Biometrika (1977), **64**, 1, *pp*. 29–35
Printed in Great Britain

Asymptotics for and against cross-validation

BY M. STONE
Department of Statistics and Computer Science,
University College London

SUMMARY

The asymptotic consistency of cross-validatory assessment and the asymptotic efficiency of cross-validatory choice is investigated both in some generality and also in the context of particular applications.

Some key words: Asymptotic consistency; Asymptotic efficiency; Cross-validatory choice and assessment.

1. INTRODUCTION

The question of the 'reliability' of cross-validatory assessment has been raised by A. P. Dawid in the discussion of Stone (1974*a*). Dawid concludes his brief investigation by expressing the 'worrying feeling' that cross-validatory assessment may not be consistent in any reasonable sense, leading to the qualified advice that

> one should be wary of over-glib use of cross-validatory assessments, since they may be wide of the mark.

This paper establishes that, within the context of 'one-item-out' cross-validation at least: (*a*) Dawid was broadly correct in his intuition concerning the possibility of asymptotic inconsistency, even though his analysis requires important modifications; (*b*) such inconsistency is often inevitable in the cross-validatory context and indeed, without it, we would be able to construct procedures with performances known, on general theoretical grounds, to be unattainably high; (*c*) for univariate estimation, the asymptotic inconsistency is often accompanied by poor performance of estimators involving cross-validatory choice.

For basic notation and definitions, see Stone (1974*a*, *b*).

2. ABSOLUTE ASSESSMENT

While an investigation of cross-validatory asymptotics may be envisaged that remained 'data-analytic' by postulating infinite data sequences with specified properties, it is easier to admit the convenient fictions of probability models and associated parameters. The minimum modelling we will require is:

Model assumption I. Given x, the value of y for an item (x, y) has distribution P_x; the y values of different items are independently distributed, given the x values.

To analyze the case where $L(y, \hat{y})$, the loss in predicting y by $\hat{y}$, is quadratic, we need the following.

Model assumption II. The previous assumption I with y real and

$$y_i = \eta_i + e_i, \quad \eta_i = E(y_i|x_i), \quad \operatorname{var}(e_i|x_i) = \sigma_i^2 \quad (i = 1, \ldots, n).$$

Implicit assumptions of finite expectations will be made wherever necessary.

A function of primary interest in judging the use of $\hat{y}(x; \alpha, S)$, for prediction of y for item (x, y) based on data-base $S = \{(x_i, y_i) | i = 1, \ldots, n\}$, is

$$E[L\{\tilde{y}, \hat{y}(x; \alpha, S)\} | S], \tag{2·1}$$

in which $\tilde{y}$ is distributed according to P_x. Unfortunately, no results of any generality have been found for (2·1); so we compromise by considering the definition

$$L(\alpha, S) = \frac{1}{n} \sum_{i=1}^{n} E\{L(\tilde{y}_i, \hat{y}_{-i}) | S_{-i}\}, \tag{2·2}$$

where S_{-i} denotes S with (x_i, y_i) removed, $\hat{y}_{-i}$ denotes $\hat{y}(x_i; \alpha, S_{-i})$, and where $\tilde{y}_i$ is assigned the distribution P_{x_i}.

The first connexion between $C(\alpha, S)$, our absolute assessment measure for α, given by

$$C(\alpha, S) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{y}_{-i})$$

and $L(\alpha, S)$, which is not an observable statistic, is that under assumption I

$$E[\{C(\alpha, S) - L(\alpha, S)\} | x_1, \ldots, x_n] = 0.$$

A stronger connexion, holding under fairly weak conditions, is that as $n \to \infty$

$$\operatorname{plim} \{C(\alpha, S) - L(\alpha, S)\} = 0, \tag{2·3}$$

where the limit in probability is defined for the random variables $y_1, y_2, \ldots$ corresponding to some fixed sequence of x values $x_1, x_2, \ldots$. The interpretation and occurrence of (2·3) is best illustrated under II with $L(y, \hat{y}) = (y - \hat{y})^2$, when we have

$$C(\alpha, S) = \frac{1}{n} \Sigma e_i^2 - \frac{2}{n} \Sigma e_i(\hat{y}_{-i} - \eta_i) + \frac{1}{n} \Sigma (\hat{y}_{-i} - \eta_i)^2, \tag{2·4}$$

$$L(\alpha, S) = \frac{1}{n} \Sigma \sigma_i^2 + \frac{1}{n} \Sigma (\hat{y}_{-i} - \eta_i)^2. \tag{2·5}$$

For (2·3) we require as $n \to \infty$

$$\operatorname{plim} \frac{1}{n} \Sigma (e_i^2 - \sigma_i^2) = 0, \tag{2·6}$$

$$\operatorname{plim} \frac{1}{n} \Sigma e_i(\hat{y}_{-i} - \eta_i) = 0. \tag{2·7}$$

Condition (2·6) is a weak condition on the errors $\{e_i\}$ which is satisfied under homoscedasticity for example. In considering condition (2·7) we invoke a definition.

