\theta_1, \infty)$ *such that, for almost all* $t_1 < t < t_2$,

$$\int f(t|\theta)\pi_1(d\theta) < \infty \tag{3.2}$$

and

$$\phi(t) = \frac{\int f(t|\theta)\pi_0(d\theta)}{\int f(t|\theta)\pi_0(d\theta) + \int f(t|\theta)\pi_1(d\theta)} . \tag{3.3}$$

Conversely, if ϕ *is admissible then there exist a truncation set* $[t_1, t_2]$, *a probability measure* π_0 *supported on* $[\theta_0, \theta_1]$ *and a* σ*-finite measure* π_1 *supported on* $(-\infty, \theta_0] \cup [\theta_1, \infty)$ *such that* (3.2) *and* (3.3) *hold for* $t \in (t_1, t_2)$.

Theorem 3.3: *In the one-sided problem* (1.5), *let* ϕ *be an admissible estimator under* L_2*-loss. Then there is a nonincreasing function* ϕ' *equivalent to* ϕ. *Assume without loss of generality that* ϕ *is nonincreasing and that* $\mathcal{C} = [t_1', t_2']$ *is a truncation set for* ϕ *such that if* $t_1' < t < t_2'$ *then* $0 < \phi(t) < 1$. *Let* $t_1' < t_0 < t_2'$. *There exist* σ*-finite measures* π_0 *on* $(-\infty, \theta_0]$ *and* π_1 *on* $[\theta_0, \infty)$ *such that*

$$1 = \int e^{t_0\theta - \Psi(\theta)}[\pi_0(d\theta) + \pi_1(d\theta)] , \tag{3.4}$$

and ϕ *is given by* (3.3) *for* $t \in (t_1, t_2)$, *both integrals of* (3.3) *being finite.*

As mentioned before, the essential point of Theorems 3.2 and 3.3 is that the complete class is given by the generalized Bayes rules (almost). Thus, to establish admissibility, one would first check to see if a rule is generalized Bayes. In the next section we apply this strategy to the p-value.

4. Admissibility Considerations Under L_2 Loss

We now return to exploration of the behavior of the p-value, and find that under L_2 loss the results are quite interesting. There is a dichotomy occurring in the fate of the p-value, one that, perhaps, is reflected in the dissenting views of Berger and Sellke (1987) and Casella and Berger (1987a). In the one-sided testing problem the p-value is, in many cases, admissible against the loss L_2 of (2.5), showing that the p-value is a reasonable measure of accuracy, a notion that agrees with Casella and Berger (1987a).

In the two-sided case, however, the answers are a bit more involved, in that the p-value is inadmissible but difficult to dominate. We are able to show that the usual p-value is not in the complete class of Theorem 3.2 (the two-sided problem), demonstrating its inadmissibility. This fact is consonant with the results of Berger and Sellke (1987) and

Berger and Delampady (1987) concerning the failings of the p-value in the two-sided problem. However, there is an interesting occurrence in the two-sided point null normal case: Although the p-value is inadmissible, it cannot be dominated by any proper Bayes estimator.

4.1.1 Examples of Admissibility in the One-Sided Problem

We present a number of examples in which the p-value is generalized Bayes, hence admissible. This property probably carries over to other distributions, but in the following cases the admissibility of the p-value can be easily established.

Theorem 4.1: *For the one-sided hypothesis testing problem of* (1.5), *with loss function* L_2 *of* (2.5), *let* $X_1,\cdots,X_n$ *be iid* $n(\theta,1)$. *The p-value* $p(\overline{x}) = P_{\theta_0}(\overline{X} > \overline{x}) = 1 - \Phi\Big(\sqrt{n}(\overline{x} - \theta_0)\Big)$, *is admissible, where* $\overline{X}$ *is the mean of* $X_1,\cdots,X_n$ *with observed value* $\overline{x}$.

Proof: Using sufficiency, we can assume n = 1. Note that the p-value is generalized Bayes with respect to the Lebesgue measure prior (it is also a limit of Bayes rules against the sequence of $n(\theta_0,r)$ priors). Furthermore, the (generalized) Bayes risk of the p-value is finite. Therefore the p-value is admissible. □

We now establish the admissibility of the p-value for some discrete distributions using a similar method. We summarize these results in the following theorem.

