RapidX Upload

1 **Rapid #: -9618569**

Status	Rapid Code	Branch Name	Start Date
New	PAU	Van Pelt	9/5/2015 12:53:20 PM
Pending	NAM	Science	9/5/2015 12:53:38 PM
Batch Not Printed	NAM	Science	9/5/2015 2:10:24 PM

CALL #:	**QA 276 A1 S75**
LOCATION:	**NAM :: Science :: SCIENCE Per**
TYPE:	Article CC:CCL
JOURNAL TITLE:	Statistics & Risk Modeling
USER JOURNAL TITLE:	Statistics
NAM CATALOG TITLE:	Statistics & decisions
ARTICLE TITLE:	ESTIMATING THE TRUTH INDICATOR FUNCTION OF A STATISTICAL HYPOTHESIS UNDER A CLASS OF PROPER LOSS FUNCTIONS
ARTICLE AUTHOR:	J. T. Gene Hwang
VOLUME:	15
ISSUE:	2
MONTH:	
YEAR:	1997-02-01
PAGES:	103
ISSN:	2193-1402 NAM ISSN: 0721-2631
OCLC #:	
CROSS REFERENCE ID:	[TN:981266][ODYSSEY:130.91.116.120/VPL]
VERIFIED:	

BORROWER: **PAU :: Van Pelt**

RAPIDILL

Statistics & Decisions 15, 103 – 128 (1997)
© R. Oldenbourg Verlag, München 1997

ESTIMATING THE TRUTH INDICATOR FUNCTION OF A STATISTICAL HYPOTHESIS UNDER A CLASS OF PROPER LOSS FUNCTIONS

J. T. Gene Hwang[1] and Robin Pemantle[2]

Revised: August 6, 1996

Abstract

We address the question of estimating an indicator function. The specific application of this framework is the estimation of the truth indicator function of a statistical hypothesis. We consider a class of "proper loss functions"; it can be argued that this is the class of all the reasonable loss functions. A general belief is that optimality (such as admissibility) of an estimator will remain the same for various proper loss functions. Under certain conditions, we justify the assertion through admissibility. We apply the results to evaluate the p-value and find that it is inadmissible for two-sided tests and admissible for one-sided tests. This result lends strong support for the result of Hwang, Casella, Robert, Wells and Farrell (1992) under quadratic loss, since our results are for a wide class of proper loss functions.

Key Words and Phrases: Hypothesis testing, p-values, decision theory.

AMS 1980 Subject Classification: 62F99, 60C99, 62A99

Research of both authors is partially supported by the Mathematical Sciences Institute, Cornell University.

[1] The author is formerly known as Jiunn T. Hwang. His research is partially supported by National Science Foundation Grant No. DMS 88-09016.

[2] Research partially supported by a National Science Foundation Postdoctoral Fellowship.

1. Introduction.

In testing a statistical hypothesis

$$\theta \in H,$$

where H is a subset of the parameter space, the practitioner will typically want to know how likely the hypothesis is to hold or how much evidence the data provides regarding its truth. We shall refer to the hypothesis as hypothesis H or H in short. A typical frequentist answer will be that one can reject H under such and such level. Actually, a better answer is to provide the p-value, i.e., the smallest α such that H can be rejected. Especially when p-value is small, it is not unusual that the p-value is thought of as an estimate of the probability that $\theta \in H$, which is from the frequentist's point of view either zero or one, i.e., the indicator function

$$I_H(\theta) = \begin{cases} 1 & \text{if } \theta \in H \\ 0 & \text{otherwise.} \end{cases}$$

This indicator function is called the truth indicator function in this paper.

Recent studies about the validity of p-value seem to be quite controversial. See Lindley (1982). In particular, the studies in Berger and Sellke (1987) and Casella and Berger (1987) make different conclusions. By comparing the p-value to the Bayes posterior probability of $\theta \in H$, the former study concluded that p-value is bad and tends to be too small for two sided tests (see also Berger and Delampady (1987)); and the study of Casella and Berger (1987) argued that p-value can be reconcilable with a Bayesian's approach for one sided tests. Throughout the paper, the one-sided test refers to the case where $H = (-\infty, \theta_0)$ for some θ_0 and the two-sided test refers to the case $H = [\theta_0, \theta_1]$ where θ_0 and θ_1 are interior points of the parameter space.

A decision theory was called for to determine the validity of the p-value by considering the problem of estimating $I_H(\theta)$ under loss functions including the quadratic loss

$$(1) \qquad Q(\theta, p(x)) = (I_H(\theta) - p(x))^2.$$

The result was reported in Hwang, Casella, Robert, Wells and Farrell (1992), abbreviated as HCRWF below. Under Q, it was found that the p-value was admissible for the one-sided test and inadmissible for the two-sided test. This result resolves the controversy to some extent. Earlier Schaafsma (1989) and Schaafsma, Tolboom and Van Der Meulen (1989) also studied the formulation. In a recent thesis, Van Der Meulen (1992) has some related results. A related but different subject about confidence set estimation can also be found in Brown and Hwang (1991) and Hwang and Brown (1991) and the references therein. This approach falls in the general area of the frequentist post-data inference which is surveyed in Goutis and Casella (1995).

Since the quadratic loss function is quite arbitrary, there seems to be a need to reconsider the problem under other reasonable loss functions. Furthermore, it is generally believed that results should not vary for different loss functions as long as they are proper. The definition of proper losses is given in Section 2, in which we also argue (by considering a weatherman) that this defines the class of all reasonable loss functions. We prove a theorem in Section 4 which states that admissibility under Q implies the same under virtually any other bounded proper loss functions and vice versa. Direct application of the theorem and the result of HCRWF implies that for the two-sided tests, the p-value is inadmissible under virtually any other bounded proper loss functions. For one-sided tests, the admissibility of p-value therefore also holds for virtually any bounded loss function, as well as other unbounded loss functions as established in Section 3. These results, therefore, provide overwhelming evidence that the fate of the p-value (interpreted as an estimator of $I_H(\theta)$) depends on whether one-sided or two-sided testing is involved.

