INADMISSIBILITY OF THE USUAL ESTIMATOR FOR THE VARIANCE OF A NORMAL DISTRIBUTION WITH UNKNOWN MEAN

CHARLES STEIN

1. Introduction

It was proved by Hodges and Lehmann [3] (see also Girshick and Savage [2]) that if $Z_1, \cdots, Z_n$ are independently normally distributed with known mean ζ and unknown variance σ^2, the estimator φ_0 defined by

$$\varphi_0(Z_1, \cdots, Z_n)=\frac{1}{n+2}\sum (Z_i-\zeta)^2 \tag{1}$$

is admissible for estimating σ^2 with squared error as loss. This means that (for fixed ζ) there is no estimator φ such that, for all σ

$$E_\sigma[\varphi(Z_1, \cdots, Z_n)-\sigma^2]^2 \leqq E_\sigma[\varphi_0(Z_1, \cdots, Z_n)-\sigma^2]^2 , \tag{2}$$

with strict inequality for some σ. We shall see that, if ζ is unknown, the apparently natural estimator φ_1 defined by

$$\varphi_1(Z_1, \cdots, Z_n)=\frac{1}{n+1}\sum (Z_i-\bar{Z})^2 , \tag{3}$$

with

$$\bar{Z}=\frac{1}{n}\sum Z_i \tag{4}$$

is inadmissible, in the sense that there exists an estimator φ such that, for all ζ and σ,

$$E_{\zeta,\sigma}[\varphi(Z_1, \cdots, Z_n)-\sigma^2]^2 < E_{\zeta,\sigma}[\varphi_1(Z_1, \cdots, Z_n)-\sigma^2]^2 . \tag{5}$$

Such an estimator φ may be defined by

$$\varphi(Z_1, \cdots, Z_n)=\min \left\{\frac{1}{n+1} \sum (Z_i-\bar{Z})^2 , \frac{1}{n+2}\sum (Z_i-\zeta_0)^2\right\} \tag{6}$$

where ζ_0 is any fixed number. Of course this φ will not be admissible either since the admissible estimators are limits of Bayes solutions and so must be analytic. It is interesting to observe that the estimator φ defined by (6) may be obtained by first testing the hypothesis $\zeta=\zeta_0$ at an appropriate significance level and using the estimate (1) with $\zeta=\zeta_0$ if the hypothesis is accepted and the estimate (3) if the hypothesis is rejected. The idea of using such an estimator is not new, but it does not seem to have been observed that the risk of this estimator (6) is less than that of the usual one for all parameter values.

The main part of the paper is concerned with a more general situation involving an arbitrary number of unknown means. This includes, as a special case, the situation occurring often in the analysis of variance when we are faced with the question of whether to include an interaction term in the estimate of the variance. Even in this case the improvement obtained by using φ given by (6) rather than φ_1 seems likely to be slight, but there is some hope that a judicious choice of a Bayes solution among estimators invariant under scale may yield a substantial improvement when the number of unknown means is an appreciable proportion of the number of observations. So far, I have had no success with this approach, partly, I fear, because I find it hard to take the problem of estimating σ^2 with quadratic loss function very seriously.

2. Proof of the result announced in the title

Let $X_1, \cdots, X_n, Y_1, \cdots, Y_k$ be independently normally distributed real random variables with common unknown variance σ^2 and means given by

$$(7) \qquad EX_i=0, \qquad EY_j=\eta_j ,$$

where the η_j are unknown. We consider the problem of estimating σ^2, say, by $\hat{\sigma}^2$ with loss function L given by

$$(8) \qquad L((\eta, \sigma^2), \hat{\sigma}^2)=\left(\frac{\hat{\sigma}^2}{\sigma^2}-1\right)^2 .$$

(We have divided the squared error by σ^4 in order to make the loss function invariant under the transformations $\sigma^2\to a^2\sigma^2$, $\hat{\sigma}^2\to a^2\hat{\sigma}^2$ with $a\neq 0$. For questions of admissibility this makes no difference.) It is tempting to try to justify the estimator φ_2 given by

$$(9) \qquad \varphi_2(X_1, \cdots, X_n, Y_1, \cdots, Y_k)=\frac{1}{n+2}\sum X_i^2$$

by the following argument (see Blackwell and Girshick [1], Ch. 11). The problem is invariant under the transformations

$$X_i \to aX_i, \quad Y_j \to a(Y_j+b_j)\,, \tag{10}$$

$$\eta_j \to \eta_j + b_j, \quad \sigma^2 \to a^2\sigma^2\,, \tag{11}$$

$$\hat{\sigma}^2 \to a^2\hat{\sigma}^2\,, \tag{12}$$

so it seems reasonable to ask that the estimator φ should also be invariant under these transformations, that is

$$\begin{aligned}\varphi(aX_1, \cdots, aX_n, a(Y_1+b_1), \cdots, a(Y_k+b_k))\\ =a^2\varphi(X_1, \cdots, X_n, Y_1, \cdots, Y_k)\,.\end{aligned} \tag{13}$$