Definition. We have mean square consistency of $\hat{y}(x; \alpha, S)$ if as $n \to \infty$

$$\operatorname{plim} \frac{1}{n} \Sigma (\hat{y}_{-i} - \eta_i)^2 = 0. \tag{2·8}$$

Schwarz's inequality then shows that (2·7) obtains under the combination of mean square consistency and

$$\Sigma e_i^2 / n = O_p(1).$$

However the occurrence of (2·8) removes any interest we may have in (2·3) as a justification for looking at $C(\alpha, S)$. For (2·8) and (2·5) imply that $L(\alpha, S)$ is asymptotically not dependent on α; while $C(\alpha, S)$ is consistent for $L(\alpha, S)$, the latter does not asymptotically depend on the choice of α; alternatively, we can describe the position by saying that $C(\alpha, S)$ is asymptotically equivalent to $\Sigma e_i^2/n$ only. We will see later, by changing from absolute to comparative assessment, that something can be salvaged from this case. Meanwhile we observe that mean square consistency is not necessary for (2·7). To see this, consider the case where $\hat{y}(x; \alpha, S)$ is the least squares predictor obtained by fitting the equation $y = \beta_1 c_1(x) + \ldots + \beta_p c_p(x)$. Appendix 1 establishes that (2·7) holds if

$$\sum_{i=1}^{n} \{\delta_i \sigma_i/(1-A_{ii})\}^2 = o(n^2), \quad \sum_{i \neq j}\sum \left\{A_{ij}\sigma_i\sigma_j\left(\frac{1}{1-A_{ii}} + \frac{1}{1-A_{jj}}\right)\right\}^2 = o(n^2), \tag{2·9}$$

where A is the symmetric matrix of orthogonal projection onto the subspace of R^n generated by $c_j = \{c_j(x_1), \ldots, c_j(x_n)\}'$ for $j = 1, \ldots, p$ and $\delta = \eta - A\eta$ is the vector deviation of the expectation of $(y_1, \ldots, y_n)'$ from this subspace. Conditions (2·9) are really quite weak since: (i) the idempotency of A implies

$$\sum_{j:j \neq i} A_{ij}^2 = A_{ii}(1-A_{ii}),$$

and (ii) $0 < A_{ii} \leqslant 1$ with $\operatorname{tr}(A) = A_{11} + \ldots + A_{nn} < p$. Moreover, for this case,

$$E\left\{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_{-i}-\eta_i)^2\right\} = E\left\{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{y}_{-i}-\eta_i-A_{ii}e_i}{1-A_{ii}}\right)^2\right\} \geqslant \frac{1}{n}\sum_{i=1}^{n}\delta_i^2,$$

so that we can usually expect not to have mean square consistency if as $n \to \infty$

$$\liminf \Sigma\delta_i^2/n > 0, \tag{2·10}$$

which condition is quite consistent with (2·9). For example, suppose x is real and we fit the equation $y = \beta_1 + \beta_2 x$ with η actually quadratic in x. With homoscedasticity in assumption II and $x_1, x_2, \ldots$ chosen evenly from the interval (0, 1), we have (2·9), (2·10) and therefore (2·3) holding, with $L(\alpha, S)$, and therefore $C(\alpha, S)$, informative about the usefulness of fitting a straight line to a model that is really quadratic. Our example is just a special case of the use of an approximating, but not exact, linear model.