2. Proper Loss Functions.

We are going to examine some natural restrictions on loss functions in order to see what may reasonably be termed the most general class of loss functions for estimators of binary events. We take as a model a weatherman trying to predict whether it will rain by using a probability . We then give the weatherman a penalty which depends on the probability and on whether it actually rains. The penalty should be large if it rains and his asserted probability of rain is near 0 or if it fails to rain when the forcasted probability of rain is near 1. We hope that by comparing the penalties incurred by this weatherman to the penalties incurred by other weathermen, we can tell in the long run which is better at predicting the weather. There are, of course, many other possible examples such as predicting the outcome of a ball game, the outcome of an election of a particular candidate or the outcome of any event as long as there are two outcomes and exactly one will happen.

The above model also applies to different statistical problems. In particular, in a hypothesis testing scenario we could substitute the event $\theta \in H$ for rain and the p-value (or any other estimator) for the weatherman's probability of rain. We want to keep the loss functions as general as possible, so as to be able to apply the theory to weather, testing hypotheses problems, estimation of confidence of sets, or other yet unnamed problems in decision theory. For ease of exposition, we will continue using weather prediction as an example.

When the forcasted probability is p, let $\ell_1(p)$ and $\ell_0(p)$ denote, respectively, the loss incurred when $I_H = 1$ (if it rains) or $I_H = 0$ (if it does not). Equivalently, we are using the loss function

$$L(\theta,p) = I_H(\theta)\ell_1(p) + (1-I_H(\theta))\ell_0(p). \tag{1}$$

The loss function is said to be proper if

$$\min_{0\le a\le 1} (p\ell_1(a)+(1-p)\ell_0(a)) = p\ell_1(p)+(1-p)\ell_0(p). \tag{2}$$

The motivation behind the definition is as follows. Suppose that the weatherman's subjective probability of rain is p. Then his expected penalty for predicting a according to the subjective probability is $p\ell_1(a)+(1-p)\ell_0(a)$. The weatherman will do best in the long run by predicting an a to minimize this expression. If the expression is not minimized at $a = p$, the weatherman will be forced to lie. Thus a proper loss function is one which allows a Bayesian not to lie. A strictly proper loss function is one that forces a Bayesian to tell the truth, i.e., one for which the minimum in the left hand side of equation (2) is uniquely achieved by p. The term proper is taken from Lindley (1982) and from Winkler and Murphy (1968) and the idea goes back at least to de Finetti (1962).

Clearly, ℓ_1 must be nonincreasing and ℓ_0 must be nondecreasing for the penalty to make any sense at all and hence these are assumed throughout the paper. Also, adding constants to ℓ_0 or ℓ_1 does not change the ordering of the class of decision procedures. Therefore, let

$$\ell_1(1) = \ell_0(0) = 0, \tag{3}$$

so all penalties are nonnegative. Since the values of ℓ_1 and ℓ_0 at points of discontinuity may be chosen in any of several ways, another convenient normalization condition is

$$\lim_{a\downarrow p} \ell_1(a) = \ell_1(p), \quad \lim_{a\downarrow p} \ell_0(a) = \ell_0(p). \tag{4}$$

We allow the possibility that $\ell_0(1)$ or $\ell_1(0)$ is infinite: if the forecasted probability of rain is truly zero, the weatherman will have made precautions that raining does not occur.

Theorem 1. For the function L in (1) with ℓ_1 and ℓ_0 satisfying conditions (3) and (4) to be proper, it is necessary and sufficient that there be a finite measure M on [0,1] such that $\ell_0(p) = \int_{(0,p]} \frac{dM(z)}{1-z}$ and $\ell_1(p) = \int_{(p,1)} \frac{dM(z)}{z}$. From this it follows that $\int_0^1 \ell_0(p)dp = \int_0^1 \ell_1(p)dp = M([0,1])$.

We omit the proof of the above theorem. Characterization of this type can be found in, for example, Schervish (1989), as well as Van Der Meulen (1992). Our characterization is slightly different and the necessary part is slightly stronger.

Throughout the paper, we use λ to represent the Lebesgue measure.

Example. Suppose $M = \lambda$. Then $\ell_1(p) = -\ell n(p)$ and $\ell_0(p) = -\ell n(1-p)$ so $L(\theta,p) = -\ell n|I_H-(1-p)|$. This is called the entropy or logarithmic loss function (Winkler and Murphy 1968). Note that the penalties are infinite for a mistaken prediction of certainty, which may be reasonable in certain situations. Suppose $dM/d\lambda = 2x(1-x)$. This is the familiar quadratic loss and is denoted as $Q(I_H(\theta),p) = (I_H(\theta)-p)^2$, which dated back to Brier (1950). An example of a proper loss function that is not absolutely continuous is when $M = \delta(p_0)$, $0 < p_0 < 1$. Then $\ell_1(p) = 1/p_0$ for $p < p_0$ and zero otherwise, while $\ell_0(p) = 0$ for $p < p_0$ and $1/(1-p_0)$ otherwise. This corresponds to a decision problem where there will be only two actions, one if the forecasted probability is at least p_0, and the other if the forecasted probability is less than p_0 (see Berger 1985).

3. Admissibility Results.

In this section we will investigate conditions under which an estimator p(X) is admissible. It is known in HCRWF that any admissible estimator with respect to Q is generalized Bayes (or possibly with some minor modification). We will therefore focus on these estimators and give, in Theorem 2, a sufficient condition under which they are admissible with respect to L. In Theorems 3 and 4, we will also provide a device to check whether p-values are generalized Bayes.

Assume that the observation X has a probability density function $f_\theta(X)$ (p.d.f.) with respect to a measure μ. Here θ is the unknown parameter ranging in the parameter space.