From this we easily conclude that φ must be of the form

$$\varphi(X_1, \cdots, X_n, Y_1, \cdots, Y_k)=C\sum X_i^2\,. \tag{14}$$

But it is a trivial part of the result mentioned at the beginning of the paper that, for all parameter values, the best choice of C in (14) is $1/(n+2)$.

However, we shall look at a somewhat larger class of estimators. Let

$$S=\sum_1^n X_i^2 \tag{15}$$

$$T=\sum_1^k Y_j^2\,, \tag{16}$$

let ϕ be an arbitrary positive-valued function of a positive variable, and let

$$\varphi(X_1, \cdots, X_n, Y_1, \cdots, Y_k)=\phi\left(\frac{S}{S+T}\right)(S+T)\,. \tag{17}$$

The estimator φ_2 is obtained from this by setting $\phi=\phi_2$ given by

$$\phi_2(u)=\frac{u}{n+2}\,. \tag{18}$$

The estimators (17) are those that depend only on the sufficient statistic $(\sum X_i^2, Y_1, \cdots, Y_k)$ and are invariant under orthogonal transformation of the Y_j and scale change. We shall show that, for any ϕ, the estimator φ^* given by

$$\varphi^*(X_1, \cdots, X_n, Y_1, \cdots, Y_k)=\phi^*\left(\frac{S}{S+T}\right)(S+T) \tag{19}$$

where

$$\phi^*(u)=\min\left[\phi(u), \frac{1}{n+k+2}\right] \tag{20}$$

is better than φ given by (17) in the sense that

$$(21)\qquad E_{\eta,\sigma}\left[\frac{\varphi^*(X_1,\cdots,X_n,Y_1,\cdots,Y_k)}{\sigma^2}-1\right]^2$$
$$\leqq E_{\eta,\sigma}\left[\frac{\varphi(X_1,\cdots,X_n,Y_1,\cdots,Y_k)}{\sigma^2}-1\right]^2$$

for all (η, σ) with strict inequality for all (η, σ) unless $\psi^*=\psi$.

The distribution of S/σ^2 is a central χ^2 distribution with n degrees of freedom and that of T/σ^2 is a non-central χ^2 distribution with k degrees of freedom and non-centrality parameter

$$(22)\qquad \lambda=\frac{\sum\eta_j^2}{\sigma^2}\,.$$

But (see, for example, Mann [7], p. 68), we may imagine that there is an auxiliary random variable L distributed, independent of S, according to a Poisson distribution with mean $\lambda/2$ such that T, given S and L, has a central χ^2 distribution with $k+2L$ degrees of freedom. Clearly the risk function of φ or φ^* depends only on λ so that we may assume $\sigma=1$ in proving (21). Then

$$E\,[\varphi(X_1,\cdots,X_n,Y_1,\cdots,Y_k)-1]^2=E\left[\psi\left(\frac{S}{S+T}\right)(S+T)-1\right]^2$$
$$(23)\qquad =E\,E\left\{\left[\psi\left(\frac{S}{S+T}\right)(S+T)-1\right]^2\,|\,L\right\}$$
$$=E\left\{\psi^2\left(\frac{S}{S+T}\right)E((S+T)^2\,|\,L)-2\psi\left(\frac{S}{S+T}\right)E(S+T\,|\,L)+1\right\}\,,$$

since, given L, $S/(S+T)$ is independent of $S+T$.
But

$$(24)\qquad E(S+T\,|\,L)=n+k+2L\,,$$

and

$$(25)\qquad E((S+T)^2\,|\,L)=(n+k+2L)(n+k+2L+2)\,.$$

Continuing with (23) we have

$$E[\varphi(X_1,\cdots,X_n,Y_1,\cdots,Y_k)-1]^2$$
$$(26)\qquad =E\left\{\psi^2\left(\frac{S}{S+T}\right)(n+k+2L)(n+k+2L+2)\right.$$
$$\left.-2\psi\left(\frac{S}{S+T}\right)(n+k+2L)+1\right\}$$

$$=E\left\{(n+k+2L)(n+k+2L+2)\left[\phi\left(\frac{S}{S+T}\right)-\frac{1}{n+k+2L+2}\right]^2\right.$$
$$\left.+\frac{2}{n+k+2L+2}\right\}.$$

The desired result (21) follows since

$$\left[\phi^*\left(\frac{S}{S+T}\right)-\frac{1}{n+k+2L+2}\right]^2 \leqq \left[\phi\left(\frac{S}{S+T}\right)-\frac{1}{n+k+2L+2}\right]^2 \tag{27}$$

for all $S/(S+T)$ and $L=0, 1, 2, \cdots$.

