EFFICIENCY ROBUSTNESS OF ESTIMATORS

PAUL SWITZER
STANFORD UNIVERSITY

1. Introduction

This introduction sets out some general available procedures to get large sample efficient estimates of a location parameter when the governing distribution f is well specified. The next section reviews some attempts to relax the f specification. The third section discusses how one chooses from a given repertoire of competing estimators. It is there advocated that their respective estimated standard errors be used to govern the choice and two methods are presented for estimating standard errors nonparametrically for this purpose. Some Monte Carlo comparisons are presented in Section 4 using sample sizes of 30, 60, and 120 together with a short tail and long tail f. The possibility of using sample determined weightings of selected estimators is also briefly explored.

Our discussion of efficiency robustness is set in the context of estimating a location parameter θ, say the center of a symmetric distribution on the real line. We use $\hat{\theta}$ to generically denote a translation invariant estimator of θ. If f is the density function of the distribution and is sufficiently well specified, then there are various general methods available to obtain large sample efficient estimators. Some of these are:

(A) the maximum likelihood estimator, that is, the value of θ which maximizes $L(\theta) = \Pi_1^N f(X_i - \theta)$ where $X_1, X_2, \cdots, X_N$ are a sample of size N from f. For example, if f is Laplace then $\hat{\theta}$ is the sample median; if f is normal then $\hat{\theta}$ is the sample mean; if f is logistic then the MLE is not easily obtained.

(B) the Pitman estimator, namely, $\hat{\theta} = \int \theta L(\theta) d\theta / \int L(\theta) d\theta$, where $L(\theta)$ is the sample likelihood as above. For example, if f is normal then $\hat{\theta}$ is the sample mean; if f is uniform on an interval of fixed length then $\hat{\theta}$ is the midrange.

(C) the midpoint of a symmetric confidence interval for θ based on the locally most powerful rank test for the specified translation family f. The confidence probability is a fixed α, and the LMPRT depends on f through $J(u) = f'(F^{-1}(u))/f(F^{-1}(u))$ for $0 < u < 1$. For example, if f is Laplace then $\hat{\theta}$ is the average of a symmetric pair of the ordered X values; if f is logistic then $\hat{\theta}$ is the average of a symmetric pair of the ordered Walsh averages. Walsh averages are averages of pairs of X values. With $\alpha \to 0$ this is the method of Hodges and Lehmann [5].

(D) a specified weighted average of the ordered X values, namely, $\hat{\theta} =$

$\Sigma_1^N w(i/N + 1)X_{(i)}$ where $w(u)$ is proportional to $J'(u)\int_0^u J(v)\,dv$ and $J(u)$ is as given in (C) above (see [3]). For example, if f is Laplace then $w(u)$ is zero except at $u = \frac{1}{2}$ and $\bar{\theta}$ is therefore the sample median; if f is logistic then $w(u) = 6u(1 - u)$.

Generally speaking, these four estimation methods require different regularity conditions on f to safeguard their large sample efficiency. However, the conditions overlap considerably and will not be a concern here. In small samples, they can be regarded as competitors when they do not coincide.

Now consider this embarrassing situation. I have advised a client to use a particular $\hat{\theta}$ based on his specification of the shape of the density function f from which his sample was to be drawn. This $\hat{\theta}$ was justified by showing it was derived by one of the four methods above. He draws his sample and dutifully computes $\hat{\theta}$ and an estimate of its standard error $s(\hat{\theta})$. Not completely trusting my advice, he also computes his pet estimate $\hat{\theta}'$ and notes that its estimated standard error $s(\hat{\theta}')$ is much smaller than $s(\hat{\theta})$. He then sues me for malpractice.

The apparent poor showing of my recommended asymptotically efficient $\hat{\theta}$ could be ascribed to several sources of which three are: (i) the sampling variability in $s(\hat{\theta})$ and $s(\hat{\theta})$ could make the event $s(\hat{\theta}') \gg s(\hat{\theta})$ not particularly surprising even when the actual standard error of $\hat{\theta}$ is the smaller; (ii) for the sample size at hand, say $N = 100$, the asymptotically efficient $\hat{\theta}$ is indeed inferior to $\hat{\theta}'$ for the given f family; (iii) the f family was inappropriately specified and $\hat{\theta}$ is not even an asymptotically efficient estimator. It is to the third of these that our attention will now be turned.