We define $\alpha_\pi(X)$ to be a generalized Bayes estimator with respect to $\pi(\theta)$ if $\alpha(X) = \alpha_\pi(X)$ minimizes the posterior risk,

$$EL(\theta,\alpha(X))|X = \int L(\theta,\alpha(X))f_\theta(X)d\pi(\theta)/\int f_\theta(X)d\pi(\theta).$$

This is an expectation integrated against the conditional distribution of θ given X, called the posterior distribution. When $\pi(\theta)$ is a probability measure, the generalized Bayes estimator is called the Bayes estimator.

A sufficient condition for the admissibility of a generalized Bayes estimator is that it is unique and it has a finite generalized Bayes risk,

$$\int R(\theta,\alpha_\pi)d\pi(\theta),$$

where for an arbitrary estimator $\alpha(X)$,

$$R(\theta,\alpha) = \int L(\theta,\alpha(x))f_\theta(x)d\mu$$

is the risk function. Under a strictly proper loss, the generalized Bayes estimator can be uniquely determined to be

$$\alpha_\pi(x) = P(\theta \in H|X = x).$$

Putting these together and using Fubini's theorem, the generalized Bayes risk is

$$(1) \qquad \int [\ell_1(\alpha_\pi(x))\alpha_\pi(x) + \ell_0(\alpha_\pi(x))(1-\alpha_\pi(x))]f_m(x)d\mu$$

where $f_m(x)$ denotes the marginal p.d.f. of X with respect to μ, i.e.,

$$f_m(x) = \int f_\theta(x)d\pi(\theta).$$

In summary, we have the following theorem.

Theorem 2. Assume that L is a strictly proper loss function. The generalized Bayes estimator $\alpha^\pi(x)$ is admissible if (1) is finite.

It is usually rather straightforward to check whether (1) is finite. The condition does hold for many generalized Bayes rules based on one-sided tests.

Example 1. Let $H = (-\infty,\theta_0)$ and assume that X has a $N(\theta,1)$ distribution where $-\infty < \theta < \infty$ is the parameter space. (The case of several i.i.d. normal observations with known variance can be reduced to this canonical form by considering the sample mean.) We can further assume that $\theta_0 = 0$ by a location transformation. The p-value is then $p(X)$ where $p(x) = P_0(X > x) = 1-\Phi(x)$, and Φ is the cumulative distribution function of $N(0,1)$. It is easy to demonstrate that $p(X)$ is generalized Bayes w.r.t. the Lebesgue measure λ (see continuation of Example 1 below). Hence $p(X)$ is admissible under L if (1) is finite where $\alpha^\pi(x)$ is replaced by $p(x)$. This is equivalent to

$$(2) \qquad \int \ell_1(1-\Phi(x))(1-\Phi(x))dx + \int \ell_0(1-\Phi(x))\Phi(x)dx < \infty.$$

Assume that M is the measure corresponding to L in Theorem 1. It is easy to show that a sufficient condition for the finiteness of (2) is

$$(3) \qquad \int_{1/2}^{1} |\ell n(1-t)|^{1/2}dM(t) + \int_0^{1/2} |\ell n\ t|^{1/2}dM(t) < \infty.$$

Condition (3) is not a very strong assumption. In particular it is satisfied by the quadratic loss and the logarithmic loss. In general, it suffices to have $M(0,a]$ and $M[1-a,1]$ going to zero as fast as a^ε for some $\varepsilon > 0$.

The more difficult question is under what situation the p-value for the one-sided test, $H = (-\infty,\theta_0)$, is generalized Bayes. This problem is addressed here. Let us suppose that $T(x)$ is a one-dimensional summary statistic and also the test is rejected if $T(x) > c$. Assume that the p-value is $P_{\theta_0}(T(X) \geq T(x)) = p(T(x))$.

Theorem 3. Assume that there exists a transformation $h(\cdot,\cdot)$ such that $h(Z_1,Z_2)$ is strictly increasing in Z_1 and is strictly decreasing in Z_2. Further suppose that the sampling distribution of $h(T(X),\theta_0)$ is identical to the posterior distribution of $h(T(X),\theta)$ given $T(X)$ with respect to a generalized prior distribution. Then the p-value is generalized Bayes with respect to such a prior.

Proof: Let the cumulative distribution function of the common distribution be denoted by F. Hence the p-value is

$$\begin{aligned} P_{\theta_0}(T(X) \geq T(x)) &= P_{\theta_0}(h(T(X),\theta_0) \geq h(T(x),\theta_0)) \\ &= 1-F(h(T(x),\theta_0)). \end{aligned}$$

Similar calculation shows that the generalized Bayes estimator is also given by the last expression, completing the proof.

The above theorem does not provide a way to construct the prior. For the continuous case, the following theorem is useful for such a derivation. Assume that h is differehtiable and $h_1'(\cdot,\cdot)$ and $h_2'(\cdot,\cdot)$ are respectively the partial derivatives of h with respect to the first and second coordinates. Further let $h_1^{-1}(\cdot,\cdot)$ denote the inverse function of h with respect to the first variable while the other variable is held fixed. Hence, by definition, $h_1(h_1^{-1}(Z_1,Z_2),Z_2) = Z_1$. Similarly $h_2^{-1}(Z_1,Z_2)$ is defined such that $h_2(Z_1,h_2^{-1}(Z_1,Z_2)) = Z_2$.

Theorem 4. Under the above assumptions about h and the assumption that T has a p.d.f. $f_\theta(t)$ with respect to the Lebesgue measure, the condition of Theorem 3 (stated in the second sentence) is equivalent to the condition that we may write

$$(4) \qquad \frac{f_{\theta_0}(h_1^{-1}(\eta,\theta_0))}{|h_1'(h_1^{-1}(\eta,\theta_0),\theta_0)|} \Big/ \frac{f_{h_2^{-1}(t,\eta)}(t)}{|h_2'(t,h_2^{-1}(t,\eta))|} = \pi(h_2^{-1}(t,\eta))\cdot g(t)$$

for some nonnegative functions π and g. When (4) holds then the p-value is generalized Bayes with respect to the prior with density π.