Condition (2·10) cannot be said to possess more than the minimal unrealism of any asymptotic condition. Keep taking observations for any imperfect linear model and one will surely have (2·10). Such imperfection is a characteristic of linear models that do not merely play safe with data but try to combine them in potentially useful ways. The condition of mean square consistency (2·8) can be regarded as a condition of asymptotic conventional good behaviour of $\hat{y}(x; \alpha, S)$. Let us, for brevity, describe the latter as 'good' if it is mean square consistent and 'bad' if it is not. Asymptotic investigations in statistics are usually concerned only with 'good' procedures; the statistical viewpoint that supports such a restriction is one that would describe mean square consistency as a weak, easily obtainable property and would require 'asymptotics' to be directed at the search for 'optimal' predictors within some class of mean square consistent predictors. However, the underlying theme of the present study is the investigation of the asymptotic behaviour of a particular form of cross-validatory assessment and its possible usefulness in the comparison of different predictors for finite values of n. A sequence of n values tending to infinity is a theoretical artefact and we are not going to be able in practice to decide which are 'good' and which are 'bad' predictors. Thus the asymptotic behaviour of 'bad' predictors is arguably of as great interest as that of 'good' predictors.

3. Comparative assessment

Consider now situations where mean square consistency must be regarded as realistic, that is problems involving one or more 'good' predictors. For these, we will now show how $C(\alpha, S)$

may be used to provide an informative comparative, rather than absolute, assessment of performance. For two predictors, indexed by α_1 and α_2, define

$$\Delta C(S) = C(\alpha_1, S) - C(\alpha_2, S), \quad \Delta L(S) = L(\alpha_1, S) - L(\alpha_2, S), \tag{3·1}$$

$$\hat{y}_{-i1} = \hat{y}(x_i;\, \alpha_1, S_{-i}), \quad \hat{y}_{-i2} = \hat{y}(x_i;\, \alpha_2, S_{-i}).$$

With model assumption II and quadratic $L(y, \hat{y})$, we have

$$\Delta L(S) = \frac{1}{n}\sum_{i=1}^{n} (\hat{y}_{-i1} - \eta_i)^2 - \frac{1}{n}\sum_{i=1}^{n} (\hat{y}_{-i2} - \eta_i)^2, \tag{3·2}$$

$$\Delta C(S) = \Delta L(S) + \frac{2}{n}\sum_{i=1}^{n} e_i(\hat{y}_{-i2} - \hat{y}_{-i1}). \tag{3·3}$$

The mixed situation involving the comparison of a 'good' and a 'bad' predictor is exemplified by the k group problem (Stone, 1974*a*) in which the 'good' predictor corresponds to the use of the appropriate group average while the 'bad' predictor corresponds to the use of the overall average only The special case of equally replicated groups can be conveniently treated as a special case of the following example.

Example 3·1: *nested symmetric linear models and the magic number* 2. Suppose that, for two nested linear models, the conditions of symmetry, given by $A_{11}(1) = \ldots = A_{nn}(1)$ and $A_{11}(2) = \ldots = A_{nn}(2)$, are satisfied for the two least squares predictors α_1 and α_2 with projection matrices $A(1)$ and $A(2)$ respectively. Then

$$\Delta C(S) = \frac{\text{RSS}_1}{n(1-p/n)^2} - \frac{\text{RSS}_2}{n\{1-(p+q)/n\}^2}, \tag{3·4}$$

where $\operatorname{tr} A(1) = p$, $\operatorname{tr} A(2) = p+q$ and RSS_1 and RSS_2 are the respective residual sums of squares (Stone, 1974*a*, equation (3·22)). More revealingly, (3·4) can be reexpressed as

$$\Delta C(S) = \frac{nq}{(n-p)^2}\,\text{MS}\left(F - 2 - \frac{q}{n-p-q}\right), \tag{3·5}$$

where $\text{MS} = \text{RSS}_2/(n-p-q)$ and $F = (\text{RSS}_1 - \text{RSS}_2)/(q\text{MS})$ is the customary F statistic for nested linear models.

If condition (2·10) obtains for predictor α_1 but not for α_2, F will usually be $O_p(n)$ and $\Delta C(S)$ will be positive and stochastically bounded away from zero, whereas if, for predictor α_1, $\delta_1 = \delta_2 = \ldots = 0$, $\Delta C(S)$ will usually be asymptotically zero. The cross-validatory choice of α_1 or α_2 is seen to be determined by the sign of $F - 2 - q/(n-p-q)$; if $F > 2 + q/(n-p-q)$ we use α_2, but otherwise α_1. As $n \to \infty$ for fixed p and q, the magic number 2 emerges as the critical value of F.

Leonard & Ord (1976) have also found the asymptotic critical value 2 for the k group problem by two quite different approaches, one classical, one Bayesian; they curiously describe the latter approach, based on an improper prior, as 'more sophisticated'.

It is straightforward to verify that, for Example 3·1 with model assumption II and homoscedasticity, if (2·10) obtains for α_1 but not α_2 then $\Delta C(S)$ and $AL(S)$ are each asymptotically $\Sigma\delta_i^2/n$, where $\delta = \eta - A(1)\,\eta$.

