ADDITIONAL REFERENCES

BERGER, J. O. (1985a). In defense of the Likelihood Principle: Axiomatics and Coherency. In *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 33–66. North-Holland, Amsterdam.

BERGER, J. O. (1988). An alternative: The estimated confidence approach. Discussion of "Conditionally acceptable frequentist solutions" by G. Casella. In *Statistical Decision Theory and Related Topics IV* (S. S. Gupta and J. O. Berger, eds.) **1** 85–90. Springer, New York.

Comment

Lawrence D. Brown

It is a pleasure and an embarrassment to read a historical story in which one plays an integral role. From my perspective the story has been accurately related, but I do have some miscellaneous comments to make which are related to the general topic.

THE LOSS FUNCTION

The point estimation segment of this article deals exclusively with the loss function (1.5)—i.e., $L(\delta, \sigma^2) = ((\delta/\sigma^2) - 1)^2$. Although this loss is relatively easy to handle analytically it seems somewhat inappropriate for a broad range of applications. Let me repeat informally some thoughts I tried to convey formally in Brown (1968).

Strictly from a qualitative point of view, the loss (1.5) is very skewed. Note that $\lim_{\delta\to\infty} L(\delta, \sigma^2) = \infty$ but $\lim_{\delta\to 0} L(\delta, \sigma^2) = 1$. Hence overestimation of σ^2 is much more severely penalized than underestimation. Furthermore, the best invariant estimator for this loss is $S^2/(\nu + 2)$, which is smaller than the maximum likelihood value of $S^2/(\nu + 1)$, or the intuitively appealing and best unbiased estimator, which is S^2/ν. One rationalization for this discrepancy could be that the intuition supporting use of S^2/ν is in error; that (1.5) is the actual loss and that therefore $S^2/(\nu + 2)$ is to be preferred to S^2/ν (and Brewster and Zidek's (2.20) is then to be preferred to $S^2/(\nu + 2)$).

However, another interpretation is possible. Note that historically use of S^2/ν was proposed and, presumably, found generally satisfactory without elicitation of or reference to a specific loss function. If indeed S^2/ν is satisfactory among invariant procedures perhaps that is because it matches the actual (but subconscious) loss function measuring the experimenters' preferences. Thus one asks, "For what loss is the best unbiased estimator, S^2/ν, also best invariant?".

Stein (1964) found a loss function for which S^2/ν is best invariant. It is

$$(1) \qquad L_S(\delta, \sigma^2) = \delta/\sigma^2 - \ln(\delta/\sigma^2) - 1.$$

Note that $L_S(\delta, \sigma^2) \ge 0$ and attains the value 0 uniquely at $\delta = \sigma^2$. Also, $L_S(\delta, \sigma^2)$ is strictly convex in δ and $\lim_{\delta\to 0} L_S(\delta, \sigma^2) = \lim_{\delta\to\infty}(L_S(\delta, \sigma^2) = \infty$. Thus L_S has a number of pleasing qualitative properties.

In Brown (1968) more was established about L_S. It was shown that for virtually any problem of estimating a single scale parameter the best unbiased estimator is also best invariant for this loss, and the loss function L_S is *the only loss function possessing this global property* (up to affine transformations, which do not affect admissibility). Thus a belief in the suitability among invariant estimators of the best unbiased estimator is *equivalent to* a belief in the suitability of L_S. In summary, my own feeling is that the loss L_S is the most appropriate for *general* studies of estimation of scale parameters. (Of course, other loss functions may be appropriate in *specific* applications.)

The story related for loss (1.5) by Maatta and Casella applies equally to the loss L_S. The analog of Stein's estimator, (2.4), is $\tilde\delta(\bar X, S^2) = \tilde\phi(Z^2)S^2$, where

$$(2) \qquad \tilde\phi(Z^2) = \min\left(\frac{1}{\nu}, \frac{1 + Z^2}{\nu + 1}\right).$$

Under L_S this estimator dominates the usual S^2/ν.