Proof: The density of the sampling distribution of $\eta = h(T(X),\theta)$ is

$$\frac{f_{\theta_0}(h_1^{-1}(\eta,\theta_0))}{|h_1'(h_1^{-1}(\eta,\theta_0),\theta_0)|}$$

where $\theta = \theta_0$. The posterior density of θ is proportional to $f_\theta(t)\pi(\theta)$ with proportionality being only a function of t, denoted by $g(t)$. Hence the posterior density of η is proportional to

$$\frac{f_{h_2^{-1}(t,\eta)}(t)\pi(h_2^{-1}(t,\eta))}{|h_2'(t,h_2^{-1}(t,\eta))|} .$$

Hence it is obvious that the assumption of Theorem 3 is equivalent to (4). The rest of the theorem follows easily.

Example 1 (continuation). Take a transformation $\eta = X-\theta$. Then (4) is proportional to one and hence taking $\pi(\theta) = 1$ will generate the p-value as the generalized Bayes estimator by Theorem 4, as also similarly concluded in Example 1 above for this normal case. The normality assumption is not important, however. In general, consider the case where the distribution of X has θ as the location parameter. We may use the transformation $\eta = X-\theta$ and apply Theorem 4 to establish the same conclusion. Hence the p-value is generalized Bayes with respect to $\pi(\theta) = 1$.

Example 2. Assume that X_i are i.i.d. gamma distributed with a scale parameter $\theta > 0$. Consider only estimators based on the sufficient statistic $T = \Sigma X_i$. To apply Theorem 4, consider a transformation $\eta = T/\theta$. The following derivation works for any scale parameter distribution with $T \sim f_\theta(t) = f(t/\theta)/\theta$. Direct calculation shows that (4) equals η also $h_2^{-1}(t,\eta) = \frac{t}{\eta}$. This leads to choosing $\pi(\theta) = 1/\theta$. Hence Theorem 3 implies that the p-value is generalized Bayes w.r.t. such a prior.

Despite the applications of Theorems 3 and 4 in Examples 1 and 2, the authors are not sure whether there are other applications.

Now we turn to investigate the admissibility of the p-value for testing $H_0:\theta \le \theta_0$ in Example 2. We assume without loss of generality that $\theta_0 = 1$. The p-value is then

$$P_1(T > t) = \int_t^\infty x^{\alpha-1} e^{-x} dx/\Gamma(\alpha)$$

where $\alpha > 0$ is assumed to be known. Note that $P_1(T > t)$ is of order $t^{\alpha-1}e^{-t}$ for a large t. Now we apply Theorem 1. Here

$$m(t) = \int_0^\infty e^{-t/\theta}\ t^{\alpha-1}/[\Gamma(\alpha)\theta^{\alpha+1}]d\theta = \frac{1}{t}\int_0^\infty e^{-z}z^{\alpha-1}/\Gamma(\alpha)dz$$

which is proportional to t^{-1}. Hence the finiteness of (1) is equivalent to

$$\text{(5)} \qquad \int_0^\infty [\ell_1(p_1(t))p_1(t)+\ell_0(p_1(t))(1-p_1(t))]\frac{1}{t}\, dt < \infty.$$

It can be shown that this is equivalent to

$$\text{(6)} \qquad \int_1^\infty \ell_0(p_1(t))\frac{1}{t}\, dt < \infty.$$

Using the fact that $p_1(T > t)$ is smaller than $e^{-t/2}$ as $t \to \infty$, one can show that (6) is implied by

$$\text{(7)} \qquad \int_0^{1/2} (\ell n(|\ell n(t)|))dM(t) < \infty.$$

This condition is weaker than (3) and holds for most proper loss functions.

In summary, it follows from Theorem 2 that the p-value is admissible whenever (6) or (7) hold. Interestingly, the p-value is one at $t = 0$ and yet it is still generalized Bayes and admissible. The situation is quite different from what was discussed in HCRWF for the two sided case.

4. A relation between the quadratic loss and other proper losses for two-sided problems.

Under certain conditions, it is shown that admissibility (inadmissibility) of an estimator under Q implies admissibility (inadmissibility) under most of other proper loss functions. Here Q is the quadratic loss defined in (1) of the Introduction. One

application of this theorem is to two-sided testing hypothesis problems. In HCRWF, it was shown that for exponential families and under the quadratic loss function, that the p-values corresponding to one-sided tests are typically admissible while those corresponding to two-sided tests are typically inadmissible. By using the theorems of this section, it then follows that these results also hold for many other proper losses.

The admissibility result here applies to any estimators admissible under Q, but to a slightly smaller class of loss functions than in Section 3.

Although our arguments can presumably be generalized to many other parametric assumptions, we will focus only on the exponential family of distributions. Hence assume that the p.d.f. of the random vector X w.r.t. μ_0 is

$$g_0(\theta)e^{\theta T(x)}.$$

The natural parameter space N is an interval. The testing hypothesis problem we will focus on is

$$H_0: \theta \in H = [\theta_0, \theta_1]$$

where $\theta_0 \leq \theta_1$ are assumed to be interior points of N. We will estimate $I_H(\theta)$ under proper loss functions. We will consider only non-randomized estimators which are functions of the one-dimensional sufficient statistic $T(X)$. Note that the distribution of $T(X)$ also belongs to an exponential family having the p.d.f.

$$f_\theta(t) = e^{\theta t}g(\theta)$$

with respect to some measure μ. Below we will use the property that for $\theta \in N$, $g(\theta)$ is continuous and the expectation of the function of T is also continuous in θ as long as the expectation is finite. See Exercise 1.13.1 on p.28 of Brown (1986).

Although we consider only non-randomized estimators based on T, we do not lose any generality. This assertion is obvious for strictly convex loss functions. It is also true for any proper loss functions with a measure M such that the function $\kappa(x)$ as defined by $M(0,x]$ is strictly increasing. In fact, in such a case, all the admissible non-randomized rules based on T forms a minimal complete class due to the reason explained below. One can prove that $\ell_i(\kappa^{-1}(x))$, $i = 1,2$, are strictly convex. In other words, there is a change of scale under which ℓ_i are simultaneously strictly convex. Hence a randomized rule is dominated by its "expectation" with respect to its new scale. This rule, in turn, is dominated by its conditional expectation given T, again with respect to the new scale.

