EFFICIENCY IN NORMAL SAMPLES AND TOLERANCE OF EXTREME VALUES FOR SOME ESTIMATES OF LOCATION

J. L. HODGES, JR.
UNIVERSITY OF CALIFORNIA, BERKELEY

1. Summary

This paper presents a number of separate but interrelated results concerned with estimates for the symmetric one-sample location problem. (1) Devices are discussed which, in the normal case, increase the information obtainable by random sampling experiments by a factor of hundreds or thousands. (2) Using these devices, sampling evidence is presented that supports the asymptotic theory for a recently introduced estimate, here called T. (3) A linear estimate, called W, is proposed as a natural analog of T, and is used to check the sampling experiment. (4) The estimate T is recognized as a member of a class of estimates, and the class is explored for other members that are easier to compute. (5) One of the simplest of these, called D, is seen to correspond to the one-sample analog of Galton's test, whose null distribution is given. (6) The same samples used with T are applied to D, with closely similar results. (7) A simple numerical measure of tolerance to extreme values is proposed, and methods of evaluating it are presented in two classes of cases that cover the estimates here discussed. (8) A number of estimates, including $\overline{X}$, T, D, and the trimmed and Winsorized means, are compared with regard to normal efficiency, ease of computation, and extreme value tolerance.

2. Introduction

Consider the problem of estimating the center μ of a symmetric population on the basis of a sample $X_1, \cdots, X_n$. It was pointed out by Hodges and Lehmann [6] that, in a natural way, an estimate for μ could be formed from any of a class of rank tests of the value of μ. Perhaps the most interesting of the estimates there considered is the one which corresponds to the Wilcoxon one-sample test. This estimate, denoted here by T and defined in the next section, was shown to be asymptotically normal as $n \rightarrow \infty$, and to have attractive large-

Prepared with the partial support of the Office of Naval Research under Contract No. Nonr-222(43).

sample properties relative to $\overline{X} = (X_1 + \cdots + X_n)/n$. If we denote the asymptotic efficiency of T relative to $\overline{X}$ by $ae(T)$, then $ae(T)$ is identical with the asymptotic efficiency of the Wilcoxon test relative to the t-test. This means, in particular, that $ae(T)$ will exceed one for populations with tails somewhat heavier than the normal. Further, it is shown in [5] that no matter what the shape of the population, $ae(T) \geq .864$. Finally, in the ideal normal case for which $\overline{X}$ is the optimum estimate, $ae(T) = 3/\pi = .955$.

Since these properties are all asymptotic, it becomes important to know to what extent they hold good in samples of moderate size. For example, one would like to know, for moderate samples drawn from a normal population, how close to .955 is the efficiency of T? How near is its distribution to the limiting normal form? Some results relevant to the first of these questions may be quoted from the literature. In [6] it was noted that when $n = 3$, T becomes a linear function of the three order statistics, and has an efficiency (in the sense of variance ratio) of .979. The possibly related question of the power efficiency of the Wilcoxon test against normal shift was investigated numerically by Klotz [8]. He found, for sample sizes $5 \leq n \leq 10$ and significance levels near .05, values near .955 and even usually somewhat above this limit value.

These results are encouraging, but the case $n = 3$ is rather degenerate, and test efficiency need not reflect estimate efficiency. It was thought worthwhile to carry out a sampling experiment for an intermediate value of n, and the value $n = 18$ was chosen for the reasons given in section 4. In general, the precise study of properties of an estimate by sampling would require enormous numbers of samples. Taking advantage of special features of the normal population, it was possible to obtain with the aid of only 100 samples the close estimate of the efficiency of T reported in section 7, and also to examine the approach of the distribution of T to the normal shape (section 8). (Corresponding results are given for a variant U of the estimate T.) These findings, together with the asymptotic theory and the facts for very small samples stated in section 3, suggest that the asymptotic properties of T may be trusted for normal samples of any size. The adequacy of the sampling is supported by its use on an estimate W of known variance; this estimate has independent interest because it may be viewed as a linear analogue to the nonlinear T (section 5).