Loss L_S is explicitly considered in Brown (1968) where it is shown (as in (2.16)) that the choice

$$(3) \qquad \tilde\phi^*(Z^2) = \begin{cases} \tilde c_{0,1}(r^2,) & \text{if } Z^2 \le r^2, \\ 1/\nu & \text{if } Z^2 > r^2 \end{cases}$$

yields an estimator better than S^2/ν when

$$\tilde c_{0,1}(r^2) = 1/E_{0,1}(S^2 \mid Z^2 \le r^2).$$

The algorithm of Brewster and Zidek then applies, and shows the estimator with

$$(4) \qquad \tilde\phi^{**}(Z^2) = \tilde c_{0,1}(Z^2)$$

Lawrence D. Brown is Professor, Department of Mathematics, and Director of the Cornell Statistics Center, Cornell University, Ithaca, New York 14853-7901.

to yield a further improvement on (3). As with (2.20) this is (probably) as far as one can go. The estimator involving (4) is again generalized Bayes with respect to the prior (2.23). There are strong reasons to suspect that, like (2.20), (4) yields an estimator which is admissible, although a detailed proof of such an assertion remains to be worked out. See Dey and Srinivasan (1985) for some more recent work related to the loss L_S.

MINIMUM LENGTH INTERVALS

One assertion in the paper with which I do disagree is that which says, "It is generally accepted that length is the overriding criterion when interval estimation is concerned." There are situations where the minimum length criterion leads to patently unacceptable procedures. For example, let X be exponential with precision θ. (I.e., $2X$ is χ^2 with $n = 2$ degrees-of-freedom and scale parameter $\sigma^2 = 1/\theta$.) Then the minimum length (invariant) interval for θ is $\tilde{I}_{\text{ML}} = \{\theta: 0 \le \theta \le c_0/X\}$ where $c_0 = -\ln \alpha$. These intervals are of course what result from inverting the U.M.P. tests of the *one-sided* hypotheses $H_0: \theta \le \theta_0$ versus $H_1: \theta > \theta_0$. The intervals are as strongly biased as possible in the sense that $P_{\theta_0}(\theta_1 \in \tilde{I}) > 1 - \alpha$ for $\theta_1 < \theta_0$; and, in fact, $P_{\theta_0}(\theta_1 \in I) < P_{\theta_0}(\theta_1 \in \tilde{I})$ for any system, I, of intervals having coverage probability at most $(1 - \alpha)$. The best unbiased interval for this problem is a much more appealing general answer. It is $I_{\text{SU}} = \{\theta: (a/X) \le \theta \le (b/X)\}$ with a, b determined by (3.4).

Casella, Hwang and Robert (1989) in a very interesting recent paper also make the point that it may be desirable to consider minimizing the expectation of some (concave) function of the length rather than the length itself.

UNBIASED INTERVALS AND LIKELIHOOD RATIO TESTS

Before discussing in general what I think are desirable criteria for confidence intervals, I would like to briefly digress and mention one other appealing feature of the intervals I_{SU} defined in (3.4). Consider the generalized likelihood ratio test of $H_0: \sigma^2 = \sigma_0^2$ versus $H_1: \sigma^2 \ne \sigma_0^2$. This test rejects for large values of

$$\Lambda = \frac{\sup_{\sigma^2}\sigma^{-2}f_\nu(S^2/\sigma^2)}{\sigma_0^{-2}f_\nu(S^2/\sigma_0^2)} = \frac{k(\nu)}{f_{\nu+2}(S^2/\sigma_0^2)}$$

where $k(\nu)$ is independent of S^2, σ_0^2. When these tests are inverted to form confidence intervals one gets

$$I_{\text{LRT}} = \{\sigma^2: aS^2 \le \sigma^2 \le bS^2\}$$

where a, b satisfy (3.4)—i.e., $f_{\nu+2}(1/a) = f_{\nu+2}(1/b)$. Thus

$$(5) \qquad I_{\text{LRT}} = I_{\text{SU}}.$$

The identity of likelihood ratio confidence intervals and best unbiased ones does not hold in all situations. However, it does hold in a number of normal theory problems involving scaled chi-squared or scaled F-distributions, including the problem discussed above.