The Admissibility Theorem.

Theorem 5. Assume that the natural parameter space $N = (-\infty,\infty)$, and $p_0(T)$ is an arbitrary admissible estimator under Q, then $p_0(T)$ is admissible under any strictly proper bounded loss function L.

Proof: By Theorem 5.1 of HCRWF, $p_0(t)$ must be a "modified" generalized Bayes estimator for $t \in (t_1,t_2)$, a truncation set. That is

$$(1) \qquad p_0(t) = \frac{\int f_\theta(t)d\pi_0(\theta)}{\int f_\theta(t)d\pi_1(\theta) + \int f_\theta(t)d\pi_1(\theta)} \qquad t \in (t_1,t_2)$$

$$= 0 \qquad t \notin [t_1,t_2]$$

where π_0 is a probability supported on $[\theta_0,\theta_1]$ and π_1 is a σ-finite measure supported on $(-\infty,\theta_0] \cup [\theta_1,\infty)$. One can easily show that $p_0(t)$ is admissible under L, since it has finite modified Bayes risk by comparing to the zero estimator.

The Inadmissibility Theorem. The inadmissibility theorem is more difficult to establish and more assumptions are needed. However, unlike in Theorem 5, we make no assumption about N.

Theorem 6. Let $p(T)$ be an arbitrary estimator and $L(\theta,x)$ any proper loss function. Assume that the derivative of L with respect to x is continuous and nonzero for $x \in [0,1]$ and that its representing measure M_L has a derivative with respect to the Lebesgue measure. If $p(T)$ is inadmissible under Q, then it is inadmissible under L. □

The theorem is proved in the Appendix. Below, we sketch the main idea by constructing an estimator q_L that domains p for loss L, given an estimator of q_Q that dominates p for loss Q. For each θ, q_Q has a lower quadratic risk than p, so

$$E_\theta Q(\theta,q_Q(T)-Q(\theta,p(T)) < 0.$$

By convexity, the integrand is bounded below by $(q_Q(T)-p(T))Q'(\theta,p(T))$ where $Q'(\theta,p) = \frac{\partial}{\partial p} Q(\theta,p)$. Hence

$$(2) \qquad E_\theta(q_Q(T)-p(T))Q'(\theta,p(T)) < 0.$$

Now depending on $\theta \in [\theta_0,\theta_1]$ or not, $L'(\theta,x) \equiv \frac{\partial}{\partial x} L(\theta,x)$ is either $(-1/x)dM_L/dx$ or $[1/(1-x)]dM_L/dx$. Similar expressions hold for $Q'(\theta,x)$. Hence $Q'(\theta,x)/L'(\theta,x) = dM_Q/dM_L(x)$ independent of θ, and

$$(3) \qquad \int (q_Q(T)-p(T))\frac{dM_Q}{dM_L}(p(T))L'(\theta,p(t))f_\theta(t)d\mu(t) < 0.$$

Let $q_L = q_{L,S}$ so that

$$q_{L,S}(T)-p(T) = S(q_Q(T)-p(T))dM_Q/dM_L(p(T)),$$

where S is a positive constant. If we are allowed to differentiate under the expectation, it is obvious that the lefthand side of (3) is $\frac{d}{dS} E_\theta L(\theta, q_{L,S}(T))$ evaluated at $S = 0$. This together with the fact that $q_{L,S}(T)$ reduces to $p(T)$ at $S = 0$ and hence both have identical risk function imply that $q_{L,S}(T)$ dominates $p(T)$ with respect to L at θ for small S. Since we are interested in domination uniformly in θ, the bulk of the proof in the Appendix is concerned with this uniformity. Effort is needed to make the differentiation and the choice of q rigorous.

In fact, it appears that the argument works with any $f_\theta(x)$. However, we assume an exponential family to simplify the analytic arguments.

Putting Theorems 5 and 6 together give the following corollary.

Corollary 11. For a strictly proper loss function satisfying the conditions of Theorems 5 and 6, the class of admissible estimators are given by (1).

5. What Dominates the p-Value for the Quadratic Loss?

We have seen in Theorem 6 and in HCRWF that the p-value is an inadmissible estimator (under the quadratic loss, hence under any proper continuously differentiable loss) for the truth indicator function of a point null in the two-sided case based on a $N(\theta,1)$ observation X. Van Der Meulen (1992, p.81) had a similar inadmissibility result which applies to convex and proper loss functions. It is noted, however, in HCRWF, by looking at the size of the tails of the p-value $p(X)$, that it cannot be dominated by any proper Bayes estimator, leaving open the question of what generalized Bayes estimator dominates the p-value for the quadratic loss. A concrete example would be useful since it could be plugged into Theorem 6 to produce an estimator dominating p for any loss function satisfying the assumptions. In this section we exhibit such an estimator, q. The demonstration that q dominates p is numerical and we know of no analytic proof that q dominates p.

Let m and k be parameters to be specified later and let $\psi(\theta)$ be the positive part of $m\theta^2-k$, i.e., the maximum of 0 and $m\theta^2-k$. Let $\pi(m,k)$ be the σ-finite measure having density ψ with respect to Lebesgue measure together with a mass at zero of weight 1. For various values of m and k the generalized Bayes estimate $q(m,k)$ with prior $\pi(m,k)$ dominates the p-value. In particular, the estimator $q(.5,1)$ has 12% less quadratic risk than the p-value for nonzero values of $|\theta|$ up to 1 or greater than 5, improvement dipping to a low of 1.3% at $\theta = \pm 3.4$. When $m = .7$ and $k = 1.95$, $q(m,k)$ does only 8% better than the p-value for small θ but is more even, dropping no smaller than 2.5% at $\theta = 2.9$. In both cases, the k was chosen as a function of m to make the quadratic risk of q slightly better than that of p for $\theta = 0$. Below is a table comparing quadratic risks of the p-value to those of q for the two above choides of parameters. The table was computed by numerical integration after first obtaining the formula $q(x) = \varphi(x)/(\varphi(x)+H(x))$, where H is the integral of the continuous part of π against $\varphi(x-\theta)$ and can be written:

$$H(x) = \int_{-\infty}^{-\sqrt{k/m}} (m\theta^2-k)\varphi(x-\theta)d\theta + \int_{\sqrt{k/m}}^{\infty} (m\theta^2-k)\varphi(x-\theta)d\theta$$

$$= \int_{-\infty}^{-\sqrt{k/m}-x} \varphi(w)(mx^2-k+2mxw+mw^2)dw$$

$$+ \int_{\sqrt{k/m}-x}^{\infty} \varphi(w)(mx^2-k+2mxw+mw^2)dw$$

$$= H_0(x) + H_0(-x).$$

where

$$H_0(x) = (mz^2+m-k)\Phi(-\sqrt{k/m}-x)+m(\sqrt{k/m}-x)\varphi(\sqrt{k/m}+x).$$

θ	$R(p,\theta)$	$R(q(.5,1.004,\theta)$	$R(q(.7,1.95),\theta)$
0	.33333	.33179	.33277
0^+	.33333	.28858	.29996
.5	.30137	.26253	.27250
1	.22290	.19766	.20429
1.5	.13520	.12311	.12634
2	.06752	.06339	.06441
2.5	.02791	.02697	.02705
3	.009612	.009475	.009357
3.5	.002774	.002748	.002664
4	.0006751	.0006583	.0006248
4.5	.0001392	.0001303	.0001207
5	.0000244	.0000213	.0000192

Acknowledgement: We thank the referees, Professors L. D. Brown and Roger Farrell for many insightful comments and discussions.

Appendix

Note that all the loss functions considered in Theorem 6 are bounded for each θ. Below we assume without loss of generality that all the estimators take values in $[0,1]$, and hence they all have finite risk for every θ and every loss considered.

The proof of Theorem 6 will be based on the following lemmas whose proofs are very technical and are either omitted or sketched. See Hwang and Pemantle (1992) for the details.

Lemma A.1. Consider an estimator $p(T)$ inadmissible under the quadratic loss Q. Then there is an estimator q strictly dominating p for all θ and for some constant $0 < r < 1$ and finite cutoff points $c_- < c_+$.

$$(1) \qquad q(t) = rp(t) \quad \text{for} \quad t \notin [c_-,c_+].$$

Furthermore $q(t) \geq rp(t)$ for all t and

$$(2) \qquad \mu(t > c_+ : p(t) > 0\} > 0 \quad \text{and} \quad \mu\{t < c_- : p(t) > 0\} > 0.$$

In proving Lemma A.1, start out with an estimator q_1 which dominates p for all θ under Q. The estimtor

$$q(t) = \begin{cases} q_1(t) & t \in [c_-,c_+] \\ rp(t) & \text{otherwise} \end{cases}$$

with approprite choice of $c_\pm$ will do the work.

Proof of Theorem 6. Suppose $p(T)$ is a Q-inadmissible estimator. By Lemma 4.1, we may assume $q_Q(T)$ dominating $p(T)$ for the quadratic loss and satisfying all the other conclusions of Lemma A.1. Let

$$(3) \qquad B = \sup_z \frac{dM_Q}{dM_L}(z)$$

where M_Q and M_L are the representing measure for Q and L respectively.

The following lemma can be established by using the continuity of L and the identity $\ell_0'(t)(1-t) = \ell_1'(t)t$.

Lemma A.2. There exists a density of M_Q with respect to M_L so that B is finite.

Below we will work with such a density and hence B is finite. Construct a family of functions

$$\{q_{L,S} : 0 \leq S \leq \tfrac{1}{B}\}$$

defined by

(4) $$q_{L,S}(t) = p(t) + S(q_Q(t)-p(t))\frac{dM_Q}{dM_L}(p(t)).$$

Assume without loss of generality that q_Q and p give values in $[0,1]$. Since $q_{L,S}$ always lie between q_Q and p, so in particular it lies in the unit interval and is at least rp.

Let $R_L(\theta,q)$ represent the risk function of q under L. Hence

$$R_L(\theta,q) = E_\theta L(\theta,q(T)).$$

Lemma A.3.

(5) $$\varphi(\theta,S) \overset{\text{defn.}}{\equiv} \frac{d}{dS}R_L(\theta,q_{L,S}) = E_\theta \frac{d}{dS} L(\theta,q_{L,S}(T)).$$

From Lemma A.3

(6) $$\frac{d}{dS} R_L(\theta,q_{L,S}) = E_\theta L'(\theta,q_{L,S})(q_Q(T)-p(T))\frac{dM_Q}{dM_L}(p(T)).$$

Setting $S = 0$ in the above expression gives the left hand side of (3) in Section 4 which equals (2) in the same section. Hence $\varphi(\theta,0)$ is strictly negative for every θ. Since $q_{L,S}$ reduces to p for $S = 0$, we conclude that for every θ

(7) $$R_L(\theta,q_{L,S}) < R(\theta,p)$$

as long as S is small enough. It remains to find an S for which (7) holds for every θ. We establish this in three parts (i) $\theta \in [\theta_0,\theta_1]$ (ii) $\theta \in [a,b] \cap (\theta_0,\theta_1)^c$ and (iii) $\theta \in [a,b]^c$ where $[a,b]$ is any interval in the parameter space N with $a < \theta_0 \le \theta_1 < b$.

To do so, we need to extend the definition of $R_L(\theta,q)$. Let

$$R_L^0(\theta,q) = E_\theta \ell_0(q(t)).$$

Note that $R_L^0(\theta,q) = R_L(\theta,q)$ for all $\theta \in [\theta_0,\theta_1]^c$ and $R_L^0(\theta,q)$ is a continuous extension of R_L from $[\theta_0,\theta_1]^c$ to $(\theta_0,\theta_1)^c$.