The main drawback of the estimate T is that it is troublesome to compute except when n is small. For this reason, it seems desirable to try to simplify the estimate while retaining its good distributional properties. To this end, we give in section 9 a class of estimates of which T is a typical member, and consider two methods of finding in the class estimates that are simpler than T, but which on intuitive grounds should have similar distributional properties.

One of these, called D, is examined in section 10, using the same 100 samples. It appears that, at least for normal samples of 18, the estimate D is about as good as the more laborious T. It is shown that, according to the general principle cited above, D corresponds to a rank test that is analogous to the Galton test, and the null distribution of this test is derived.

Behavior of an estimate such as T or D in normal samples is, of course, only part of the story, and one would like to know, for example, how well it stands up in the presence of extreme values. As a contribution to this difficult problem, a simple measure of tolerance of extreme values is introduced in section 11, and this measure is evaluated for two classes of estimates, including those considered in this paper and also the trimmed and Winsorized means. Finally, in section 12 various estimates are compared with regard both to extreme value tolerance and normal efficiency. The estimate D appears to have attractive properties, and to be worth further study. In particular, it would be very desirable to discover its asymptotic distribution.

3. The estimates T and U

Let us denote the sample $X_1, \cdots, X_n$, when arranged in order of increasing size, by $Y_1 \leq \cdots \leq Y_n$. Denote by $\mathcal{A}$ the set of all pairs (i, j) such that $1 \leq i \leq j \leq n$. For each $(i, j) \in \mathcal{A}$, form the mean $M_{ij} = \frac{1}{2}(Y_i + Y_j)$. The statistic T is defined as the median of these means,

$$T = \text{med}\,\{M_{ij}\colon (i, j) \in \mathcal{A}\}. \tag{3.1}$$

If the number $\#(\mathcal{A}) = \frac{1}{2}n(n + 1)$ of these means is even, we shall, as is customary, define the median T as the value midway between the two central values of the set of means.

It seems natural to consider also the slightly different estimate that results when the identity means $M_{ii} = Y_i$ are excluded. Let $\mathcal{B}$ denote the set of all pairs (i, j) with $1 \leq i < j \leq n$, and define

$$U = \text{med}\,\{M_{ij}\colon (i, j) \in \mathcal{B}\}, \tag{3.2}$$

with the same convention if $\#(\mathcal{B}) = \frac{1}{2}n(n - 1)$ is even.

Both T and U can, of course, be defined directly in terms of the unordered observations, but the definitions as given will unify the treatment with that of section 9. Furthermore, when computing the estimates (section 6) it is more convenient to work with the ordered sample.

For certain very small sample sizes, the estimates T and U degenerate to linear functions of the order statistics. In these cases, and for normal samples, their variances can be computed from table I of Sarhan and Greenberg [10].

TABLE I

LINEAR CASES

n	T	$e(T)$	U	$e(U)$
1	Y_1	1	undefined	—
2	$\frac{1}{2}(Y_1 + Y_2)$	1	$\frac{1}{2}(Y_1 + Y_2)$	1
3	$\frac{1}{4}Y_1 + \frac{1}{2}Y_2 + \frac{1}{4}Y_3$	.979	$\frac{1}{2}(Y_1 + Y_3)$	.920
4	nonlinear	?	$\frac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4)$	1

To anchor the lower end of the range of n, these cases are summarized in table I, where $e(T)$ and $e(U)$ are the efficiencies of T and U relative to $\overline{X}$, as defined by variance ratio.

The fact that $e(U)$ is substantially higher at $n = 4$ than at $n = 3$ may be related to a "parity effect" which was influential in choosing the value of n for the sampling experiment. At $n = 3$, $\#(\mathcal{B}) = 3$ is odd, so U is a "pure" median; at $n = 4$, $\#(\mathcal{B}) = 6$ is even, and U is the average of two means M_{ij}. It seems intuitive that such averaging would improve the estimate when sampling from a normal population.