Theorem 4.2: *For the one-sided hypothesis testing problem of* (1.5), *with loss function* L_2 *of* (2.5)

a. If $f(x|\theta)$ *is binomial* (n,θ), *the p-value* $p(x) = P_{\theta_0}(X \geq x) = \sum_{k=x}^{n} \binom{n}{k} \theta_0^k (1-\theta_0)^{n-x}$ *is admissible.*

b. If $f(x|\theta)$ *is Poisson*(θ), *the p-value* $p(x) = P_{\theta_0}(X \geq x) = \sum_{k=x}^{\infty} e^{-\theta_0} \theta_0^k / k!$ *is admissible.*

Proof: For (a), consider the generalized prior density $1/\theta$, which has corresponding generalized Bayes estimator p(x). The fact that this estimator has finite generalized Bayes risk follows from the fact that the estimator

$$\delta(x) = \begin{cases} 1 \text{ if } x=0 \\ 0 \text{ otherwise} \end{cases}$$

has finite generalized Bayes risk. Thus p(x) is admissible. Part (b) can be established similarly by again considering the generalized prior density $1/\theta$. □

We also note that in both the binomial and Poisson cases, generalization to an iid sample is immediate. Therefore, in a number of cases in the one-sided testing problem, the p-value is admissible as an estimator of $I_{(-\infty,\theta_0)}(\theta)$.

4.1.2 Admissibility in the Two-Sided Problem

The complete class theorem (Theorem 3.2) gives us a powerful tool for exploring admissibility of the p-value in the two-sided problem. The following theorem, which is a corollary of Theorem 3.2, allows us to reach some decisive conclusions about inadmissibility of the p-value.

Theorem 4.3: *For the hypothesis testing problem of* (1.6), *with loss function* L_2 *of* (2.5), *suppose the estimator* $\varphi(T(x)) > 0$ *is continuous, nonconstant, and, for some value* x_0, $\varphi(T(x_0)) = 1$. *Then* φ *is inadmissible.*

Proof: If φ were admissible, then almost surely (3.3) holds. Since $\varphi > 0$ for all x, $\int f(x|\theta)\pi(d\theta) < \infty$ for almost all x. Thus, both sides of (3.3) are continuous in x, and hence equal for all x in the support of X. Since $f(x|\theta) > 0$, and $\pi \geq \pi_0$, $\varphi(T(x_0)) = 1$ implies $\pi = \pi_0$ and $\varphi(T(x)) = 1$ for all x, a contradiction. □

The result of Theorem 4.3 now allows us to answer the question of the admissibility of the p-value.

Theorem 4.4: *For the hypothesis testing problem of* (1.6), *with loss function* L_2 *of* (2.5) *and* $T(x)$ *continuous, the p-value is inadmissible.*

Proof. The proof proceeds by showing that the p-value given in (1.7) takes the value 1. For the hypotheses of (1.6), the p-value is based on a UMPU test of the form

$$\phi(x) = \begin{cases} 0 & \text{if } T(x) < c_0 \text{ or } T(x) > c_1 \\ \\ 1 & \text{if } c_0 \leq T(x) \leq c_1 \end{cases}, \tag{4.1}$$

where $\phi(x)$ is the probability of accepting H_0. The constants c_0 and c_1 are functions of α, the level of the test, that is, $c_0 = c_0(\alpha)$ and $c_1 = c_1(\alpha)$.

We first deal with the case $\theta_0 \neq \theta_1$. By Lehmann(1986, page 135), $c_0(\alpha)$ and $c_1(\alpha)$ satisfy

$$P_{\theta_0}\big(c_0(\alpha) \leq T(X) \leq c_1(\alpha)\big) = P_{\theta_1}\big(c_0(\alpha) \leq T(X) \leq c_1(\alpha)\big) = 1\text{-}\alpha. \tag{4.2}$$

Define

$$c^* = \inf\big\{T(x):\ f(x|\theta_1) \geq f(x|\theta_0) \text{ and } x \text{ in the support of } X\big\}.$$

By the continuity if T(x), $c^*$is in the support of T(x). Also define

$$c' = \frac{\Psi(\theta_1) - \Psi(\theta_0)}{\theta_1 - \theta_0},$$

for $\Psi(\cdot)$ as in (3.1). Note that $f(x|\theta_1) \geq f(x|\theta_0)$ if and only if $T(x) \geq c'$, from which it

follows that $c^* \geq c'$.

