Comment

James O. Berger

The article is an extremely lucid discussion of the variance estimation problem, highlighting the major intertwining theoretical developments. Numerous interesting conceptual issues are raised in the article but, before turning to these, the issue of importance to practitioners needs further elaboration.

The authors finish Section 2 by saying "the maximum relevant improvement of Rukhin's estimators ... is only 4%, suggesting that there would be very little practical benefit associated with these improved estimators .... However, as we shall see in Section 5, there are interesting cases where substantial improvement is possible." Turning to Section 5, we see (in Table 1) only 2%, 3%, at most 5.3%, improvement in confidence interval length. In what way are these improvements more "substantial" than Rukhin's, especially since they are achieved only for the less realistic generalized linear model with common unknown σ^2?

The author's replies will likely be that it is more difficult to obtain improvement in the "length of confidence set" problem than in the "point estimation" problem, and to reiterate that any guaranteed improvement is useful, no matter how small. I would agree with these points, but feel it is important to acknowledge that only small gains are available here. Careful study of these developments by practitioners is not very cost effective as compared to, say, careful study of the more familiar "shrinkage estimation of means" literature, in which practical improvements in excess of 50% are common. Practitioners have the time to access only a small fraction of the theory and methodology that is produced, and I have always thus felt it to be imperative to carefully indicate the likely practical potential of new developments.

Conceptually, there are many fascinating developments in the paper. For instance, starting with the "hard to take seriously" problem of estimating σ^2 with quadratic loss (which I would say is hard to take seriously because decision-theoretic estimation of σ^2 is very sensitive to the loss, and there are many different reasonable losses), the authors push the developments toward the easy-to-take-seriously problem of obtaining reduced length confidence sets. Insisting, at the same time, on good conditional performance of the resulting sets is a further bold step toward realism. All in all, the development is an exemplary illustration of how to turn an originally abstract decision-theoretic benefit into a clearly concrete inferential gain.

As a Bayesian reading the paper, I naturally got excited that the most useful confidence sets were actually generalized Bayes credible sets, and that the posterior probabilities of the sets appeared plausible. For instance, a Bayesian might well be very happy with using $\langle I_S, \gamma_0(\bar{X}, S^2)\rangle$ or $\langle I_\tau, \gamma_\tau(\bar{X}, S^2)\rangle$ (see the discussion following (4.14) and (4.15)). A frequentist might also be very happy with these reports, especially if $E_{\mu,\sigma^2}\gamma(\bar{X}, S^2) <$ Coverage Probability, for all μ, σ^2, since then γ has complete frequentist justification as a report (cf. Berger, 1988). It would be interesting to know if this is so.

Let me finish by raising a few old pet peeves:

(i) Following (3.4): "The choice of interval should depend on more than just ease of calculation, which is the only favorable factor associated with I_{ET}." To the contrary, I would argue that I_{ET} (equal-tailed interval) is often most favored because (a) it conveys the uncertainty present in each tail separately, and (b) it is invariant under all monotonic transformations.

(ii) The "betting" approach to evaluating conditional behavior: I tend to view this approach with suspicion (cf. Berger, 1985a), feeling that confidence sets are constructed to communicate something, not to provide a basis for betting. If one really takes betting on confidence sets seriously, one is quickly led, via usual axiomatics, to Bayesianism (which is the only thing that makes me at all sympathetic to the betting framework).

(iii) "Positive bias" in conditional coverage is, in several places, stated to be of lesser concern. I am uncomfortable with this. I can think of many problems in which it is much more serious, in the practical context, to underestimate coverage than to overestimate coverage. We are all taught in our statistical education that it is okay to err on the conservative side, but I am not sure that this is indeed generally true in practical decision contexts where statistics is used.

James O. Berger is the Richard M. Brumfield Distinguished Professor of Statistics. His address is Department of Statistics, Purdue University, West Lafayette, Indiana 47907.

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?".

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

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

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

Note that $L_S(\delta, \sigma^2) \geq 0$ 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

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

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

$$\tilde{\phi}^*(Z^2) = \begin{cases} \tilde{c}_{0,1}(r^2,) & \text{if } Z^2 \leq r^2, \\ 1/\nu & \text{if } Z^2 > r^2 \end{cases} \tag{3}$$

yields an estimator better than S^2/ν when

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

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

$$\tilde{\phi}^{**}(Z^2) = \tilde{c}_{0,1}(Z^2) \tag{4}$$