CRITERIA FOR CONFIDENCE INTERVALS

The source of the judgment that shorter (in some sense) confidence intervals are better is probably the intuitive feeling that shorter intervals have generally lower overall probability of covering false values. That is, the criterion actually operating is the classical one which says that one procedure, I_1, say, dominates another, I_2, if

$$(6) \qquad P_{\nu,\sigma^2}\{\sigma^2 \in I_1\} \ge P_{\mu,\sigma^2}\{\sigma^2 \in I_2\}$$

and

$$(7) \qquad P_{\nu,\sigma^2}\{\tau^2 \in I_1\} \le P_{\mu,\sigma^2}\{\tau^2 \in I_2\} \quad \text{for } \tau^2 \ne \sigma^2,$$

with strict inequality for some μ, σ^2. (An alternate valid set of criteria would be just (7) plus the validity criterion $1 - \alpha_1 \le 1 - \alpha_2$ where

$$(8) \qquad 1 - \alpha_1 = \inf_{\mu,\sigma^2} P_{\mu,\sigma^2}\{\sigma^2 \in I_i\}.)$$

The standard procedures I_{ML}, I_{SU} or I_{ET} are all admissible (i.e., cannot be dominated) in the sense of (6) and (7) or of (7) and (8). This is because they are formed as inversions of admissible families of tests. But this by itself does not mean there do not exist procedures which are mainly preferable to the standard ones. For example, it may be that the intervals of Goutis (3.18) generally perform well in relation to I_{ML}. By construction they dominate I_{ML} with respect to (6), and it could possibly be that (7) is satisfied for a wide range of values of μ, σ^2; and perhaps even that when (7) fails it does so only by a numerically insignificant margin. Whether this is so appears to require further investigation, perhaps in the nature of a numerical study.

It seems plausible that the argument in Proskin (1985) can be adapted to prove admissibility in the sense of coverage probability and expected length of the procedures of Goutis (3.18) and of Shorrock (3.10). If so then Cohen and Strawderman (1973) shows that both (3.18) and (3.10) are themselves almost admissible (and, even, admissible) in the sense of (6) and (7).

INDIFFERENCE ZONES

As a digression, let me note that there are situations in which standard procedures can be qualitatively improved as suggested above. This improvement can sometimes even be formalized mathematically through introduction of indifference zones.

For example, consider the standard one sided t-intervals: $I_{1t} = \{\mu\colon \mu \le \bar{X} + cS\}$. Let $\varepsilon > 0$. Brown and Sackrowitz (1984) construct improved intervals, I'_{1t} say, such that

$$(9)\qquad P_{\mu,\sigma^2}\{\nu \in I'_{1t}\} > P_{\mu,\sigma^2}\{\nu \in I_{1t}\} \quad \text{for } \nu \le \mu$$

and

$$(10)\quad P_{\mu,\sigma^2}\{\nu \in I'_{1t}\} < P_{\mu,\sigma^2}\{\nu \in I_{1t}\} \quad \text{whenever } \nu - \mu \ge \varepsilon.$$

The region $\{\mu, \sigma^2, \nu\colon \mu < \nu < \mu + \varepsilon\}$ is an indifference region. (Larger values of ε enable greater inequality in (9), (10).) For the standard two sided t-intervals, $I_{2t} = \{\mu\colon | \mu - \bar{X} | \le cS\}$, I believe there should be a similar dominance result when the indifference zone is of the form $\{\mu, \sigma^2, \nu\colon 0 < | \mu - \nu | < \varepsilon, \sigma^2 > \sigma_0^2\}$ for any prechosen constants ε, σ_0^2. (It is shown in Brown and Sackrowitz that I_{2t} cannot be dominated when the indifference zone is just $\{\mu, \sigma^2, \nu\colon 0 < | \mu - \nu | < \varepsilon\}$.)