We claim that for every α, $0 < \alpha < 1$, $c^* \in [c_0(\alpha), c(\alpha)]$. This then implies that $p(T(x)) = 1$ when $T(x) = c^*$, and thus, by Theorem 4.3, the p-value is inadmissible. To establish the claim, suppose to the contrary that $c^* \notin [c_0(\alpha), c(\alpha)]$ for some $\alpha > 0$, say $c^* < c_0(\alpha)$. Thus, for $T(x) \in [c_0(\alpha), c(\alpha)]$, $f(x|\theta_1) \geq f(x|\theta_0)$. From (4.2) it follows that

$$\mathbb{I}\Big(c_0(\alpha) \leq T(X) \leq c_1(\alpha)\Big)\Big(f(x|\theta_1) - f(x|\theta_0)\Big) = 0 \quad \text{a.s.},$$

which implies that whenever $c_0(\alpha) \leq T(X) \leq c_1(\alpha)$, $T(x) = c'$. However, $c' \leq c^* < c_0(\alpha)$, so $P_{\theta_0}\Big(c_0(\alpha) \leq T(X) \leq c_1(\alpha)\Big) = 0$ implying $\alpha = 0$ which is a contradiction. Thus the claim is established and the p-value is inadmissible if $\theta_0 \neq \theta_1$.

If $\theta_0 = \theta_1$, instead of $f(x|\theta_0)$ and $f(x|\theta_1)$, we consider

$$f(x|\theta^*) = \Big(E_{\theta_0} T(X)\Big) f(x|\theta_0)$$

and

$$f(x|\theta^{**}) = T(X) f(x|\theta_0).$$

Arguments similar to those above, along with (5) and (6) of Lehmann(1986, page 136), can be used to establish the inadmissibility of the p-value in this case. □

Remark: For the case of testing a point null hypothesis, where $\theta_0 = \theta_1$ and $c_0 = -c_1$, Theorem 4.4 immediately applies. In fact, it can be made to apply to a k-parameter exponential family.

Although Theorem 4.4 is negative in its assessment of p-values, we will see that it is, perhaps, not as negative as it might first appear. We now look at the special case of testing a point null hypothesis about a normal mean, and find that the p-value cannot be dominated by any proper Bayes procedure. Thus, even though the p-value is inadmissible for testing a point null hypothesis, it is quite difficult to exhibit a better estimator. As before, sufficiency allows us to consider the case of one observation.

Theorem 4.5: *For testing the hypothesis*

$$H_0: \theta = \theta_0 \quad \textit{vs.} \quad H_1: \theta \neq \theta_0 ,$$

based on one observation X from a $n(\theta,1)$ *density, and using loss function* L_2 *of* (2.5), *the p-value cannot be dominated by any Bayes rule.*

Proof: Assume, without loss of generality, that $\theta_0 = 0$. The Bayes rules for this problem are of the form

$$\phi^{\pi}(x) = \frac{\pi_0 f(x|0)}{\pi_0 f(x|0) + (1-\pi_0) \int f(x|\theta) g(\theta) \mu(d\theta)} ,$$

where $f(x|\theta)$ is a $n(\theta,1)$ density and $\int_{-\infty}^{\infty} g(\theta)\mu(d\theta) = 1$ (see Section 2), and the p-value is given by

(4.3) $$p(x) = P(|X| \geq x) = 2\Big(1 - \Phi(x)\Big) .$$

We consider three cases.

Case 1: $\pi_0 = 1$. In this case $\phi^\pi(x) = 1$. As $\theta \to \infty$, $R\Big(\theta, p(x)\Big) \to 0$ but $R\Big(\theta, \phi^\pi(x)\Big) = 1$ for $\theta \neq 0$, so $\phi^\pi(x)$ cannot dominate $p(x)$.

Case 2: $\pi_0 = 0$. In this case $\phi^\pi(x) = 0$ so $R\Big(0, \phi^\pi\Big) = 1 > R(0, p(x)) = \frac{1}{3}$ and $\phi^\pi(x)$ cannot dominate $p(x)$.

Case 3: $0 < \pi_0 < 1$. We will show that as $\theta \to \infty$, $R(\theta, p(x))$ becomes smaller than $R(\theta, \phi^\pi)$. First note that for sufficiently large $|x| > a > 0$, $\phi^\pi(x) > p(x)$. This follows from the fact that

(4.4) $$\phi^\pi(x) \geq \frac{\pi_0 f(x|\theta_0)}{\pi_0 f(x|\theta_0) + (1-\pi_0) f(x|\hat{\theta})} > p(x)$$

for sufficiently large $|x|$, where $\hat{\theta} = x$ is the maximum likelihood estimator of θ.