In the case of comparative assessment of two 'good' predictors, we find that we cannot, for very good reasons, have consistency. Rather than attempting a general treatment, it is probably as enlightening and certainly more interesting to work with a notorious example.

Example 3·2: *mean versus median*. We consider model assumption II with x absent and

$\eta_1 = \ldots = \eta_n$, that is, essentially the problem of univariate estimation. With α_1 being the sample mean and α_2 being the sample median, we find after some algebra with an even $n = 2m$

$$n\Delta L(S) = \frac{\Sigma(e_i - \bar{e})^2}{(n-1)^2} - m(e_{(m)}^2 - \bar{e}^2) - m(e_{(m+1)}^2 - \bar{e}^2), \tag{3·6}$$

$$n\Delta C(S) = n\Delta L(S) + 2\left\{\frac{\Sigma(e_i - \bar{e})^2}{n-1} + \sum_{i=1}^{m} e_{(i)}(e_{(m+1)} - \bar{e}) + \sum_{i=m+1}^{n} e_{(i)}(e_{(m)} - \bar{e})\right\}, \tag{3·7}$$

where $e_{(1)} \leqslant \ldots \leqslant e_{(m)} \leqslant e_{(m+1)} \leqslant \ldots \leqslant e_{(n)}$ is the ordering of $e_1, \ldots, e_n$. Suppose that $e_1, \ldots, e_n$ have a common distribution F with variance σ^2. Under weak assumptions on F, $n\Delta L(S)$ will be asymptotically distributed as $n(\bar{e}^2 - e_{(m)}^2)$, while, after some algebra, we find that $n\Delta C(S)$ will be asymptotically distributed as

$$2\sigma^2 - n(e_{(m)} - \bar{e})^2 - E(|e|)\, n(e_{(m+1)} - e_{(m)}). \tag{3·8}$$

From the expression $n(\bar{e}^2 - e_{(m)}^2)$, we see why $n\Delta C(S)$ cannot even be expected to estimate consistently $n\Delta L(S)$, that is, why we cannot expect $\text{plim}\{n\Delta C(S) - n\Delta L(S)\} = 0$. For, if it did, we would be able to ascertain asymptotically the sign of $\bar{e}^2 - e_{(m)}^2$ and hence that of $\bar{e}^2 - (\frac{1}{2}e_{(m)} + \frac{1}{2}e_{(m+1)})^2$. That this must be impossible is clear for the case when F is normal when knowledge of the latter sign could be used to construct a better unbiased estimator of the mean of y than $\bar{y}$, that is, $\bar{y}$ if $\Delta C(S) < 0$ and $\frac{1}{2}(y_{(m)} + y_{(m+1)})$ otherwise.

In the case of normal F, without loss of generality $\sigma^2 = 1$, and (3·8) can be used to find the asymptotic probability that $\Delta C(S) < 0$, that is, that cross-validatory choice will select α_1, the sample mean. For then $\bar{e}$ and $e_{(m)} - \bar{e}$ are independent, whence, as $n \to \infty$,

$$E\{\bar{e}(e_{(m)} - \bar{e})\} = 0, \quad E\{n(e_{(m)} - \bar{e})^2\} = E(ne_{(m)}^2) - E(n\bar{e}^2) \sim \tfrac{1}{2}\pi - 1, \tag{3·9}$$

so that, asymptotically, $n(e_{(m)} - \bar{e})^2$ is $(\frac{1}{2}\pi - 1)\chi_1^2$. Furthermore, it is intuitively ascertainable that $n^{\frac{1}{2}}(e_{(m)} - \bar{e})$ and $n(e_{(m+1)} - e_{(m)})$ are asymptotically independently distributed; it is relatively easy to establish that they are asymptotically uncorrelated. As $n \to \infty$, we find that $n(e_{(m+1)} - e_{(m)})$ is asymptotically distributed as $(2\pi)^{\frac{1}{2}} w$, where w has a standard exponential distribution, while $E(|e|) = (2/\pi)^{\frac{1}{2}}$. Then, for large n, assuming normality,

$$\text{pr}\{\Delta C(S) < 0\} \sim \text{pr}\{(\tfrac{1}{2}\pi - 1)\chi_1^2 + 2w > 2\} = 0·4992. \tag{3·10}$$