The above considerations lead to the question of whether the standard intervals I_{ML}, I_{SU}, or I_{ET} can be improved on in this sense when there is an indifference zone of the form $\{\mu, \sigma^2, \tau^2\colon | \tau^2 - \sigma^2 | > \varepsilon\}$ or of the form $\{\mu, \sigma^2, \tau^2\colon | \ln \tau^2 - \ln \sigma^2 | > \varepsilon\}$. The answer is no, as can be shown by using the Bayes representation of Kiefer and Schwartz (1965) for the associated family of tests. (This representation can be thought of here as an extension of the representation in Lehmann and Stein, 1948.)

RELEVANT AND SEMI-RELEVANT SUBSETS

Rejection of the null hypothesis $H_0\colon \mu = 0$ at suitable levels is a negatively biased semi-relevant subset for the two sided t-interval, I_{2t}. This means (for suitable k)

$$(11)\qquad P_{\mu,\sigma^2}(\mu \in I_{2t} \mid | \bar{X}/S | > k) < 1 - \alpha.$$

Property (11) is somewhat disturbing for two reasons.

First, the direction of the inequality means that the usual confidence claim—i.e., $P_{\mu,\sigma^2}(\mu \in I_{2t}) \ge 1 - \alpha$—is conditionally false. (For suitable k, $P_{0,\sigma^2}(\mu \in I_{2t} \mid | \bar{X}/S | > k) = 0$ so the inequality in (11) can be numerically significant.) On the other hand, if the conditioning set were positively biased semi-relevant this would not be disturbing since the usual confidence claim would still be conditionally valid. If the set were positively biased and relevant, as is the set $\{(\bar{X}/S)\colon | \bar{X}/S | \le k\}$, this would be worth noting but still might not be judged disturbing; it would mean that the confidence claim could be conditionally adjusted upward to $1 - \alpha + \varepsilon$.

Second, and perhaps more important here, is the nature of the conditioning set. Some statisticians are in the habit of supplying confidence intervals (or at least of paying attention to them) only when $H_0\colon \mu = 0$ is rejected. For such statisticians the conditional calculation becomes an unconditional one. That is, the intervals they produce and pay attention to have probability of coverage $<1 - \alpha$, the nominal value.

In summary, negatively biased semi-relevant sets are disturbing when the conditioning set has a natural interpretation. Otherwise they need not be as I shall emphasize below. Positively biased semi-relevant sets are never upsetting; occasionally they can be worth noting. (For example, for the intervals of Goutis (3.18) and Shorrock (3.10) the entire sample space is positively biased semi-relevant.) Relevant subsets seem generally worth noting and, of course, negatively biased ones are much more disturbing than positively biased ones.

ONE-SIDED INTERVALS

One-sided intervals virtually always admit semi-relevant subsets. This emphasizes, I think, that the nature of a semi-relevant subset is crucial in deciding on its importance (rather than establishing that one-sided intervals are virtually never acceptable).

As a canonical example consider the case where $\sigma^2 = 1$ is known, and μ is unknown. The uniformly best one sided intervals for μ are of the form $I_1 = \{\mu\colon \mu \le \bar{X} + c\}$. Then, the set $\{\bar{X}\colon \bar{X} \le 0\}$ is negatively biased semi-relevant; consequently its complement, $\{\bar{X}\colon \bar{X} \ge 0\}$, is positively biased semi-relevant.

CONDITIONAL PROPERTIES

The authors have discussed the conditional properties of relevancy and semi-relevancy. I want merely to note that there are other conditional properties which may be of interest. For example, Robinson (1979a) discusses a concept of estimated conditional coverage which leads to an admissibility criterion somewhere between nonexistence of relevant subsets and nonexistence of semi-relevant subsets. See Kiefer (1977) for various other ideas including a different discussion of estimated conditional coverage. These ideas can also be supplemented by those of guaranteed conditional coverage, or frequentist validity, as presented, for example, in Brown (1978) or Berger (1985b).

ACKNOWLEDGMENT

Preparation was supported in part by NSF Grant DMS-88-09016.