This phenomenon is easier to display in the case of the sample median $\tilde{X}$, efficiency values of which are given in table II. When n is even, $\tilde{X}$ is the average of two order statistics, and its normal efficiency is substantially higher than indicated by the adjacent odd values, by about $.6/n$ for $10 \leq n \leq 20$. For these reasons, it was thought that T and U would have a somewhat higher normal efficiency when $\#(\mathcal{A})$ and $\#(\mathcal{B})$ are even than when they are odd.

TABLE II

NORMAL EFFICIENCY OF $\tilde{X}$ FOR $n \leq 20$

n	$e(\tilde{X})$	n	$e(\tilde{X})$	n	$e(\tilde{X})$	n	$e(\tilde{X})$
1	1.000 000	6	.776 123	11	.662 784	16	.691 561
2	1.000 000	7	.678 828	12	.709 122	17	.653 257
3	.742 935	8	.743 247	13	.658 594	18	.685 630
4	.838 365	9	.668 936	14	.699 130	19	.651 454
5	.697 268	10	.722 928	15	.655 557	20	.680 855

4. The sampling design

Since precise information about normal efficiency for one value of n seemed more valuable than diffuse information for several, it was decided to concentrate on a single value, taking one large enough to escape the degenerate behavior of very small sample sizes and to reflect the sort of moderate sizes often encountered in practice. The linear check W described in section 5 was possible only for $n \leq 20$ because of the range of the Sarhan-Greenberg table. The value chosen, $n = 18$, is the largest in this range for which both T and U are pure medians; fortunately, for this value the estimate D considered in section 10 is also an unaveraged median. As explained in section 3, this choice should lead to conservative results.

The motivation for the somewhat complex sampling design will be given in terms of T, with similar remarks applying to U, W, and D. Since a linear transformation applied to the sample also affects T and $\overline{X}$, there is no loss of generality in using the standard normal population $\mathfrak{N}(0, 1)$, for which extensive tables of random deviates are available.

To obtain precise estimates of the distribution of T would require an enormous number of samples if we proceeded directly, by simply finding the values of T in each such sample. Fortunately, special features of the normal population permit a drastic economy.

Consider a sample of n from $\mathfrak{N}(0, 1)$ and let $\bar{X}$ and T denote estimates computed from this sample. Write $\Delta = T - \bar{X}$, so that $T = \bar{X} + \Delta$. If each observation X_i is increased by c, so are T and $\bar{X}$, and hence Δ is unchanged. Therefore Δ is a function only of the sample differences $X_i - X_1$, $i = 2, \cdots, n$, which implies that Δ is independent of $\bar{X}$. The distribution of T is the convolution of the distribution of Δ with the known distribution $\mathfrak{N}(0, 1/n)$ of $\bar{X}$. We may therefore proceed indirectly as follows: Draw a number of samples of size n from $\mathfrak{N}(0, 1)$. For each sample compute $\bar{X}$ and T, and thus find the value of Δ. Use these observed Δ-values to estimate the distribution of Δ. Finally, convolute the estimated distribution of Δ with the known distribution of $\bar{X}$ to obtain an estimate of the distribution of T.

The advantage of this indirect method resides in the fact that T is a highly efficient estimate, and therefore highly correlated with $\bar{X}$. This means that Δ has a spread much smaller than that of $\bar{X}$. Consequently, an estimate for the distribution of Δ, based on few samples and crude, relative to the spread of Δ, can give us an estimate for the distribution of T which is precise, relative to its much larger spread.

