The Annals of Statistics
1988, Vol. 16, No. 3, 1262–1269

LOSS FUNCTIONS FOR LOSS ESTIMATION[1]

BY ANDREW L. RUKHIN

Purdue University

A class of proper scoring functions which combine the error in a decision problem and the precision of the statistical decision rule is introduced. The Bayesian procedures with respect to these loss functions are pairs formed by the usual Bayes decision and by the expected posterior loss. A necessary and sufficient condition for admissibility under the corresponding risk is given.

1. Introduction. Consider general statistical decision problems as described by possible states of nature θ, decisions d and a loss function $W(\theta, d)$. Classical decision theory advocates making some decision $d = \delta(x)$, where x is the observation, with frequentist risk $\mathscr{R}(\theta, \delta)$. There are important situations where one would like to accompany the decision δ with an estimate, say $\gamma = \gamma(x)$, of its inaccuracy or of the loss $W(\theta, \delta(x))$. In many examples the procedure δ has constant risk $\mathscr{R}(\theta, \delta) = E_\theta W(\theta, \delta(x)) = \mathscr{R}$, but a constant estimator $\gamma(x) = \mathscr{R}$ is unreasonable. In fact one would expect γ to take smaller values for "lucky" observations x.

The idea of estimated inaccuracy of a point estimator is behind the concept of a confidence interval [cf. Savage (1954), Chapter 17]. Indeed while the midpoint of such intervals typically may serve as a point estimator, its width indicates the inaccuracy of this estimator.

The problem of estimating the risk function has been considered by Lehmann (1959) who mentioned estimated power of a test and by Sandved (1968) who found unbiased estimators of risk corresponding to quadratic loss in several estimation situations. A lot of attention was brought to this problem by Kiefer who in a series of papers (1975, 1976, 1977) developed conditional and estimated confidence theories which, in particular, provide estimates of confidence or accuracy admitting frequentist interpretability. Berger (1985a, b, c) compares the subjective Bayesian approach to this problem with the frequentist one. In particular he discussed the desirable properties of valid measures of performance of a statistical decision rule δ from the frequentist point of view. We also note that estimated standard errors and other characteristics of statistical accuracy may be of interest in nonparametric and bootstrap methods [Efron (1982) and Efron and Tibshirani (1986)].

To be able to compare two estimators of the loss, one must specify an appropriate utility function. In fact a variety of loss functions in interval estimation has been considered [see Aitchison and Dunsmore (1968), Pratt (1961), Winkler (1972) and Cohen and Strawderman (1973)].

Received June 1986; revised September 1987.

[1] Research supported by National Science Foundation Grant DMS-84-01996.

AMS 1980 *subject classifications*. Primary 62C15; secondary 62A15, 62C10.

Key words and phrases. Loss functions, posterior risk, admissibility, concavity.

In this paper for a general decision problem we give a class of loss functions which combine the decision problem error with the error of inaccuracy estimate. These loss functions are very convenient in the problem of simultaneous "decision-precision" reporting. The corresponding risk, as any risk function, has frequentist interpretability in terms of long-run frequencies. The Bayes decision-precision pairs turn out to be the usual Bayes decision δ_B for θ and the posterior loss $\gamma_B = E\{W(\theta, \delta_B)|x\}$. Since admissible pairs in statistical decision theory are typically Bayes or generalized Bayes procedures, a frequentist may accept posterior loss as an estimate of risk because of the admissibility argument.

2. Loss functions combining decision error and estimated loss error.

Denote by γ an estimator of the nonnegative loss $W(\theta, \delta)$ and assume that a loss function $\mathscr{L}(\theta; \delta, \gamma)$ which combines the decision error $W(\theta, \delta)$ and the error in estimating $W(\theta, \delta)$ by γ is desired. We develop here an axiomatic approach to determine such a loss function $\mathscr{L}$ from two conditions.