For $\theta \neq \theta_0$, the difference in risks is given by

(4.5) $$E_\theta\Big(\mathbb{I}_{\{\theta_0\}}(\theta) - \phi^\pi(X)\Big)^2 - E_\theta\Big(\mathbb{I}_{\{\theta_0\}}(\theta) - p(X)\Big)^2 = E_\theta\Big(\phi^\pi(X)^2 - p(X)^2\Big) ,$$

and from (4.4), by continuity, there exists an $\epsilon > 0$ such that $\phi^\pi(x)^2 - p(x)^2 > \epsilon$ for all $a < |x| < a+\epsilon$. Hence

(4.6) $$E_\theta\Big(\phi^\pi(X)^2 - p(X)^2\Big) \geq \epsilon P_\theta(a < |X| < a+\epsilon) - P_\theta(|X| < a) .$$

This lower bound is positive for large θ since

$$\frac{P_\theta(a < |X| < a+\epsilon)}{P_\theta(|X| < a)} \to \infty \text{ as } \theta \to \infty$$

by L'Hospital's rule. Therefore the difference in risks is strictly positive for large θ, and $\phi^\pi(x)$ cannot dominate $p(x)$. □

Thus, we are left with an interesting situation. We have an inadmissible rule, $p(x)$, that cannot be dominated by any obvious competitor based on a Bayes argument. Using a generalized Bayes estimator based on a complicated prior, Hwang and Pemantle (1990) constructed an estimator that dominates the p-value. That estimator was only constructed for that purpose, however, and will probably not gain widespread use in practice. Thus, the p-value will, no doubt, remain as an often used estimator of accuracy, and although inadmissible in the two-sided problem, may not be too bad.

5. Discussion

The formulation of hypothesis testing as a decision-theoretic estimation problem leads to results, that is, estimators, that are more satisfying than the conclusions from Neyman-Pearson theory. These estimators, which may be considered measures of evidence possess formal optimality properties. Viewing the testing problem as one of estimating an indicator of H_0, and separating it from the set estimation problem, leads to estimators that are more desirable in practice.

The failure of minimaxity to provide any interesting results for the loss L_k of (1.3), with $k > 1$, is surprising, and we are unsure how to interpret this. The fact that $\phi_0 = \frac{1}{2}$ is minimax was anticipated, but the fact that it is unique minimax was not. Therefore, we have to accept the fact that for strictly convex loss functions, if we use a data-dependent measure of evidence, we will sometimes do worse than the "no-data" rule $\phi_0 = \frac{1}{2}$. Minimaxity may prove to be a useful criterion, however, in any further decision-theoretic study using absolute error loss.

The dichotomy between the (rather straightforward) one-sided problem and the (more involved) two-sided problem is illustrated by the fate of the p-value. It is generally admissible in the one-sided case (being a limit of Bayes rules) but inadmissible in the two-sided case (not corresponding to any generalized Bayes rule). These conclusions are in line with, and partially explain, the opposing arguments of Berger and Sellke (1987) and Berger and Delampady (1988), who contended that the p-value is unreasonable in two-sided problems, and Casella and Berger (1987), who contended that the p-value can be reasonable in one-sided problems.

What is even more startling, however, is the inability of any Bayes rule to dominate the p-value in the two-sided point null problem. Unless we can find a practical dominating estimator, this gives the p-value a position enjoyed by few estimators (the positive-part James-Stein estimator comes to mind), an inadmissible estimator for which it is difficult to exhibit a dominating estimator. Thus, even in the two-sided case, the p-value could be a viable measure of evidence against H_0.

These conclusions about the p-value are tied to the use of squared error loss, the loss L_2 of (2.5). This may be a cause for criticism, for we are somewhat unsure of what the conclusions will be if other losses are used. However, this is a loss that results in the Bayes rules being given by the posterior probabilities, a perfectly natural situation shared by other proper losses. This leads us to believe that, for any proper loss, the results presented here would continue to be valid.

Acknowledgments. We thank L.D. Brown and Richard Royall for helpful discussions, A. Rukhin and S. Gutmann for showing us some references we missed, and R. Pemantle for pointing out a simpler proof of Theorem 4.1. We also thank the Associate Editor and referee, whose comments greatly improved this paper.

References

Berger, J. O. (1985a). *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York.