This indirect approach depends on special features of the normal population, but an additional refinement is of more general applicability. When estimating the distribution of Δ, we may classify the samples into strata, and draw separate samples from each stratum. If k_s samples are drawn from stratum s, resulting in values Δ_{sj} of Δ, $j = 1, 2, \cdots, k_s$, these values may be used to estimate the conditional distribution of Δ in stratum s. If P_s is the probability that a random sample of n comes from stratum s, then the weights P_s may be used to combine the estimates of the conditional distributions, to produce an estimate of the (unconditional) distribution of Δ.

Such stratification is feasible only if two conditions are met: it must be possible to calculate the probabilities P_s, and it must be possible to obtain samples which are randomly drawn conditionally from each stratum. By general principles of stratified sampling, stratification is really effective only if a third condition holds: the conditional distributions of Δ in the different strata must be substantially different.

An outstanding feature of T is its insensitivity to extreme values (section 11). On the other hand, $\bar{X}$ is rather sensitive to them, especially if they occur on only one end of the sample. We may therefore expect $|\Delta|$ to be large when the sample has values far out in one tail but not in the other. In contrast, $\bar{X}$ and T will tend to be close, and $|\Delta|$ to be small, if the sample is nearly symmetric, and especially if it lies in a narrow range.

A method of stratification which meets all three conditions consists in dividing the axis into a finite number of intervals, and classifying the samples into strata according to the numbers H_i of observations from the various intervals. The probabilities P_s of such strata can be calculated from the multinomial distribution. It is easy to see at a glance in which stratum a sample falls. Finally, the values of H_i in the extreme intervals will characterize the range and asymmetry of the sample.

In the experiment as performed, the axis was divided into seven intervals by the points ± 2.5, ± 2.0, ± 1.8. Let $H_1, H_2, \cdots$ denote the numbers of the $n = 18$ observations falling in $(-\infty, -2.5)$, $(-2.5, -2.0)$, $\cdots$. In terms of these numbers, twelve strata were defined by the following conditions:

Stratum 1: $H_3 + H_4 + H_5 = 18$,
Strata 2–7: $H_1 + H_7 = 0$, $H_2 + H_6 > 0$;
Stratum 2: $(H_2 = 1, H_3 > 0, H_6 = 0)$ or $(H_6 = 1, H_5 > 0, H_2 = 0)$,
Stratum 3: $(H_2 = 1, H_3 = 0, H_6 = 0)$ or $(H_6 = 1, H_5 = 0, H_2 = 0)$,
Stratum 4: $H_2 = H_6 = 1$,
Stratum 5: $(H_2 = 2, H_6 = 0)$ or $(H_6 = 2, H_2 = 0)$,
Stratum 6: $(H_2 = 3, H_6 = 0)$ or $(H_6 = 3, H_2 = 0)$,
Strata 8-12: $H_1 + H_7 > 0$;
Stratum 8: $(H_1 = 1, H_2 > 0, H_7 = 0)$ or $(H_7 = 1, H_6 > 0, H_1 = 0)$,
Stratum 9: $(H_1 = 1, H_2 = 0, H_7 = 0)$ or $(H_7 = 1, H_6 = 0, H_1 = 0)$,
Stratum 10: $H_1 = H_7 = 0$,
Stratum 11: $(H_1 = 2, H_7 = 0)$ or $(H_7 = 2, H_1 = 0)$.

The probability of any stratum can easily be expressed in terms of the standard normal cumulative Φ by simple formulas. For example,

$$P_4 = P_5 = 306[\Phi(2.5) - \Phi(2.0)]^2[\Phi(2.0) - \Phi(-2.0)]^{16}. \tag{4.1}$$

The probabilities of the twelve strata are shown in table III.