The first condition is that for fixed γ, i.e., in the case when one does not have to estimate loss $W(\theta, \delta)$, the utility function $\mathscr{L}$ should be equivalent to W. If this equivalence is defined by the expected utility, then [see DeGroot (1970), Section 7.9]

$$\mathscr{L}(\theta; \delta, \gamma) = a(\gamma)W(\theta, \delta) + b(\gamma) \tag{2.1}$$

with positive function a. Thus we consider loss functions $\mathscr{L}$ only of the form (2.1).

According to the second condition for fixed δ, i.e., when decision δ is specified, $\mathscr{L}$ as a function of γ, $\gamma \geq 0$, must be uniquely minimized at

$$\gamma_{\min} = W(\theta, \delta). \tag{2.2}$$

This condition just means that for a fixed δ, $\mathscr{L}$ is indeed a loss function for estimating $W(\theta, \delta)$.

THEOREM 1. *Any loss function $\mathscr{L}$ of the form* (2.1) *with differentiable functions a and b, such that $\gamma a(\gamma) \to 0$ as $\gamma \to 0$, and for which* (2.2) *holds has the form*

$$\mathscr{L}(\theta; \delta, \gamma) = f'(\gamma)W(\theta, \delta) - f'(\gamma)\gamma + f(\gamma) + c. \tag{2.3}$$

Here c is a constant and f is an increasing concave function,

$$f'(\gamma) > 0, \qquad f''(\gamma) < 0, \qquad \gamma f'(\gamma) \to 0 \quad \textit{as } \gamma \to 0.$$

PROOF. Condition (2.2) implies that

$$a'(\gamma)\gamma = -b'(\gamma).$$

Put $a(\gamma) = f'(\gamma)$. Then

$$b(\gamma) - b(0) = -\int_0^\gamma f''(t)t\,dt = -\gamma f'(\gamma) + \int_0^\gamma f'(t)\,dt = -\gamma f'(\gamma) + f(\gamma) - f(0)$$

and representation (2.3) obtains with $c = b(0) - f(0)$.

The function $\mathscr{L}$ of the form (2.3) has minimum at $W(\theta, d) = w$ if and only if for all γ,

$$(w - \gamma)f'(\gamma) \geq f(w) - f(\gamma),$$

which implies the concavity of f. □

Henceforth we assume that the utility function (2.3) is normalized by the condition $c = 0$. Rukhin (1985) studied the loss (2.3) in the particular case $f(\gamma) = \gamma^{1/2}$.

We give now a statistical interpretation of the function f. Assume that

$$E_\theta W(\theta, \delta(x)) = \gamma,$$

i.e., that the risk of δ is constant. If one uses an estimator γ_1 of the loss,

$$\gamma_1(x) \equiv \gamma,$$

then the risk $\mathscr{R}(\theta; \delta, \gamma_1)$ of the pair (δ, γ_1) is

$$\mathscr{R}(\theta; \delta, \gamma_1) = E_\theta \mathscr{L}(\theta; \delta, \gamma_1) = f'(\gamma)\gamma - f'(\gamma)\gamma + f(\gamma) = f(\gamma).$$

In other terms $f(\gamma)$ is the value of the combined risk of (δ, γ) if δ has constant risk equal to γ.

Notice also that the differentiability condition in Theorem 1 can be considerably relaxed. In fact it suffices to assume lower semicontinuity of the functions a and b in (2.1).

The most important feature of loss functions (2.3) is that the Bayes procedure (δ_B, γ_B) has the following form. The rule δ_B is just the Bayes decision corresponding to the loss $W(\theta, \delta)$ and γ_B coincides with the posterior loss,

$$\gamma_B(x) = E\{W(\theta, \delta_B)|x\}. \tag{2.4}$$

Loss functions possessing this property are called proper scoring rules [cf. Savage (1971) and Hogarth (1975)]. From the frequentist point of view the use of Bayes procedures and (some of) their limits is motivated by the admissibility argument, and the combined loss function $\mathscr{L}$ allows decision-theoretical comparison of different pairs (δ, γ). In particular the corresponding risk function can be used to define a natural notion of admissibility.