This probability is, surprisingly, less than $\frac{1}{2}$; cross-validation chooses the median in more than 50 % of large normal samples. Note that the distribution of $\Delta C(S)$ must be skew because at the same time we find $E\{n\Delta C(S)\} \sim -0·57$. However the asymptotic efficiency of the cross-validatory estimator is

$$\lim_{n\to\infty}\left\{n\int_{\Delta C(S)<0} \bar{e}^2 f(e_1, \ldots, e_n)\, de_1 \ldots de_n + n\int_{\Delta C(S)>0} e_{(m)}^2 f(e_1, \ldots, e_n)\, de_1 \ldots de_n\right\}. \tag{3·11}$$

Appendix 2 shows that this efficiency is 0·870, which compares with that of the median at $2/\pi = 0·637$, and that of a completely random procedure that selects the mean with probability 0·4992, and otherwise the median, with efficiency $0·4992 + (1 - 0·4992)2/\pi = 0·818$. Sceptics of these asymptotic calculations may be reassured by the Monte Carlo results: for $n = 300$, the mean was selected on 1020, or 51 %, out of 2000 trials while (averaged squared error)$^{-1}$ was 0·871.

Consider the effect of a change to $L(y, \hat{y}) = |y - \hat{y}|$. As a caution against drawing any general conclusions from these findings, it is shown in Appendix 3 that, if we merely change from quadratic to modulus loss in Example 3·2, we obtain under normality $\text{pr}\{\Delta C(S) < 0\} \sim 0·4327$ and the asymptotic efficiency of the cross-validatory estimator is 0·711. In this case, a completely random procedure that selects the mean with probability 0·4327, and otherwise the median, has efficiency 0·794, superior to cross-validatory choice.

Another asymptotic calculation that is even more adverse to cross-validatory choice concerns a case where the sample mean has zero asymptotic efficiency.

Example 3·3: *exponential error distribution.* As for Example 3·2, consider model assumption II with x absent and $\eta_1 = \ldots = \eta_n$. However, instead of F normal, suppose that (a) the error density is $f(e) = \exp\{-(e+1)\}$ for $e > -1$ and $f(e) = 0$, otherwise, (b) the choice lies between $\alpha_1 \sim$ sample mean and $\alpha_2 \sim m + 1 - 1/n$, where $m = \min(y_1, \ldots, y_n)$. Note that the estimator α_2 is optimal for quadratic L in the class $m + c_n$, where c_n is a function of n only. Appendix 4 shows that, in this case, the mean is chosen with asymptotic probability 0·157 and the resulting 'choice' estimator has asymptotic mean squared error $0{\cdot}574/n$. The perversity of the 'choice' estimator is asymptotically that of an estimator that, even with the knowledge of $|\bar{e}|$, insisted on choosing α_1, the sample mean, when $\bar{e}^2 > 2$ and α_2 otherwise.

Finally, note that Lunts & Brailovskiy (1967) consider the asymptotic behaviour of a measure very similar to $C(\alpha, S) - L(\alpha, S)$, developed for the pattern recognition context.

APPENDIX 1

Proof of (2·9)

From Stone (1974*a*, p. 126), $\hat{y}_{-i} = \{\hat{y}(x_i; \alpha, S) - A_{ii} y_i\}/(1 - A_{ii})$, whence the left-hand side of (2·7) is

$$\operatorname*{plim}_{n\to\infty} \left(\frac{1}{n} \sum_{i<j}\sum q_{ij} e_i e_j - \frac{1}{n} \sum e_i \delta_i/(1 - A_{ii})\right), \tag{A 1}$$

where $q_{ij} = A_{ij}\{(1 - A_{ii})^{-1} + (1 - A_{jj})^{-1}\}$. Both expressions in (A 1) have zero expectation while (2·9) states their respective variances, whence the result.

APPENDIX 2

Mean versus median: Squared loss

Writing $dF_n = f(e_1, \ldots, e_n)\, de_1 \ldots de_n$, the operand in (3·11) equals

$$\int n e_{(m)}^2 \, dF_n - \int_{\Delta C(S) < 0} n(e_{(m)}^2 - \bar{e}^2)\, dF_n,$$

whose first integral is asymptotically $\frac{1}{2}\pi$. Since normality implies the independence of $\bar{e}$ and $\{e_i - \bar{e}\ (i = 1, \ldots, n)\}$ of which $e_{(m)} - \bar{e}$ and $\Delta C(S)$ are functions, the second integral equals

$$\int_{\Delta C(S) < 0} n(e_{(m)} - \bar{e})^2 \, dF_n,$$

which is asymptotically

$$(\tfrac{1}{2}\pi - 1) \int_{(\frac{1}{2}\pi - 1)\chi_1^2 + 2w > 2} \chi_1^2 f(\chi_1^2)\, e^{-w}\, d\chi_1^2\, dw = 0{\cdot}4215.$$

The asymptotic efficiency is then $(\frac{1}{2}\pi - 0{\cdot}4215)^{-1} = 0{\cdot}870$.