TABLE III

STRATA PROBABILITIES AND SAMPLE NUMBERS

s	P_s	k_s	s	P_s	k_s	s	P_s	k_s
1	.432 479	15	5	.039 740	8	9	.135 645	15
2	.056 798	10	6	.003 673	2	10	.009 661	2
3	.213 000	25	7	.013 128	2	11	.009 661	4
4	.039 740	4	8	.045 116	10	12	.001 359	3

One hundred samples were drawn, allocated among the strata as shown by the numbers k_s in table III. This allocation was governed by the desire to insure adequate representation of the strata in which it was anticipated that $|\Delta|$ would be large and variable, at the expense of strata where $|\Delta|$ might tend to be small and constant. Thus stratum 1, consisting of samples from the interval $(-2, 2)$, constitutes 43% of all samples. But, since $|\Delta|$ should tend to be small here, only $k_1 = 15$ of the 100 samples were taken from stratum 1. On the other hand, the samples in stratum 11 have two extreme observations, $(|X_i| > 2.5)$ at one end not balanced by any at the other end. Here $|\Delta|$ should be large and variable, so $k_{11} = 4$ samples were drawn; a number that by proportional allocation would correspond to an experiment of $k_{11}/P_{11} = 414$ samples instead of 100.

A random sample from one of the strata can be obtained by either of two methods. One may draw unrestricted random samples of size 18 until a sample is obtained which falls in the stratum. Alternatively, one may independently draw observations from the various intervals and combine them. Both methods were used. The 87 samples from strata 1–5, 8, and 9 were obtained from [12] as follows. Beginning on page 142, the first 18 deviates in each column were regarded as a sample from $\mathfrak{N}(0, 1)$. The first $k_1 = 15$ of these samples which fell in $(-2, 2)$ were used as the samples from stratum 1, and so on. In table V, the page and column numbers of each of these 87 samples are given, permitting the reader to check any value. Because samples of the types of strata 6, 7, and 10–12 are rare, these 13 samples were drawn by the alternative method. For strata 6 and 7, the few extreme values needed were drawn from page 1 of [12] taking the first observations with $2 < |X_i| < 2.5$; these were combined with the required numbers of observations $|X_i| < 2$ from the columns indicated in table V. Similarly the values $|X_i| < 2$ for strata 10 and 11 were taken from pages 3 and 4, and the values for stratum 12 from page 1. For the five samples in the catch-all strata 7 and 12, an independent randomization was used to determine the number of extreme values to be taken.

5. A linear estimate

It is always comforting when using random samples to be able to check the quality of the samples by using them to estimate a known quantity, especially one that is closely related to the quantity under investigation. From the tables of variances and covariances of normal order statistics, one can readily compute the variance of any linear combination $W = \sum w_i Y_i$ of the 18 order statistics. If $\sum w_i = 1$, the variance of W can be estimated by the method discussed in section 4, and then be compared with the correct value.

The check is relevant to the extent that weights w_i can be chosen to make W behave like T (and hence like the closely related U). An heuristic argument, given below, suggests that in the normal case a good choice is $w_i \propto \varphi(x_i)$ where $x_i = E(Y_i)$. Using for simplicity weights that closely approximate these, let us define

$$(5.1) \qquad W = \frac{1}{66}(2M_{1,18} + 4M_{2,17} + 6M_{3,16} + 7M_{4,15} + 8M_{5,14} + 9M_{6,13} + 10M_{7,12} + 10M_{8,11} + 10M_{9,10}).$$

The success of this choice of weights is reported in section 7. We now give its heuristic motivation, dealing with the more general problem of a smooth positive density f symmetric about zero. This argument has independent interest in that it suggests the reason for the good properties of the estimate T.