As an example let us consider the situation where θ can be estimated correctly with positive probability. Namely let $D_\theta = \{x\colon \delta(x) \neq \theta\}$ and assume that $P_\theta(D_\theta) = \gamma$, $0 < \gamma < 1$, for all θ. Also assume that for $\bigcup_\theta D_\theta = C$, one has

$$P_\theta(C) = w, \qquad 0 < w < 1.$$

Under zero–one loss the risk of δ is constant, but the constant estimator $\gamma_1(x) \equiv \gamma$ is inadmissible under any loss function (2.3). Indeed let

$$\begin{aligned} \gamma_0(x) &= \gamma/w, \qquad x \in C, \\ &= 0, \quad \text{otherwise}. \end{aligned}$$

Then as we already noticed,

$$\mathscr{R}(\theta;\delta,\gamma_1)=f(\gamma)$$

and

$$E_\theta f'(\gamma_0)W(\theta,\delta)=f'(\gamma/w)\gamma,$$

$$E_\theta[f'(\gamma_0)\gamma_0-f(\gamma_0)]=f'(\gamma/w)\gamma-f(\gamma/w)w-f(0)(1-w).$$

Therefore for all θ, because of the concavity of f,

$$\mathscr{R}(\theta;\delta,\gamma_0)=wf(\gamma/w)+f(0)(1-w)<f(\gamma)=\mathscr{R}(\theta;\delta,\gamma_1).$$

A particular case of this situation happens in an example considered by Berger (1985a, b). Let $x=(x_1,x_2)$ with independent x_1 and x_2, such that

$$P_\theta(x_i=\theta-1)=1-P_\theta(x_i=\theta+1)=p.$$

Consider the procedure

$$\begin{aligned}\delta(x)&=(x_1+x_2)/2\quad\text{if }|x_1-x_2|=2\\&=x_1+1\quad\text{if }x_1=x_2.\end{aligned}$$

Then

$$D_\theta=\{x\colon x_1=x_2=\theta+1\},\qquad P_\theta(D_\theta)=(1-p)^2,$$

$$C=\{x\colon x_1=x_2\},\qquad P_\theta(C)=p^2+(1-p)^2=w.$$

In this example constant estimator $\gamma_1(x)\equiv(1-p)^2$ is clearly unreasonable. Indeed if $|x_1-x_2|=2$ one is certain that $\delta(x)=\theta$, while if $x_1=x_2$ the exact value of θ is unknown. The inadmissibility of γ_1 under (2.3) should be contrasted with its admissibility under loss $(\gamma-W(\theta,\delta))^2$ [see Berger (1985a)], where W is zero–one loss function and δ is fixed.

The important feature of this and similar examples can be extracted as the following simple result.

THEOREM 2. *Assume that for some prior distribution the posterior loss,*

$$\gamma_0(x)=E\{W(\theta,\delta)|x\},$$

possesses the following property: For all θ,

$$E_\theta f'(\gamma_0)[W(\theta,\delta)-\gamma_0]\le 0 \tag{2.5}$$

and

$$\sup_\theta E_\theta\gamma_0=\bar\gamma<\infty. \tag{2.6}$$

Then γ_0 improves upon any constant estimator $\gamma_1(x)\equiv\gamma$ with $\gamma\ge\bar\gamma$ in the sense of (2.3).

PROOF. Under condition (2.5),

$$\mathscr{R}(\theta;\delta,\gamma_1)>\mathscr{R}(\theta;\delta,\gamma_0)$$

if

$$E_\theta f(\gamma_0) < f(\gamma). \tag{2.7}$$

Because of Jensen's inequality,

$$E_\theta f(\gamma_0) < f(E_\theta \gamma_0) \geq f(\bar{\gamma})$$

so that (2.7) follows because of the monotonicity of f.