Define

$$\varphi_0(\theta,S) = \frac{d}{dS} R_L^0(\theta,q_{L,S}).$$

Similar to (5) we have

(8) $$\varphi_0(\theta,S) = E_\theta \frac{d}{dS}(q_{L,S}(T)) = E_\theta \ell_0'(q_{L,S}(T))\frac{dM_Q}{dM_L}(p(T))$$

and hence $\varphi_0(\theta,0) < 0$ for $\theta \in [\theta_0,\theta_1]^c$. It can also be shown that $\varphi_0(\theta,0) < 0$ for $\theta = \theta_0$ or θ_1. This can be established similarly to (2) and (3) of Section 4 as long as one starts with a q_Q such that $R_L^0(\theta,q_Q) < R_L^0(\theta,p(t))$ for $\theta \in (\theta_0,\theta_1)^c$. Putting these together, we have

(9) $$\varphi_0(\theta,0) < 0 \quad \text{for} \quad \theta \in (\theta_0,\theta_1)^c.$$

We have the following property of $\varphi_0(\theta,S)$.

Lemma A.4. $\varphi_0(\theta,S)$ is jointly continuous in $\theta \in (\theta_0,\theta_1)^c$ and S.

Now we return to the proof of (7). We work with case (ii) first. For each θ, let $u(\theta)$ be the largest number u such that $\varphi_0(\theta,S) < 0$ for every $S < u$. By (9) and Lemma A.4, $u(\theta) > 0$. Note that

(10) $$\varphi_0(\theta,u(\theta)) = 0.$$

We claim that for $\theta \in [a,b] \cap (\theta_0,\theta_1)^c$, $u(\theta)$ is bounded below by $u_* > 0$. This then implies that for $0 < S < S_2$, where $S_2 = \frac{1}{2} u_*$, $\varphi_0(\theta,S) < 0$ for every $\theta \in [a,b] \cap [\theta_0,\theta_1]^c$ and therefore $q_{L,S}$ dominates p for case (ii).

To establish that $u(\theta)$ is bounded away from zero, assume that this is not the case. Then $(\theta,u(\theta))$ has a limiting point $(\theta_*,0)$ for some $\theta_* \in [a,b] \cap (\theta_\theta,\theta_1)^c$ by compactness. Hence $\varphi_0(\theta_*,0) = 0$ by (10) and by continuity of φ_0. This, however, contradicts (9). Inequality (7) is now established for θ in case (ii).

Case (i) can be similarly proved by considering L_1 instead of L_0. Hence there exists $S_2 > 0$ so that $R_L(\theta,q_{L,S})-R_L(\theta,p)$ is strictly negative for all θ and $0 < S < S_2$.

Now we only need to deal with case (iii). If N is compact, taking $[a,b] = N$ completes the proof of the theorem. If N is not compact, we have three cases:

$$N = [\theta_L,\theta_U),\ (\theta_L,\theta_U) \text{ or } (\theta_L,\theta_U].$$

All the three cases are similar and hence we consider only the first case. The lower end point causes no problem since we can handle by taking $a = \theta_L$. The only problem is as $\theta \to \theta_U$. Note that θ_U can be infinite or finite which will be considered below.

Now for $\theta < \theta < \theta$, write the difference in risks between $q_{L,S}$ and p as

(11) $$R_L(\theta,q_{L,S})-R_L(\theta,p)$$

$$= \int_{t\le c_+} f_\theta(t)[\ell_0(q_{L,S}(t))-\ell_0(p(t))]d\mu(t)$$

$$+ \int_{t>c_+} f_\theta(t)[\ell_0(q_{L,S}(t))-\ell_0(p(t))]d\mu(t)$$

where $c_\pm$ were defined in Lemma A.1. For $t > c_+$,

$$\ell_0(q_{L,S}(t))-\ell_0(p(t)) \le \frac{-2(p(t)-q_{L,S}(t))}{B} (rp(t))$$

$$= -2S(1-r)rp^2(t)\frac{dM_Q}{dM_L} (p(t))/B.$$

By (2), there exists a positive number $\epsilon > 0$ such that the set

$$D = \{t > c_+ + \epsilon ; p^2(t)\frac{dM_Q}{dM_L} (p(t))/B > \epsilon)$$

has a positive measure. Hence the second term on the right hand side of (11) is bounded above by

(12) $$-2S(1-r)\epsilon P_\theta(T > c_+ + \epsilon).$$

Now consider the first term on the right hand side of (11). Let θ^* be a fixed point with $\theta^* < \theta_U$. For θ such that $\theta^* < \theta < \theta_U$, and $t < c_+$,

$$f_\theta(t) = g(\theta)e^{\theta t} = g(\theta^*)e^{\theta^* t}\cdot e^{t(\theta-\theta^*)} \frac{g(\theta)}{g(\theta^*)}$$

$$\le g(\theta^*)e^{\theta^* t}e^{c_+(\theta-\theta^*)} \frac{g(\theta)}{g(\theta^*)}.$$

Hence the aforementioned first term is bounded above by

$$(13)\quad e^{c_+(\theta-\theta^*)}\,\frac{g(\theta)}{g(\theta^*)}\int_{t<c_+} f_{\theta^*}(t)(\ell_0(q_{L,S}(t))-\ell_0(p(t)))d\mu(t).$$

Arguments similar to those leading to (6) establish the finiteness of $\frac{d}{dS}\,EI_{(-\infty,c_+)}(T)\ell_0(q_{L,S}(T))\Big|_{S=0}$. This implies that the integral in (13) is bounded above by kS for S small enough where k is some constant independent of S and θ.