Let us consider the points (i, j) as arranged in rows, associating with each diagonal point $(i, n + 1 - i)$ the points $(i + a, n + 1 - i + a)$ for $a = 0, \pm 1, \pm 2, \cdots$. For a large, but small compared with n, there is a simple approximate relation between the means at $(i, n + 1 - i)$ and at $(i + a, n + 1 - i + a)$. It can be seen that $Y_{i+a} - Y_i$ and $Y_{n+1-i+a} - Y_{n+1-i}$ are both approximately equal to $a/nf(x_i)$. Hence,

$$M_{i+a,n+1-i+a} \doteq M_{i,n+1-i} + a/nf(x_i). \tag{5.2}$$

Now $M_{i+a,n+1-i+a}$ is an increasing function of a, rising from below T to above T. Let $b(i)$ denote that value of a for which the mean is closest to T, so that $M_{i+b(i),n+1-i+b(i)} \doteq T$. Substitution in (5.2) gives

$$b(i) \doteq nf(x_i)[T - M_{i,n+1-i}]. \tag{5.3}$$

By the definition of T, there are as many points with $M_{ij} > T$ as with $M_{ij} < T$. Imposing this condition on (5.3) gives T as, approximately, a weighted average of the Y_i with weights proportional to $f(x_i)$.

While no attempt has been made to rigorize this intuitive discussion, it does help to explain the success of the estimate T, by suggesting that it will tend to behave like a weighted average of the order statistics with weights that are small in the tails of a population with long tails. In practice, of course, f is unknown; the virtue of T is that it accomplishes a reasonable weighting without requiring knowledge of f.

6. The sampling results

For each of the 100 samples of 18, drawn as described in section 4, the estimates $\overline{X}$, U, T, and W were computed. The values of the differences $U - \overline{X}$, $T - \overline{X}$, and $W - \overline{X}$, multiplied by 1000, are shown in table V, as well as the values of $1000(D - \overline{X})$ for the estimate D to be described in section 10. To simplify the table, the signs of all four differences were changed if $U - \overline{X}$ happened to be negative; this is permissible because of the symmetry of all distributions about zero.

The computations of T and U were carried out by the method of Høyland [7], and may be illustrated on the first sample of stratum 1. The first 18 deviates of ([12], p. 142, column 2) (each multiplied by 1000 to avoid decimals) are arranged in decreasing order as shown in table IV, from $Y_{18} = 1.686$ to $Y_1 = -1.632$. Each number in the body of the table is the sum S_{ij} of Y_i below it and Y_j to its left. (The arrangement of the Y_i's in a "broken line" pattern instead of diagonally saves space and also brings the Y's nearer their sums.) Linear combinations of the nine diagonal sums $S_{9,10} = -.004, \cdots, S_{1,18} = .054$ provide the values $\overline{X} = -.0049$ and $W = .0195$, and hence $W - \overline{X} = .024$.

To find U, one must find the median of the sums S_{ij}, $i < j$, and this is done by trial and error, guided by the fact that $S_{ij} \geq S_{k\ell}$ whenever $i \geq k$ and $j \geq \ell$. The line in table IV, passing through the sum $S_{7,12} = .100$, goes as often above as below the row of diagonal sums, and thus divides the 153 sums into 76.5

TABLE IV

COMPUTATION OF T AND U

1686										366	54
616									−152	−704	
572									−196		
571								189			
571							200	189			
341						147	−30				
198					100	4					
	172			125	74						
		25	−4								
			−29								
				−47	−98	−194	−371	−382	−768	−1320	−1632

that are above the line, all at least equal to .100, and the 76.5 that are below the line, none greater than .100. Therefore $U = \frac{1}{2}(.100) = .0500$ and $U - \overline{X} = .055$.

To find T, we add in the 18 identity sums $S_{ii} = 2Y_i$. The dashed extension of the line passes between $S_{10,10} = 2(.025)$ and $S_{11,11} = 2(.172)$. As this line has 8 of the newly added sums above it and 10 below it, the median of the augmented set will be the sum next smaller than $S_{7,12}$. By inspection this is seen to be $S_{7,11} = .074$, so that $T = \frac{1}{2}(.074) = .0370$, and $U - \overline{X} = .042$.

As $U - \overline{X}$ is positive, no sign change is required. The entries in the first row of table V give the differences, multiplied by 1000 to simplify the setting. These entries form the bases of the computations reported in sections 7, 8, and 10.