Berger, J. O. (1985b). The frequentist viewpoint and conditioning. In *Proceedings of the Berkeley Conference in Honor of J. Neyman and J. Kiefer* (L. LeCam and R. Olshen, eds.). Wadsworth: Monterey, CA, pp. 15-44.

Berger, J. O. (1986). Discussion of "Conditionally Acceptable Frequentist Solutions" by G. Casella. In *Statistical Decision Theory and Related Topics IV, Vol.* 1 (S.S. Gupta and J.O. Berger, eds.), pp. 85-90.

Berger J. O. and Wolpert, R. W. (1984). *The Likelihood Principle*, 2nd edition. IMS Monograph Series, Institute of Mathematical Statistics, Hayward, CA.

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

Berger, J. O. and Delampady, M. (1988). Testing precise hypotheses (with discussion). *Statistical Science* 2, 317-352.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *J. Amer. Statist. Assoc.* 57, 269-306.

Blyth, C R. (1951). On minimax statistical decision procedures and their admissibility. *Ann. Math. Statist.* 22, 22-42.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families.* IMS Monograph Series, Institute of Mathematical Statistics, Hayward, CA.

Brown, L. D. and Hwang, J. T. (1989). Admissibility of confidence estimators. Technical Report, Statistics Center, Cornell University, Ithaca, NY. To appear in the proceedings of the Taipai Symposium.

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

Casella, G. and Berger, R. L. (1987b). Comment on "Testing precise hypotheses by Berger and Delampady." *Statistical Science* 2, 344-347.

Casella, G., Hwang, J.T., and Robert, C. (1989). Loss functions for set estimation. Technical Report BU-999-M, Biometrics Unit, Cornell University, Ithaca, NY.

Casella, G., Hwang, J.T., and Robert C. (1990). A paradox in decision theoretic set estimation. Technical Report BU-1086-M, Biometrics Unit and Statistics Center, Cornell University, Ithaca, New York.

Casella, G. and Wells, M.T. (1990). Comparing p-values to Neyman-Pearson tests. Technical Report BU-1073-M, Biometrics Unit and Statistics Center, Cornell University,

Ithaca, New York.

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

Dickey, J. (1977). Is the tail area useful as an approximate Bayes factor? *J. Amer. Statist. Assoc.* 72, 138-142.

Farrell, R. H. (1968). Towards a theory of generalized Bayes tests. *Ann. Math. Statist.* 39, 1-28.

George, E. I. and Casella, G. (1989). Empirical Bayes confidence estimation. Technical Report BU-1062-M, Biometrics Unit, Cornell University, Ithaca, NY.

Gutmann, S. (1984). Loss functions for p-values and simultaneous inference. Technical Report No. 43, Statistics Center, Massachusetts Institute of Technology, Cambridge, MA.

Hwang, J. T. and Brown, L. D. (1989). The validity admissibility criterion for confidence estimators. Technical Report, Statistics Center, Cornell University, Ithaca, NY. To appear in the *Annals of Statistics.*

Hwang, J. T. and Pemantle, R. L. (1990). Evaluation of estimators of statistical significance under a class of loss functions. Technical Report, Statistics Center, Cornell University, Ithaca, New York.

Kiefer, J. (1977). Conditional confidence statements and confidence estimators (with discussion). *J. Amer. Statist. Assoc.* 72, 789-808.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, 2nd edition. New York: John Wiley.

Lindley, D. V. (1985). *Making Decisions*, Second Edition. New York: John Wiley.

Lindley, D. V. (1957). A statistical paradox. *Biometrika* 44, 187-192.

Lu, K. L. and Berger, J. O. (1989). Estimated confidence procedures for multivariate normal means. *J. Statist. Plann. Inf.* 23, 1-19.

Robert, C. and Casella, G. (1990). Improved confidence estimators for the usual multivariate normal confidence set. Technical Report BU-1041-M, Biometrics Unit, Cornell University, Ithaca, NY.

Robinson, G. K. (1979a). Conditional properties of statistical procedures. *Ann. Statist.* 7, 742-755.

Robinson, G. K. (1979b). Conditional properties of statistical procedures for location and scale families. *Ann. Statist.* 7, 756-771.

Royall, R. (1986). The effect of sample size on the meaning of significance tests. *Amer. Statist.* 40, 313-315.