In another example considered by Kiefer (1976) and Berger (1985b) where conditions (2.5) and (2.6) hold, x is a normal random variable with mean θ and unit variance. Assume that the hypothesis H_0: $\theta \leq -\varepsilon$ (ε fixed positive) has to be tested against H_1: $\theta > \varepsilon$.

Consider the test δ which rejects H_0 if $x > 0$. Under zero–one loss,

$$E_\theta W(\theta, \delta) = P_{|\theta|}(x < 0) = \Phi(-|\theta|) \leq \Phi(-\varepsilon).$$

If, say $\varepsilon = 2$, $\Phi(-2) = 0.0228$, but it seems to be rather unreasonable to state that H_0 is rejected with error probability not exceeding 0.0228 when the observed value of x is 0.

Motivated by the fact that δ is a Bayes test against the prior distribution assigning equal mass to $\theta = -\varepsilon$ and $\theta = \varepsilon$, we put

$$\gamma_0(x) = E\{W(\theta, \delta)|x\} = 1/(1 + e^{2\varepsilon|x|}).$$

An easy calculation shows that

$$E_\theta f'(\gamma_0)(W(\theta, \delta) - \gamma_0)$$
$$= (2\pi)^{-1/2} \int_0^\infty f'(\gamma_0(x))\gamma_0(x)[e^{2\varepsilon x} - e^{2|\theta|x}]e^{-(x+|\theta|)^2/2}\,dx \leq 0$$

and

$$E_\theta \gamma_0 \leq E_\varepsilon \gamma_0 = \Phi(-\varepsilon).$$

Thus (2.5) and (2.6) are met and any "silly" constant estimator $\gamma(x) \equiv \gamma$, $\gamma \geq \Phi(-\varepsilon)$, is inadmissible. □

3. Admissibility criterion. In this section it is assumed that the sample space $\mathscr{X}$ is Euclidean space, $\mathscr{X} = \mathbb{R}^n$, the decision space $\mathscr{D}$ is an open convex subset of $\mathbb{R}^m$ and the parameter space Θ is a separable locally compact metric space. We make the measurability and regularity assumptions of (i)–(v) of Theorem 1 of Farrell (1968). We suppose that W is a continuous loss function over $\Theta \times \mathscr{D}$ which is strictly convex in d and that there exist positive densities $p_\theta(x)$ with respect to some measure μ.

THEOREM 3. *Under assumptions* (i)–(v) *of Farrell* (1968) *the pair* (δ_0, γ_0) *is an admissible procedure under loss* (2.3) *if and only if there exists a sequence* G_k, $k = 1, 2, \dots$, *of finite measures over* Θ *such that for any compact subset E* *of* Θ, $G_k(E) \geq 1$, $k = 1, 2, \ldots$, $\sup_k G_k(C) < \infty$ *for compact* C *and*

$$\int_{\mathscr{X}}\int_{\Theta}[W(\theta,\delta_0) - W(\theta,\delta_k)]\,f'(\gamma_0)p_\theta(x)\,d\mu(x)\,dG_k(\theta) \to 0, \tag{3.1}$$

$$\int_{\mathscr{X}}\int_{\Theta}[f(\gamma_0) - f(\gamma_k) - (\gamma_0 - \gamma_k)f'(\gamma_0)]\,p_\theta(x)\,d\mu(x)\,dG_k(\theta) \to 0, \tag{3.2}$$

where δ_k, γ_k *are Bayes rules against* G_k.

In particular if (δ_0, γ_0) *is admissible under the loss* (2.3), *then* δ_0 *is admissible under the loss* $W(\theta, \delta)f'(\gamma_0)$ *and* γ_0 *is admissible under the loss* $f'(\gamma)E_\theta W(\theta, \delta_0) - f'(\gamma)\gamma + f(\gamma)$.