2. Relaxed model specifications as an approach to robustness

For some time now, and especially in recent years, there has been a search for estimating procedures which retain high efficiency simultaneously for widely differing specifications f. Commonly used estimators like the mean or the median certainly would not qualify, the former having low efficiency for long tail f and the latter for short tail f, generally speaking. Instead, a variety of uncommon procedures have been advocated, some of which are reviewed below, not necessarily in historical order.

The approaches taken by Gastwirth [3] and Hogg [6] can be illustrated with the following example. Suppose f is allowed to be either a normal or a Laplace density. Using approach (D), say, we can find the functions J_G and J_L, respectively, which yield efficient estimators for each of the two families. Gastwirth's idea is to look for something intermediate between J_G and J_L which will have high relative efficiency regardless of whether f is normal or Laplace (or some convex combination of the two). This idea can be extended to three or more f families, to include Cauchy and logistic f, say. If this is done, then Gastwirth suggests that $\hat{\theta}$ be a weighted average of the $33\frac{1}{3}$rd, 50th and $66\frac{2}{3}$rd percentiles of the sample, using weights 0.3, 0.4, 0.3, respectively. The relative efficiency of this $\hat{\theta}$ never falls below 80 per cent for any f belonging to one of the four mentioned families.

Hogg, on the other hand, feels that with samples of moderate size it should be possible to distinguish whether f is normal or Laplace from the sample itself. He suggests using the fourth sample moment to discriminate between the two families. Then, use either the J_G or J_L based estimator, whichever is indicated by the fourth moment. This approach, which can also be extended to three or more f families, yields estimators which have full asymptotic efficiency for any f belonging to one of the admitted families.

Full efficiency within the context of a few well-specified f families can also be achieved, in a straightforward manner, by a somewhat extended use of the maximum likelihood method. Once again if f is either normal or Laplace, then the estimator

$$\hat{\theta} = \begin{cases} \text{mean}, & \text{if } (\sum|X_i - \text{median}|)^2 > \dfrac{n\pi}{2e}(\sum|X_i - \text{mean}|^2), \\ \text{median}, & \text{otherwise.} \end{cases} \tag{1}$$

is the MLE for θ when both families are entertained together and the scale parameter is unknown. But note, should the real f fall outside this pair of families, neither formula (1) nor Hogg's method will necessarily indicate whether the mean or median has higher relative efficiency.

A quite different approach is taken by Huber [7]. There it is assumed that a tightly specified f governs the bulk of the sample data, while a loosly specified f' governs the remainder. This is Tukey's contamination model, in particular when f is taken to be normal and f' has longer tails than f. The robustness of any estimator $\hat{\theta}$ was gauged by Huber to be its maximum asymptotic variance as the contaminating f' ranged over the class $\mathscr{F}$ of its admitted possibilities.

For f normal (scale unknown) and f' any symmetric density centered at θ, the most robust $\hat{\theta}$ is found by minimizing $\sum_1^N \rho(X_i - \theta)$ where $\rho(t) = \max(\frac{1}{2}t^2, \lambda t - \frac{1}{2}t^2)$ and λ is related to the contamination proportion ε. Huber has shown that taking $\lambda = 2.0$ will do quite well for any ε less than 20 per cent. Tukey, some time ago [11], suggested using either trimmed or Winsorized means, both of which are close to Huber's estimator. The trimmed mean ignores a specified number of extreme observations in computing a sample average, whereas the Winsorized mean pulls in the extreme observations in a specified way.

In 1955, Stein [10] suggested that large sample uniformly efficient estimating procedures could be concocted without specifying very much at all about f. Hájek [4] and van Eeden [12] carried through on this suggestion. Specifically, one can estimate the whole J function of (C) so that $\int_0^1 [\hat{J}(u) - J(u)]^2 du \to 0$ in probability, with only mild restrictions on f. This enables one to act as though J were known and thereby, using (C) or (D), say, to construct a uniformly fully efficient estimator $\hat{\theta}$. It is not hard to believe that the J function will not be well estimated generally unless the sample size is huge. In the meantime, such estimators remain to be tried on samples of moderate size, and the more modest objectives of the preceding paragraphs still command attention.

3. Comparison of procedures as an approach to robustness

It is possible to make a good case *against* the use of a particular procedure $\hat{\theta}$. Specifically, if a competing procedure $\hat{\theta}'$ is much better for some f and never much worse for any f, then there is little point in using $\hat{\theta}$. For example, Hodges and Lehmann [5] showed that the median of the Walsh averages (H-L estimator) can be infinitely more efficient than the sample mean, if f has infinite variance. On the other hand, the sample mean is never more than 125/108 times as efficient as the H-L estimator for any continuous symmetric f. This pretty much rules out the sample mean. Indeed, there exist estimators which are never less efficient than $\bar{X}$ in large samples, for example, use procedure (C) with the normal scores test.