Schaarfsma, W., Tobloom, J., and Van der Menlen, B. (1989). Discussing truth or falsity by computing a q-value. In *Statistical Data Analysis and Inference* (Y. Dodge, ed.). North Holland: Elsevier Science.

Stein, C. (1956). The admissibility of Hotelling's T^2 test. *Ann. Math. Statist.* 27, 616-623.

Thompson, P. M. (1989). Admissibility of p-value rules. Ph.D. Thesis, Department of Statistics, University of Illinois, Urbana, IL.

Appendix: Proofs of Theorems 3.2 and 3.3

We first establish some preliminary notation and lemmas.

Lemma A.1: *For L_2 loss, if ϕ' is as good as ϕ and ϕ has truncation set $[t_1, t_2]$ then $\phi(t) = \phi'(t)$, for $t < t_1$ or $t > t_2$.*

Proof: The standard argument for this result is to use Stein (1956). □

In order to prove the best possible result it is necessary to modify the parameter space. We discuss this for the two-sided problem. Let

$$I_0 = \{(0,\theta): \theta_0 \le \theta \le \theta_1] \ ,$$

and

$$I_1 = \{(1,\theta): \theta \le \theta_0 \quad \text{or} \quad \theta \ge \theta_1\} \ .$$

Define

$$R_1(\theta,\phi) = E_\theta\Big(\mathbb{1}_{[\theta_0,\theta_1]}(\theta) - \phi(X)\Big)^2 \ ,$$

$$R_2[(i,\theta),\phi] = E_\theta\Big(\mathbb{1}_{I_0}(0,\theta) - \phi(X)\Big)^2 \ , \quad (i,\theta) \in I_0 \ ,$$

and

$$R_2[(i,\theta),\phi] = E_\theta\Big(1 - \mathbb{1}_{I_1}(1,\theta) - \phi(X)\Big)^2 \ , \quad (i,\theta) \in I_1 \ .$$

Because $R_2(\cdot,\phi)$ can be obtained as a limit of $R_1(\cdot,\phi)$ the following lemma is easy to verify.

Lemma A.2: *ϕ is admissible for risk R_1 if and only if ϕ is admissible for risk R_2.*

Remark: In the two-sided problem with modified parameter space, when π_0 and π_1 are finite measures, the Bayes estimators are given by equation (3.3). In particular, the constant valued estimators $\phi(t) = \alpha$ are Bayes estimators in the modified problem, hence are admissible.

In the following proofs we write θ rather than (i,θ) and use (3.3).

Proof of Theorem 3.2: Assume $\phi(X)$ is any admissible rule. From Brown (1986, Proposition 4A.7 and Theorem 4A.12) there exists a sequence of finite priors (μ_{0j}, μ_{1j}) concentrated on finite subsets such that the Bayes estimators $\phi^{\mu_j}(t)$ converge to $\phi(t)$ almost surely. The special case $\phi \equiv 0$ is obvious and we assume $E_\theta\phi[T(X)] > 0$. By the dominated convergence theorem $\lim_{j\to\infty} E_\theta\phi^{\mu_j}[T(X)] = E_\theta\phi[T(X)]$ so that for all large j, $\phi^{\mu_j}(t) > 0$ with positive measure. Hence $\mu_{0j}([\theta_0, \theta_1]) > 0$ and by renormalization we assume $\mu_{0j}([\theta_0, \theta_1]) = 1$.

The convex set

$$\text{(A.1)} \qquad \mathcal{C}' = \left\{ t : \limsup_{j\to\infty} \int e^{t\theta-\Psi(\theta)} \mu_{1j}(d\theta) < \infty \right\}$$

is an interval. Clearly from (3.3) for almost all t not in $\mathcal{C}$, $\phi(t) = 0$. Let $\mathcal{C} = [t_1, t_2]$ be the closure of $\mathcal{C}'$ and use $\mathcal{C}$ as a truncation set.

From (3.3), for almost all $t_1 < t < t_2$,

$$0 < \lim_{j\to\infty} \phi^{\mu_j}(t) = \phi(t) .$$

If necessary by choice of subsequence we may assume μ_{0j} converges to a probability measure μ_0 weakly so that

$$\lim_{j\to\infty} \int f(t|\theta)\mu_{0j}(d\theta) = \int f(t|\theta)\mu_0(d\theta) .$$

It then follows for almost all $t_1 < t < t_2$ that $\lim_{j\to\infty} \int f(t|\theta)\mu_{1j}(d\theta)$ exists and is finite.