Now compare (13) to (12). If θ_U is finite, one can show by using Fatou's lemma and the fact $\theta_U \notin N$ that as $\theta \to \theta_U$, $g(\theta) \to 0$ and hence (13) converges to zero. Similarly one can show $P_\theta(T > c+\varepsilon) \to 1$ as $\theta \to \theta_U$. Therefore (12) is larger in magnitude than (13) for all S, $0 < S < S_3$, where S_3 is some finite number, and for θ sufficiently close to θ_U and hence (11) is negative for such S. What if $\theta_U = \infty$? We can use arguments similar to those leading to (13) to show that

$$(14)\quad P_\theta(T > c_+ + \varepsilon) \geq e^{(c_+ + \varepsilon)(\theta-\theta^*)}\,g(\theta)\int_{t>c_+ + \varepsilon} g(\theta^*)e^{\theta^* t}dt.$$

Using this lower bound, (12) is bounded above by $-2S(1-r)r$ times (14). Since there is an extra term $e^{\varepsilon\theta}$ in (14), the ratio of (14) to (13) without the integral term approaches ∞ as $\theta \to \infty$. Hence the rest of the argument is similar to the case when θ_U is finite.

In summary, $q_{L,S}$ dominates p for all θ and S such that $0 < S < \mathrm{Min}(S_1,S_2,S_3)$ and hence p is inadmissible. Q.E.D.

References

[1] Berger, J. O. (1985). Statistical Decision Theory and Bayesian Analysis. Springer-Verlag, New York.

[2] 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.

[3] Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses (with discussion). Statistical Science 2, 317-352.

[4] Brier, G. W. (1950). Verification of forecasts expressed in terms of probabilities. Month. Weather Rev. 78, 1-3.

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

[6] Brown, L. D. and Hwang, J. T. (1991). Admissibility of confidence estimators. Proceedings of the 1990 Taipei Symposium in Statistics, June 28-30, 1990.

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

[8] de Finetti, B. (1962). Does it make sense to speak of "good probability appraisers"? In The scientist speculates: an anthology of partly baked ideas, New York, Basic Books.

[9] de Finetti, B. and Savage, L. J. (1965). The elicitation of subjective probabilities. Unpublished manuscript.

[10] Goutis, C. and Casella, G. (1995). Frequentist post-data inference, International Stat. Review 63, pp.325-344.

[11] Hwang, J. T. and Brown, L. (1991). Estimated confidence approach under the validity constraint criterion. Annals of Statistics, 19, 1964-1977.

[12] Hwang, J. T., Casella, G., Robert, C., Wells, M. and Farrell, R. (1992). Estimation of accuracy of testing, Annals of Statistics, 20, 490-509.

[13] Hwang, J. T. and Pemantle, R. (1992). Estimating the truth of a statistical hypothesis under a class of proper loss functions. Cornell Stat. Center Technical Report.

[14] Lindley, D. V. (1982). Scoring rules and the inevitability of probability. International Stat. Review 50, pp. 1-26.

[15] Royden, H. L. (1988). Real analysis. Macmillan Publishing Company, New York.

[16] Schaafsma, W. (1989). Discussing the truth or falsity of a statistical hypothesis H and its negation A. Int. Workshop on Theory and Practice in Data Analysis, Proceedings, 150-166.

[17] Schaafsma, W., Tolboom, J. and Van Der Meulen, E. A. (1989). Discussing truth or falsity by computing a Q-value. Statistical Data Analysis and Inference, 85-100.

[18] Schervish, M. (1989). A general method for comparing probability assessors. Ann. Stat. 17, 1856-1879.

[19] Van Der Meulen, E. A. (1992). Assessing weights of evidence for discussing classical statistical hypotheses. Ph.D. thesis, University of Groningen, The Netherlands.

[20] Winkler, R. and Murphy, A. (1968). "Good" probability assessors. J. Appl. Meteor. 7, 751-758.

J. T. Gene Hwang
Department of Mathematics
Cornell University
Ithaca, NY 14853

Robin Pemantle
Department of Mathematics
University of Wisconsin
Madison, WI 53706

Statistics & Decisions 15, 129 – 140 (1997)
© R. Oldenbourg Verlag, München 1997

EXPANSION OF BAYES RISK FOR ENTROPY LOSS AND REFERENCE PRIOR IN NONREGULAR CASES

Subhashis Ghosal[1] and Tapas Samanta

Received: Revised version: *June 27, 1996*

Abstract

Lindley's measure of information in a sample about a parameter is given by the average Kullback-Leibler distance between the posterior and the prior. This is also equal to the Bayes risk when one estimates the density using the entropy loss. In this paper, an asymptotic expansion of this measure is obtained for a one-parameter family of discontinuous densities. This expansion is then used to obtain the reference prior in the sense of Bernardo.

1. Introduction

Let $X_1, X_2, \ldots, X_n$ be independent observations each having a distribution P_θ with a density $f(x;\theta)$ with respect to a fixed dominating measure where $\theta \in \Theta$, an open subset of $\mathbb{R}$. Consider a prior on Θ having a density $\pi(\cdot)$ with respect to the Lebesgue measure. Lindley's measure of information (see Lindley [15]) $I(\pi; X^n)$ in $X^n = (X_1, \ldots, X_n)$ about θ is given by the average relative entropy or Kullback-Leibler distance between the posterior distribution of θ given X^n and the prior π (see Sec. 2). This measure is also equal to the average (with respect to π) relative entropy distance between the distribution of X^n given θ and the marginal distribution of X^n and indeed is the Bayes risk when one estimates the density of X^n given θ using the entropy loss (see Aitchison [1]).

In Section 2 of this paper, we obtain an asymptotic expansion of this Bayes risk (or measure of information) $I(\pi; X^n)$ for a family of nonregular cases. We restrict our attention to the cases which, by results of Ghosh *et al.* [11], are essentially the only cases where the posterior distributions converge. Our treatment is similar to that of Clarke and Barron [9] who obtained an expansion of the entropy risk for the regular cases. Results similar to those of Clarke and Barron [9] were obtained earlier by Ibragimov and Has'minskii [13]. For extensions to non

[1] Research supported by the National Board of Higher Mathematics, Department of Atomic Energy, Bombay, India.

AMS 1991 *subject classification:* Primary 62F25.

Key words and phrases: Asymptotic expansion, Bayes risk, discontinuous densities, Kullback-Leibler number, posterior distribution, reference prior.