In a similar vein, Bickel [1] compared the Hodges-Lehmann estimator with both trimmed means and Winsorized means. He found that the H-L estimator retains its *relative* robustness against these two challengers as well.

It would seem from the preceding considerations that one might confidently use the H-L estimator in preference to the competitors entertained in Bickel's paper. Two concerns remain. First, we may regret somewhat the efficiency lost by not using the simple mean, say, in those few situations when it is actually somewhat better than H-L. Second, the H-L estimator could itself have near zero efficiency relative to a third estimator, even when it is distinctly better than the mean.

The proposal of this paper is that the sample itself should be used to distinguish which one of several competing estimators is most efficient, for the unknown f from which the sample was drawn. To be able to use the sample in this way requires that the competing θ estimators be such that their standard errors can also be estimated without making use of the unknown shape f. In spirit, the proposal is unlike those of the previous section where the object was to pin down f. Rather, it addresses itself to the malpractice suit of Section 1 by standing opposed to arguments which fix upon a procedure based exclusively on *a priori* considerations. However, such considerations could be profitably used to set up the collection of competing estimators; for example, the collection should contain only estimators whose efficiency relative to one another ranges widely from very small to very large numbers as f ranges over a set of reasonable possibilities.

Specifically, let $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$ be three sequences of competing estimators defined for every sample size N. The dependence on N is suppressed in the notation. For example, these might be the H-L estimator, the median, and the midrange, respectively, or they might be three trimmed means with different trimming proportions. Let S_1, S_2, S_3 be sequences of nonparametric estimators of the standard errors of $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$ (more will be said about the S in a moment). Then the recommended estimator of the location parameter θ is

$$\hat{\theta} = \sum_{1}^{3} I_i \hat{\theta}_i \tag{2}$$

where

$$I_i = \begin{cases} 1 & \text{if } \min(S_1, S_2, S_3) = S_i, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

That is, $\hat{\theta}$ is equivalent to one of $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$ depending on which of the three has the smallest estimated standard error.

Now, for convenience, assume that the $\hat{\theta}_i$ were chosen so that $\sqrt{N}(\hat{\theta}_i - \theta)$ has a limiting normal distribution with zero mean, for $i = 1, 2, 3$. If $\sigma_i^2(f)$ denotes the variance of each of the three limiting distributions, then the large sample efficiency of $\hat{\theta}_i$ relative to $\hat{\theta}_j$ is $\sigma_j^2(f)/\sigma_i^2(f) = e_{i,j}(f)$, say. Hence, if $e_{i,j}(f)$ is greater than one for some f, then $\hat{\theta}_i$ is relatively more efficient than $\hat{\theta}_j$ for that f family. It is now further assumed that the standard error estimates were chosen so that NS_i^2 consistently estimates $\sigma_i^2(f)$, for each i and any f belonging to a *large* class $\mathscr{F}$: it is in this sense that we referred to the S as nonparametric. For example, if $\hat{\theta}_1$ is the sample mean, then NS_1^2 is the sample variance and has this property for any f having a finite variance. Some general methods for obtaining nonparametric S will be given shortly.

Let $\hat{\theta}(f)$ denote the most efficient of the three competiors for a given f. That is, if $e_{1,2}(f) > 1$ and $e_{1,3}(f) > 1$, then $\hat{\theta}(f) = \hat{\theta}_1$ for that f. With the assumptions of the preceding paragraph, it follows that $\sqrt{N}(\hat{\theta} - \theta)$ has the same limiting distribution as $\sqrt{N}(\hat{\theta}(f - \theta)$, for every $f \in \mathscr{F}$. This makes the proposed estimator $\hat{\theta}$ of (2) as efficient in large samples as $\hat{\theta}(f)$. So $\hat{\theta}$ is never less efficient than any of the initial competitors $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$, (be they H-L, van Eeden, or whatever), and it is always more efficient than some two of them. Should $e_{1,2}(f) \cong 1$ and $e_{1,3}(f) > 1$, then $\hat{\theta}$ will bounce between $\hat{\theta}_1$ and $\hat{\theta}_2$, but this is of no concern since both are equally good and better than $\hat{\theta}_3$.