ADDITIONAL REFERENCES

BERGER, J. O. (1985b). The frequentist viewpoint and conditioning. *Proc. Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. M. LeCam and R. A. Olshen, eds.) **1** 15–44. Wadsworth, Monterey, Calif.

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

BROWN, L. D. and SACKROWITZ, H. (1984). An alternative to Student's t test for problems with indifference zones. *Ann. Statist.* **12** 451–469.

CASELLA, G., HWANG, J. T. and ROBERT, C. (1989). Loss functions for set estimation. Technical Report, Cornell Statistics Center.

DEY, D. K. and SRINIVASAN, C. (1985). Estimation of a covariance matrix under Stein's loss. *Ann. Statist.* **13** 1581–1591.

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

KIEFER, J. and SCHWARTZ, R. (1965). Admissible Bayes character of T^2-, R^2-, and other fully invariant tests for classical multivariate normal problems. *Ann. Math. Statist.* **36** 747–770.

LEHMANN, E. L. and STEIN, C. (1948). Most powerful tests of composite hypotheses. *Ann. Math. Statist.* **19** 495–516.

Comment

Arthur Cohen

A historical perspective on one of the more fascinating and intriguing theoretical results of statistics is most welcome. I have some comments that are concerned with rounding out the story and generalizations.

If one requires an estimator which is both location invariant and scale equivariant, then the best equivariant estimator of σ^2 with respect to squared error loss is $S^2/(n+1)$, and of course it is admissible within its class. For the confidence interval problem, if the vector loss L_1 = (0–1 for correct coverage or not, length) is replaced by the vector loss L_2 = (0–1 for correct coverage or not, 1–0 for covering false values or not), then the "usual" confidence interval is admissible. This latter fact follows from the duality of hypothesis testing and confidence intervals. In this problem and in several other interesting problems, the following pattern holds for the "usual" procedure: admissible as a test and hence admissible as a confidence interval for the vector loss L_2; inadmissible as a confidence interval for the vector loss L_1 and inadmissible as a point estimator for squared error (or other) loss. Table 1 indicates some problems where this pattern holds. A stands for "admissible" and I for "inadmissible." For the problem of estimating the normal mean vector see Stein (1956) and Brown (1966). For the common mean problem see Brown and Cohen (1974) and Cohen and Sackrowitz (1977). For the normal quantile problem see Zidek (1971), and for the Poisson problem see Clevenson and Zidek (1975).

There is a substantial amount of work in decision theory on estimating a normal covariance matrix or generalized variance. Again the "usual" estimators are inadmissible for reasonable loss functions. In some cases, the sample mean can be used as in the univariate case to get "help." This is the situation in papers by Sinha and Ghosh (1987) and Sarkar (1989). In other cases there is a "Stein" or dimensional effect and improvements can be made even without using information from the sample mean (see the survey paper of Lin and Perlman, 1985).

The statement in the paper that Stein knew his estimator was not admissible is a bit confusing. Stein may have speculated that the generalized Bayes estimators form a complete class as is the case of some one parameter exponential family models (see Sacks, 1963). The basis of the conjecture then is that, since it cannot be generalized Bayes because it lacks smoothness properties, it is inadmissible. As it turns out Stein's estimator is easily beaten and that is why it is inadmissible. It is not known whether the class of generalized Bayes estimators for problems with unbounded nuisance parameters is a complete class except in isolated examples where it is not true.

Although Brown (1968) is already referenced, it is important to note that his paper contains many results

TABLE 1
Admissibility status of "usual" procedure

Problem	Type of inference			
	Confidence set			Point estimation squared error loss
	Testing	Loss L_2	Loss L_1	
Normal variance	A	A	I	I
Normal mean vector of dimension 3 or more	A	A	I	I
Common mean of two independent normals	A	A	I	I
Normal quantile	A	A		I
Independent Poisson parameters of dimension 2(3)	A	A		I

Arthur Cohen is Professor, Department of Statistics, Hill Center, Busch Campus, Rutgers University, New Brunswick, New Jersey 08903.