Proof. Define for any integrable function $h(x, \theta)$, $\mathscr{E}_k h(x, \theta) = \int\int h(x, \theta)p_\theta(x)\,d\mu(x)\,dG_k(\theta)$. According to Theorem 1 of Farrell (1968), (δ_0, γ_0) is admissible if and only if

$$\rho_k = \mathscr{E}_k\{\mathscr{L}(\theta; \delta_0(x), \gamma_0(x)) - \mathscr{L}(\theta; \delta_k(x), \gamma_k(x))\} \to 0. \tag{3.3}$$

Because of the property of iterated expected value,

$$\begin{aligned}
\rho_k &= \mathscr{E}_k\{f'(\gamma_0)[W(\theta,\delta_0) - W(\theta,\delta_k)] + [f'(\gamma_0) - f'(\gamma_k)]W(\theta,\delta_k) \\
&\qquad -\gamma_0 f'(\gamma_0) + \gamma_k f(\gamma_k) + f(\gamma_0) - f(\gamma_k)\} \\
&= \mathscr{E}_k\{f'(\gamma_0)[W(\theta,\delta_0) - W(\theta,\delta_k)] + [f'(\gamma_0) - f'(\gamma_k)]\gamma_k \\
&\qquad -\gamma_0 f'(\gamma_0) + \gamma_k f'(\gamma_k) + f(\gamma_0) - f(\gamma_k)\} \\
&= \mathscr{E}_k f'(\gamma_0)[W(\theta,\delta_0) - W(\theta,\delta_k)] \\
&\qquad + \mathscr{E}_k[f(\gamma_0) - f(\gamma_k) - (\gamma_0 - \gamma_k)f'(\gamma_0)].
\end{aligned} \tag{3.4}$$

Since both terms in the right-hand side of (3.4) are nonnegative, (3.3) holds if and only if (3.1) and (3.2) are valid, which completes the proof. □

Clearly (3.1) means the admissibility of δ_0 as an estimator of θ under rescaled loss function $L_0(\theta, \delta) = W(\theta, \delta)f'(\gamma_0)$ (which involves the observation x).

Formula (3.2) means that γ_0 is an admissible estimator of the parametric function $\varphi(\theta) = E_\theta W(\theta, \delta_0)$ under loss function $L_1(\theta, \gamma) = \varphi(\theta)f'(\gamma) - \gamma f'(\gamma) + f(\gamma)$. Indeed an easy calculation shows that

$$\mathscr{E}_k\{L_1(\theta,\gamma) - L_1(\theta,\gamma_k)\} = \mathscr{E}_k\{f(\gamma) - f(\gamma_k) + f'(\gamma)(\gamma_k - \gamma)\},$$

and the conclusion follows from Farrell's theorem.

Notice that separate admissibility of δ_0 under $L_0(\theta, \delta)$ and of γ_0 under $L_1(\theta, \gamma)$ does not imply the admissibility of (δ_0, γ_0) under $L(\theta, \delta, \gamma)$.

It is known [cf. Berger and Srinivasan (1978)] that if $p_\theta(x) = \beta(\theta)\exp\{\theta' x\}$ and $W(\theta, d) = \|\theta - d\|^2$, then any admissible estimator has the form

$$\delta(x) = \nabla \log \hat{G}(x),$$

where G is a σ-finite measure with support in the closure of the natural parameter space and

$$\hat{G}(x) = \int \exp\{\theta x\}\, dG(\theta) \tag{3.5}$$

is the Laplace transform of G.

It is easy to see that the corresponding loss estimator has the form,

$$\gamma(x) = \nabla^2 \log \hat{G}(x) = \sum \frac{\partial^2}{\partial x_i^2} \log \hat{G}(x). \tag{3.6}$$

A modification of the proof of Theorem 2.1 of Berger and Srinivasan (1978) shows that any admissible pair δ, γ under (2.3) has the form (3.5) and (3.6) for some σ-finite measure supported by the closure of the natural parameter space. Formula (3.6) is convenient for the calculation of admissible loss estimators in an exponential family.