We have thus established a desirable large sample property for $\hat{\theta}$, namely, we do as well as if we had known which of $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$ was best to begin with. However, its desirability hinges on how large we can make $\mathscr{F}$ and how well it does in samples of moderate size. For if $\mathscr{F}$ is too restricted then our $\hat{\theta}$ will show little advantage over the estimators of Gastwirth and Hogg. On the other hand, even when $\mathscr{F}$ is quite large, if enormous sample sizes are needed, then we may as well try van Eeden's uniformly efficient estimator. Recall that $\hat{\theta}$ of (2) is *not* uniformly fully efficient in $\mathscr{F}$: it is merely as efficient in $\mathscr{F}$ as the best of a preselected small collection of trial estimators, and in small samples it is necessarily *less* efficient than the best of the trial estimators.

As we remarked earlier, the choice of initial trial competitors $\hat{\theta}_i$ will be necessarily influenced by the ability to get nonparametric estimates of their standard errors. Here are two general procedures:

(E) Suppose a trial estimator $\hat{\theta}_i$ is constructed by the method (C), that is, by taking the midpoint of a level α nonparametric confidence interval for θ. Under quite general conditions on f, $\sqrt{N}(\hat{\theta}_i - \theta)$ will have an asymptotic normal distribution with asymptotic variance consistently estimated by a constant times the

squared length of the confidence interval. This constant depends on the chosen α and N, but not on f (see Sen [9]). For example, when $\hat{\theta}_i$ is based on the Wilcoxon test, then the length of the confidence interval has the required consistency property for any f such that $\int f^2$ is finite.

(F) Assume the sample can be divided into K blocks of equal size $n = N/K$. In each of these blocks compute your favorite $\hat{\theta}_i$ estimates based on samples of size n; denote these by $\hat{\theta}_i^k$, $k = 1, 2, \cdots, K$ and $i = 1, 2, 3$ say. Then take the overall $\hat{\theta}_i$ to be the average of the block estimates, that is, $\hat{\theta}_i = \Sigma_1^k \hat{\theta}_i^k/K$. Its standard error is estimated nonparametrically by taking the sample variance of the $\hat{\theta}_i^k$, that is $S_i^2 = \Sigma_1^k (\hat{\theta}_i^k - \hat{\theta}_i)^2/K(K-1)$. Such S_i are consistent in the required sense provided only that the individual $\hat{\theta}_i^k$ has finite variance and that $K \to \infty$ with N. Estimators based on dividing up the data have been proposed by Box [2].

The second of the above two general procedures should be preferred for its simplicity. There will be some loss in efficiency due to the partitioning of the data, but this is usually slight. If both n and K are allowed to grow with N, and if $\sqrt{N}(\hat{\theta}_i^k - \theta)$ are themselves asymptotically normal, then there is no efficiency loss in large samples by dividing up the data. This second procedure also resembles the jack-knife method of estimation which might well be used in its place because it, too, produces an estimate of the standard error. See Miller [8] and the references therein contained for some caveats on the use of the jack-knife.

4. Some small sample Monte Carlo calculations

Let the total sample size N be divisible by six for purposes of the succeeding illustration. Divide the data into $K = N/6$ equal groups (at random). In each group k we compute the three midranges, namely,

$$\begin{aligned} \hat{\theta}_1^k &= \tfrac{1}{2}[X_{(3)} + X_{(4)}], \\ \hat{\theta}_2^k &= \tfrac{1}{2}[X_{(2)} + X_{(5)}], \\ \hat{\theta}_3^k &= \tfrac{1}{2}[X_{(1)} + X_{(6)}]. \end{aligned} \qquad k = 1, 2, \cdots, K. \tag{4}$$

The three competing trial estimators $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$ are the respective averages of these midranges, as outlined in (F). The nonparametric estimates of their respective standard errors S_i are also as given in (F) and are proportional to $\Sigma_1^k (\hat{\theta}_i^k - \hat{\theta}_i)^2$, $i = 1, 2, 3$.

Samples of size $N = 30, 60$, and 120 were drawn from a short tail distribution (uniform), a normal distribution and a long tail distribution (contaminated normal). The contaminated normal was a 90 per cent to 10 per cent mixture of a standard normal and three times a standard normal, respectively. Each sample was replicated 800 times. Table I shows the resulting Monte Carlo estimates of the variances of $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$ for each of the three sample sizes and each of three parent distribution models. The table entries have been scaled for ease of presentation, so horizontal comparisons should not be made. The standard errors of the table entries themselves run about 5 per cent.