Thus, if $t_1 < t_2$, then standard arguments may be used to show there exists a limiting σ-finite measure μ_1 (and a subsequence if necessary), such that $\mu_{1j} \to \mu_1$ and if $t_1 < t < t_2$ then

$$\text{(A.2)} \qquad \lim_{j\to\infty} \int e^{\theta t-\Psi(\theta)} \mu_{ij}(d\theta) = \int e^{\theta t-\Psi(\theta)} \mu_i(d\theta) , \qquad i = 1,2 .$$

Thus for almost all $t_1 < t < t_2$, $\phi(t)$ can be expressed as in (3.3), establishing the second part of Theorem 3.2. The first part of Theorem 3.2 follows from Lemma A.2 and the uniqueness of the generalized Bayes estimator as minimizing the generalized Bayes risk. □

Proof of Theorem 3.3: In the one-sided problem of (1.5), for density functions (3.1), by modification of the functions b and Ψ, we can assume without loss that $\theta_0 = 0$ since we can write

$$b(x)\, e^{\theta T(x)-\Psi(\theta)} = \left(b(x)\, e^{\theta_0 T(x)}\right) e^{(\theta-\theta_0)T(x)-\Psi[(\theta-\theta_0)+\theta_0]} .$$

Assume $\phi(t)$ is admissible. Bayes rules are given for the modified parameter space by

$$\phi^{\pi}(t) = \frac{\int_{-\infty}^{0} f(t,\theta)\pi_0(d\theta)}{\int_{-\infty}^{0} f(t,\theta)\pi_0(d\theta) + \int_{0}^{\infty} f(t,\theta)\pi_1(d\theta)} .$$

It follows at once that in the exponential case $\phi^{\pi}(t)$ is nonincreasing. ϕ as an almost sure limit of Bayes estimators is thus equal almost surely to a nonincreasing function. Without loss of generality assume ϕ is nonincreasing.

Let $\mathcal{C} = \{t : 0 < \phi(t) < 1\}$. Then $\mathcal{C}$ is an interval. We assume $\mathcal{C}$ contains two distinct points, hence has nonvoid interior. Define $\gamma^{\pi}(t)$ by

$$\gamma^{\pi}(t) = \frac{\int_{0}^{\infty} f(t,\theta)\pi_1(d\theta)}{\int_{-\infty}^{0} f(t,\theta)\pi_0(d\theta)} .$$

Thus $\phi^{\pi} = 1/1+\gamma^{\pi}(t)$. If $t \in \mathcal{C}$, then $0 < \gamma^{\pi}(t) < \infty$.

Let $\{\pi_{0n}, \pi_{1n}\}$ be a sequence of finite measures such that $\pi_{0n}(-\infty, \infty)+\pi_{1n}(-\infty, \infty) = 1$ with corresponding Bayes estimators $\phi^{\pi_n}(x) \to \phi(x)$ almost surely. Define γ^{π_n} as above, and let

$$\pi_n(A) = \pi_{0n}(A) + \pi_{1n}(A) \quad \text{and} \quad \lambda_n(A) = \int_A e^{t_0\theta - \Psi(\theta)} \pi_n(d\theta),$$

where t_0 is in the interior of $\mathcal{C}$ by hypothesis. Renormalize so that $\lambda_n(-\infty, \infty) = 1$. Then the sequence $\{\lambda_n\}$ is tight. To show this, let $\epsilon > 0$ and $a_n \to \infty$ such that $\lambda_n[(a_n, \infty)] \geq \epsilon$. Take $t \in \mathcal{C}$, $t > t_0$. Then

$$\sup_n \int_{-\infty}^{0} e^{(t-t_0)\theta} \lambda_n(d\theta) \leq 1$$

and

$$\limsup_{n\to\infty} \int_0^{\infty} e^{(t-t_0)\theta} \lambda_n(d\theta) \geq \epsilon \limsup_{n\to\infty} e^{(t-t_0)a_n} = +\infty .$$

Hence $\limsup_{n\to\infty} \gamma^{\pi_n}(t) = +\infty$, which is a contradiction.

If $a_n \to -\infty$ and $\mu_n[(-\infty, a_n)] \geq \epsilon$, take $t \in \mathcal{C}$, $t - t_0 < 0$. Then

$$\limsup_{n\to\infty} \int_{-\infty}^{0} e^{(t-t_0)\theta} \lambda_n(d\theta) \geq \limsup_{n\to\infty} \epsilon e^{(t-t_0)a_n} = +\infty$$

and

$$\sup_n \int_0^{\infty} e^{(t-t_0)\theta} \lambda_n(d\theta) \leq 1 .$$

Thus $\liminf_{n\to\infty} \gamma^{\pi_n}(t) = 0$, which is again a contradiction.