APPENDIX 3

Mean versus median: Modulus loss

With modulus loss

$$n\Delta C(S) = \sum_{i=1}^{n} \left| e_i - \frac{n\bar{e} - e_i}{n-1} \right| - \sum_{i=1}^{n} |e_{(i)} - e_{(m+1)}| - \sum_{i=m+1}^{n} |e_{(i)} - e_{(m)}|,$$

which, after some analysis, is found to have the same asymptotic distribution as

$$u_n v_n - (\tfrac{1}{2}\pi)^{\frac{1}{2}} w_n + E(|e|),$$

where $u_n = \Sigma \operatorname{sgn} e_i/\sqrt{n}$, $v_n = \sqrt{n}(e_{(m)} - \bar{e})$ and $w_n = n(e_{(m+1)} - e_{(m)})/(2\pi)^{\frac{1}{2}}$. Under normality we find that w_n is asymptotically standard exponential, independent of (u_n, v_n), which are asymptotically bivariate normal with means $(0, 0)$, variances $(1, \frac{1}{2}\pi - 1)$ and covariance $(\frac{1}{2}\pi)^{\frac{1}{2}} - (2/\pi)^{\frac{1}{2}}$. Straightforward calculation then gives $\operatorname{pr}\{\Delta C(S) < 0\} = 0{\cdot}4327$, while $n \times$ (mean squared error of the 'choice' estimator) is asymptotically

$$\begin{aligned} \int_{\Delta C(S)<0} n\bar{e}^2 \, dF_n + \int_{\Delta C(S)<0} n e_{(m)}^2 \, dF_n &\sim \tfrac{1}{2}\pi - \int_{\Delta C(S)<0} n(e_{(m)} - \bar{e})^2 \, dF_n \\ &= \tfrac{1}{2}\pi - \int_{u_n v_n - (\frac{1}{2}\pi)^{\frac{1}{2}} w_n + (2/\pi)^{\frac{1}{2}} + (2/\pi)^{\frac{1}{2}} < 0} v_n^2 f(u_n, v_n, w_n) \, du_n \, dv_n \, dw_n \\ &\sim 1{\cdot}406 \end{aligned}$$

on asymptotic evaluation.

APPENDIX 4

Exponential loss

We find

$$n\Delta C(S) = \frac{2n-1}{(n-1)^2} \sum_{i=1}^{n} (y_i - \bar{y})^2 - (m - y_{(2)})^2 - \frac{2(n-2)}{(n-1)} (y_{(2)} - m) - n\left(\bar{y} - m - \frac{n-2}{n-1}\right)^2,$$

where $m = y_{(1)} \leqslant \ldots \leqslant y_{(n)}$ is the ordering of $y_1, \ldots, y_n$, from which it follows that $n\Delta C(S)$ is asymptotically $2 - n\bar{e}^2$ or $2 - z^2$, where z is $N(0, 1)$. Then $\operatorname{pr}\{\Delta C(S) < 0\} \sim \operatorname{pr}(z^2 > 2) = 0{\cdot}157$, while $n \times$ (mean squared error of the 'choice' estimator) is

$$\int_{\Delta C(S)<0} n\bar{e}^2 \, dF_n + \int_{\Delta C(S)>0} n(e_{(1)} + 1 - n^{-1})^2 \, dF_n \sim \int z^2 \phi(z) \, dz = 0{\cdot}574.$$

REFERENCES

LEONARD, T. & ORD, K. (1976). An investigation of the *F*-test procedure as an estimation short-cut. *J. R. Statist. Soc.* B **38**, 95–8.

LUNTS, A. L. & BRAILOVSKIY, V. L. (1967). Evaluation of attributes obtained in statistical decision rules. *Eng. Cybern.* no. 3, 98–109.

STONE, M. (1974*a*). Cross-validatory choice and assessment of statistical predictions (with discussion). *J. R. Statist. Soc.* B **36**, 111–47,

STONE, M. (1974*b*). Cross-validation and multinomial prediction. *Biometrika* **61**, 509–15.

[*Received May* 1976. *Revised July* 1976]