TABLE I

SCALED VARIANCES OF VARIOUS ESTIMATORS
These variances were each computed from 800 independent Monte Carlo replications of each sample.

		Uniform		Normal		Contaminated Normal	
$N = 30$	$\hat{\theta}_1$	1.61	(7%)	1.16	(35%)	1.39	(45%)
	$\hat{\theta}_2$	1.05	(18%)	1.15	(34%)	1.55	(41%)
	$\hat{\theta}_3$	0.54	(75%)	1.46	(31%)	4.48	(14%)
	$\hat{\theta}$	0.74		1.27		1.52	
	$\hat{\theta}^*$	1.47		2.56		2.78	
$N = 60$	$\hat{\theta}_1$	1.62	(1%)	1.11	(34%)	1.36	(48%)
	$\hat{\theta}_2$	1.09	(6%)	1.15	(39%)	1.55	(46%)
	$\hat{\theta}_3$	0.58	(93%)	1.52	(27%)	4.34	(6%)
	$\hat{\theta}$	0.68		1.23		1.45	
	$\hat{\theta}^*$	0.82		1.28		1.69	
$N = 120$	$\hat{\theta}_1$	1.56	(0%)	1.23	(33%)	1.52	(51%)
	$\hat{\theta}_2$	1.07	(1%)	1.14	(44%)	1.55	(48%)
	$\hat{\theta}_3$	0.58	(99%)	1.46	(23%)	4.20	(1%)
	$\hat{\theta}$	0.58		1.22		1.54	
	$\hat{\theta}^*$	0.72		1.11		1.47	

The results are as expected. For example, $\hat{\theta}_3$ is strongly favored by the short tail distribution and strongly disfavored by the long tail distribution. The suggested procedure (2) was also applied to these three competing estimators using the S_i as computed from (F) for each sample. The variance of the resulting estimator $\hat{\theta}$ is shown in the table. It is necessarily larger than the variance of the best of $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$, but for the sample sizes used here it performs nearly as well for each of the three f families. The percentage figures in parentheses indicate the relative frequencies with which each of the three competitors were used when they were thrown into procedure (2).

5. Miscellanea

In using the recommended estimator $\hat{\theta}$ of (2), several reasonable questions arise. What if this estimator, itself, was used as one of its component competitors? To use it in this way one would need a nonparametric estimator of *its* standard error, say S. But whatever value we give to S should not be less than min (S_1, S_2, S_3), because $\hat{\theta}$ is either equal to the best of the three competitors or something worse. It follows that the inclusion of $\hat{\theta}$, itself, does not affect the procedure (2).

The general problem of getting an estimate S of the standard error of $\hat{\theta}$ is not taken up here in detail. In a sample of moderate size, $S = \min (S_1, S_2, S_3)$

will be slightly, but optimistically, biased. Corrections could be obtained by using the approximate joint normality of the competing $\hat{\theta}_i$; this is likely to lead to complicated calculations involving the covariances, which must then also be estimated nonparametrically.

If data groupings as in (F) are used, then we can indeed estimate nonparametrically the covariance between $\hat{\theta}_i$ and $\hat{\theta}_j$ by using the sample covariance,

$$S_{i,j} = \sum_{k=1}^{K} (\hat{\theta}_i^k - \hat{\theta}_i)(\hat{\theta}_j^k - \hat{\theta}_j)/K(K-1). \tag{5}$$

Having the $S_{i,j}$ in hand might also suggest that we can improve on the recommendation of (2) by using an "almost optimally weighted" linear combination of the initial competitors $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$, rather than using weights which are always zero or one as in (2). Specifically, take

$$\hat{\theta}^* = \sum \hat{w}_i \hat{\theta}_i, \qquad \sum w_i = 1, \tag{6}$$

where the weights $\hat{w}_i$, possibly negative, are estimates of the optimal weights w_i we would use if we know the covariances precisely. The w_i satisfy the following linear equations (using T initial competitors)

$$\sum_{j=1}^{T-1} \operatorname{Cov}(\hat{\theta}_T - \hat{\theta}_i, \hat{\theta}_T - \hat{\theta}_j) w_j = \operatorname{Cov}(\hat{\theta}_T - \hat{\theta}_i, \hat{\theta}_T), \tag{7}$$