3. Concluding remarks

The result of this paper is an illustration of a rather ill-defined general technique. The problem we start with is invariant under a transformation group $\mathscr{G}$ (in the present case, the group of translations and orthogonal transformations of the Y_j and a scale change in all the variables simultaneously), and the usual procedure is best among those procedures invariant under $\mathscr{G}$. To find out whether the usual procedure is admissible it is often helpful to look for a better one in the class of procedures invariant under a sub-group $\mathscr{H}\subset\mathscr{G}$. If $\mathscr{H}$ is a normal (invariant) subgroup of $\mathscr{G}$, the problem reduced in this way will be invariant under the quotient group $\mathscr{G}/\mathscr{H}$, and unless this group is fairly complicated we shall not ordinarily find a procedure invariant under this group $\mathscr{H}$ that is better than the usual one. However, if $\mathscr{H}$ is not a normal subgroup of $\mathscr{G}$, the group will not operate on the reduced problem, and even if the reduced problem depends continuously on a single unknown real parameter, there may be a procedure invariant under $\mathscr{H}$ that is substantially better than the usual procedure. In the present case $\mathscr{H}$ consists of the orthogonal transformations of the Y_j and the scale changes and the only unknown parameter in the reduced problem is λ given by (22). Other illustrations of this technique are given in James and Stein [4] and in Stein [9]. Unlike the results of the present paper, the main results of section 2 in [4] can be seriously recommended to the practical statistician. An earlier paper of Robbins [8] contains a similar idea, although there the naive procedure is admissible but, nevertheless, poor in many situations.

One curious feature of the present problem is not present in any of the results cited above. By analogy with an argument of Stone [10] it is easy to show that, in our (unreduced) problem, the usual procedure is a pointwise limit of Bayes solutions, and, furthermore, these can be chosen so that the Bayes risks approximate the constant risk of the usual procedure. However, the usual procedure is not even a pointwise limit of Bayes procedures among those invariant under $\mathscr{H}$, since, in the notation of (17), it fails to have the property $\psi\left(\frac{S}{S+T}\right) \leq \frac{1}{n+k+2}$ possessed by all of these reduced Bayes procedures.

For some positive results on optimum properties of invariant procedures, the reader would do well to consult the papers of Kudo [6] and Kiefer [5], as well as parts of James and Stein [4].

I am indebted to Mr. J. B. Selliah for reminding me of this problem.

STANFORD UNIVERSITY

REFERENCES

[1] D. Blackwell, and M. A. Girshick, *Theory of Games and Statistical Decisions*, John Wiley and Sons, New York, 1954.

[2] M. A. Girshick, and L. J. Savage, " Bayes and minimax estimates for quadratic loss functions," *Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California, Berkeley, 1951, 53-73.

[3] J. L. Hodges, Jr., and E. L. Lehmann, "Some applications of the Cramér-Rao inequality," *Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California, Berkeley, 1951, 13-22.

[4] W. James, and C. Stein, " Estimation with quadratic loss," *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, University of California, Berkeley, I (1961), 361-379.

[5] J. Kiefer, " Invariance minimax sequential estimation and continuous time processes," *Ann. Math. Statist.*, 28 (1957), 573-601.

[6] H. Kudo, " On minimax invariant estimators of the transformation parameter," *Nat. Sci. Rep. Ochanomizu Univ.*, 6 (1955), 31-73.

[7] H. B. Mann, *Analysis and Design of Experiments*, Dover Publications, New York, 1949.

[8] H. Robbins, " Asymptotically subminimax solutions of compound statistical decision problems," *Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California, Berkeley, 1951, 131-148.

[9] C. Stein, " Multiple regression," *Contributions to Probability and Statistics Essays in Honor of Harold Hotelling*, Stanford, 1960, 424-443.

[10] M. Stone, " The posterior *t*-distribution," *Ann. Math. Statist.*, 34 (1963) 568-573.