Define $\lambda_{in}(A) = \int_A e^{t_0\theta - \Psi(\theta)} \pi_{in}(d\theta)$. The sequences $[\lambda_{in}\}$ are tight and $\lambda_{0n} \to \lambda_0$, $\lambda_{1n} \to \lambda_1$ (if necessary by taking subsequences). Thus $\lambda_0 + \lambda_1$ is a probability measure.

The assumption $0 < \gamma(t) < \infty$ implies that if $t \in$ interior $\mathcal{C}$, then (as shown above)

$$\sup_n \int e^{(t-t_0)\theta} \lambda_{in}(d\theta) < \infty, \quad i = 0,1 .$$

It then follows that if $t_1, t_2 \in$ interior $\mathcal{C}$ and $t_1 < t_2$, then the sequences

$$\tilde{\lambda}_{in}(A) = \int_A \left(e^{(t_1-t_0)\theta} + e^{(t_2-t_0)\theta}\right) \lambda_{in}(d\theta)$$

are tight. This follows since, by the preceding argument, t_0 is an arbitrary interior point of $\mathcal{C}$.

Thus assume $\tilde{\lambda}_{in} \to \tilde{\lambda}_i$, $i = 0,1$. If $t_1 < t < t_2$, and h is a bounded continuous function, then

$$\begin{aligned} &\lim_{n\to\infty} \int e^{(t-t_0)\theta} h(\theta) \lambda_{in}(d\theta) \\ &= \lim_{n\to\infty} \int \frac{e^{(t-t_0)\theta}}{e^{(t_1-t_0)\theta} + e^{(t_2-t_0)\theta}} h(\theta)\, \tilde{\lambda}_{in}(d\theta) \\ &= \int \frac{e^{(t-t_0)\theta}}{e^{(t_1-t_0)\theta} + e^{(t_2-t_0)\theta}} h(\theta)\, \tilde{\lambda}_i(d\theta) . \end{aligned}$$

For h with compact support it then follows that

$$\int e^{(t-t_0)\theta} h(\theta)\lambda_i(d\theta) = \int \frac{e^{(t-t_0)\theta}}{e^{(t_1-t_0)\theta} + e^{(t_2-t_0)\theta}} h(\theta)\,\tilde{\lambda}_i(d\theta)$$

and thus $\tilde{\lambda}_i = \left(e^{(t_1-t_0)\theta} + e^{(t_2-t_0)\theta}\right)\lambda_i$. When $h \equiv 1$ then

$$\lim_{n\to\infty} \int e^{(t-t_0)\theta}\lambda_{in}(d\theta) = \int e^{(t-t_0)\theta}\lambda_i(d\theta)$$

or

$$\lim_{n\to\infty} \gamma^{\pi_n}(t) = \gamma(t) = \frac{\int_0^\infty e^{t\theta}e^{-t_0\theta}\lambda_i(d\theta)}{\int_{-\infty}^0 e^{t\theta}e^{-t_0\theta}\lambda_0(d\theta)}\,. \qquad \square$$

Remark. This argument is still correct for the discrete exponential families. Here the interesting points of the truncation set $\mathcal{C}$ are atoms of the integrating measure. For each atom $t \in \mathcal{C}$, $0 < \lim \phi^{\pi_n}(t) < 1$ and the above argument goes through.

Figure 1: Risks for testing H_0: $\theta \leq 0$ versus H_1: $\theta > 0$, based on one observation x from a $n(\theta, 1)$ distribution. The solid line is the risk of the p-value, P(Z>x), where Z is a standard normal random variable. Bayes risks are given for the estimator $\phi_\tau(x) = P\left(Z > \left(\frac{\tau^2}{\tau^2+1}\right)^{\frac{1}{2}} x\right)$ for two $n(0, \tau^2)$ priors, $\tau^2 = .01$ (short dashes) and $\tau^2 = .1$ (close dots). The risk of the Neyman-Pearson rules are also give for α=.05 (long dashes) and α=.25 (dots). Finally, the constant risk=.25 is the risk of the minimax estimator $\phi_0 = \frac{1}{2}$.