$$i = 1, 2, \cdots, T-1, \qquad w_T = 1 - \textstyle\sum_{j=1}^{T-1} w_j.$$

These equations are symmetric in $w_1, w_2, \cdots, w_T$. Changing the w to $\hat{w}$, they can be solved by using the covariances estimated from the sample. Provided the $\hat{\theta}_i$ are linearly independent, the resulting $\hat{w}_i$ will consistently estimate the w_i, and therefore $\sqrt{N}(\hat{\theta}^* - \theta)$ will have the same asymptotic normal distribution as if we had known the w_i. In general, the asymptotic variance of $\sqrt{N}(\hat{\theta}^* - \theta)$ will be strictly less than that for any initial competitor regardless of the density f governing the data, unless the optimal weights really are zero-one.

However, one might expect the small sample behavior of $\hat{w}_i$ to be erratic causing $\hat{\theta}^*$ to be actually inferior even to the $\hat{\theta}$ of (2) which necessarily uses zero-one weights. This phenomenon is clearly demonstrated by the Monte Carlo results of Table I, particularly for $N = 30$ where only five data groups were available to estimate the three variances and three covariances. In the case of uniform f, where 0-0-1 weighting is actually optimal, $\hat{\theta}$ continues to best $\hat{\theta}^*$ even in moderately large samples. For the normal and contaminated normal f, $\hat{\theta}^*$ is about as good as $\hat{\theta}$ when $N = 60$ and is possibly somewhat superior when $N = 120$. A few unreported Monte Carlo calculations at $N = 240$ continue to show that $\hat{\theta}^*$ is slightly but noticeably better than $\hat{\theta}$.

Herman Rubin suggested in a private communication that it would be reasonable to let the number of trial competitors used with procedure (2) depend on the sample size N. For purposes of illustration we have worked here with three competitors, but Rubin's suggestion appears quite reasonable. It is in the spirit of van Eeden's proposal [12] and merits further study.

In conclusion it should be recalled that this discussion of efficiency robustness was carried through in the context of a specific one sample problem of estimating the center of a symmetric distribution. Our main interest has been to characterize various approaches to the robustness question rather than to provide specific suggestions to specific problems. The two sample shift problem, for example, could have been treated in an almost identical manner where the parameter θ is the amount of the shift.

Immediately prior to the presentation of this paper in July, 1970, the author became aware of a highly relevant unpublished thesis by L. Jaeckel dated December 1969, Statistics Department, Berkeley. It contains essentially among other things, the proposal made in our formula (2). To this extent, and to the extent of any other overlap with this paper, priority belongs to Dr. Jaeckel.

REFERENCES

[1] P. J. BICKEL, "On some robust estimates of location," *Ann. Math. Statist.*, Vol. 36 (1965), pp. 847–858.

[2] G. E. P. BOX, "Non-normality and tests on variances," *Biometrika*, Vol. 40 (1953), pp. 318–335.

[3] J. L. GASTWIRTH, "On robust procedures," *J. Amer. Statist. Assoc.*, Vol. 61 (1966), pp. 929–948.

[4] J. HÁJEK, "Asymptotically most powerful rank-order tests," *Ann. Math. Statist.*, Vol. 33 (1962), pp. 1124–1147.

[5] J. L. HODGES JR. and E. L. LEHMANN, "Estimates of location based on rank tests," *Ann. Math. Statist.*, Vol. 24 (1963), pp. 598–611.

[6] R. V. HOGG, "Some observations on robust estimation," *J. Amer. Statist. Assoc.*, Vol. 62 (1967), pp. 1179–1186.

[7] P. J. HUBER, "Robust estimation of a location parameter," *Ann. Math. Statist.*, Vol. 35 (1964), pp. 73–102.

[8] R. G. MILLER, JR., "Jackknifing variances," *Ann. Math. Statist.*, Vol. 39 (1968), pp. 567–582.

[9] P. K. SEN, "On a distribution-free method of estimating asymptotically efficiency of a class of non-parametric tests," *Ann. Math. Statist.*, Vol. 37 (1966), pp. 1759–1770.

[10] C. STEIN, "Efficient nonparametric testing and estimation," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1956, Vol. I, pp. 187–196.

[11] J. W. TUKEY, *Reports* 31–34, Statistics Research Group, Princeton University, 1949.

[12] C. VAN EEDEN, "Efficiency-robust estimation of location," *Ann. Math. Statist.*, Vol. 41 (1970), pp. 172–181.