Notice that the admissibility notion associated with the loss (2.3) is more conventional and convenient to work with than the admissibility definitions owing to Kiefer (1975) and Brown (1978) in the problems of conditional confidence estimators.

Our concluding remark is that in the case of randomized procedures $\delta = \delta_x$, the loss $W(\theta, d)$ over $\Theta \times \mathscr{D}$ should be replaced by a new loss,

$$\tilde{W}(\theta, \delta) = \int_{\mathscr{D}} W(\theta, t)\, d\delta_x(t),$$

which is defined over $\Theta \times \mathscr{P}(\mathscr{D})$, where $\mathscr{P}(\mathscr{D})$ is the collection of all probability measures over $\mathscr{D}$.

REFERENCES

Aitchison, J. and Dunsmore, I. R. (1968). Linear-loss interval estimation of location and scale parameters. *Biometrika* **55** 141–148.

Berger, J. (1985a). In defense of the likelihood principle: Axiomatics and coherency. In *Bayesian Statistics II* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 33–65. North-Holland, Amsterdam.

Berger, J. (1985b). The frequentist viewpoint and conditioning. In *Proc. of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. Le Cam and R. Olshen, eds.) **1** 15–44. Wadsworth, Monterey, Calif.

Berger, J. (1985c). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York.

Berger, J. and Srinivasan, C. (1978). Generalized Bayes estimators in multivariate problems. *Ann. Statist.* **6** 783–801.

Brown, L. (1978). A contribution to Kiefer's theory of conditional confidence procedures. *Ann. Statist.* **6** 59–71.

Cohen, A. and Strawderman, W. (1973). Admissible confidence interval and point estimation for translation or scale parameters. *Ann. Statist.* **1** 545–550.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia.

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy (with discussion). *Statist. Sci.* **1** 54–77.

Farrell, R. H. (1968). On a necessary and sufficient condition for admissibility of estimators when strictly convex loss is used. *Ann. Math. Statist.* **39** 23–28.

HOGARTH, R. M. (1975). Cognitive processes and the assessment of subjective probability distributions (with discussion). *J. Amer. Statist. Assoc.* **70** 271–294.

KIEFER, J. (1975). Conditional confidence approach in multi-decision problems. In *Multivariate Analysis IV* (P. R. Krishnaiah, ed.) 143–158. Academic, New York.

KIEFER, J. (1976). Admissibility of conditional confidence procedures. *Ann. Statist.* 4 836–865.

KIEFER, J. (1977). Conditional confidence statements and confidence estimators. *J. Amer. Statist. Assoc.* **72** 789–827.

LEHMANN, E. L. (1959). *Testing Statistical Hypotheses*. Wiley, New York.

PRATT, J. W. (1961). Length of confidence intervals. *J. Amer. Statist. Assoc.* **56** 549–567.

RUKHIN, A. L. (1985). Estimated loss and admissible loss estimators. Technical Report 85–26, Dept. Statistics, Purdue Univ.

SANDVED, E. (1968). Ancillary statistics and estimation of the loss in estimation problems. *Ann. Math. Statist.* **39** 1755–1758.

SAVAGE, L. J. (1954). *The Foundations of Statistics*. Wiley, New York.

SAVAGE, L. J. (1971). Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* **66** 783–801.

WINKLER, R. L. (1972). Decision-theoretic approach to interval estimation. *J. Amer. Statist. Assoc.* **67** 187–191.

DEPARTMENT OF MATHEMATICS AND STATISTICS
LEDERLE GRADUATE RESEARCH CENTER
UNIVERSITY OF MASSACHUSETTS
AMHERST, MASSACHUSETTS 01003