7. Efficiencies

As a first application of the sampling results reported in section 6, we give estimates of the efficiencies of T and U for normal samples of size $n = 18$. Again, the theoretical discussion will be given in terms of the estimate T, with analogous remarks holding for U.

We define, as is customary, the efficiency of T relative to $\overline{X}$ by the variance ratio $e(T) = \text{Var}(\overline{X})/\text{Var}(T) = 1/n \text{Var}(T)$. Because of the independence of $\overline{X}$ and $\Delta = T - X$, $\text{Var}(T) = 1/n + \text{Var}(\Delta)$. Suppose we have an estimate $\mathcal{E}$ for $\text{Var}(\Delta)$, to which is attached an (estimated) standard error, say $\mathcal{E} \pm \sigma$. Then $1/n + \mathcal{E} \pm \sigma$ is an estimate for $\text{Var}(T)$. Since σ is small compared to $1/n + \mathcal{E}$, we may expand linearly and use

$$\frac{1}{1 + n\mathcal{E}} \pm \frac{n\sigma}{(1 + n\mathcal{E})^2} \tag{7.1}$$

as an estimate for $e(T)$.

The desired estimates $\mathcal{E}$ and σ are obtained from our stratified sample as follows. Since Δ is symmetrically distributed about zero, $\text{Var}(\Delta) = E(\Delta^2)$. If Δ_s denotes the random variable Δ conditioned to lie in stratum s, then

$$E(\Delta^2) = \sum_s P_s E(\Delta_s^2). \tag{7.2}$$

The values of Δ_s, say $\Delta_{s1}, \cdots, \Delta_{sk_s}$, observed in the sampling experiment, give in

$$\mathcal{E}_s = \sum_j \Delta_{sj}^2/k_s \tag{7.3}$$

an (unbiased) estimate for $E(\Delta_s^2)$, with variance $\text{Var}(\Delta_s^2)/k_s$. Combining these estimates gives

$$\mathcal{E} = \sum P_s \mathcal{E}_s \tag{7.4}$$

as the desired (unbiased) estimate for $\text{Var}(\Delta)$.

Since the estimates $\mathcal{E}_s$ are independent, the variance of $\mathcal{E}$ is

$$\sum P_s^2 \text{Var}(\mathcal{E}_s) = \sum P_s^2 \text{Var}(\Delta_s^2)/k_s. \tag{7.5}$$

Regarding the Δ_{sj}^2 as observed values of Δ_s^2,

$$\sigma_s^2 = [k_s \sum_j \Delta_{sj}^4 - (\sum_j \Delta_{sj}^2)^2]/k_s(k_s - 1) \tag{7.6}$$

is the usual (unbiased) estimate for $\text{Var}(\Delta_s^2)$. Hence we may use

$$\sigma^2 = \sum P_s^2 \sigma_s^2/k_s \tag{7.7}$$

as an (unbiased) estimate for the variance of the estimate $\mathcal{E}$. Using the observed values recorded in table V, these calculations were performed for the estimates T, U, and W, giving the following estimated efficiencies:

$$\begin{aligned} e(T) &= .949 \pm .007, \\ e(U) &= .956 \pm .006, \\ e(W) &= .969 \pm .004. \end{aligned} \tag{7.8}$$

The standard errors are crude, but may serve to indicate the high precision of these estimates.

The check provided by the linear estimate W now comes into play. From

TABLE V

Values of 1000Δ for Four Estimates

p	*c*	*U*	*T*	*D*	*W*
		Stratum 1			
142	2	55	42	55	24
	4	48	48	52	42
	5	13	13	−9	8
	8	3	16	43	−18
	9	20	56	−32	12
143	2	17	18	18	−2
	3	35	35	35	12
	5	25	25	20	26
144	4	13	27	−41	33
	8	32	32	35	8
145	2	5	−4	−4	0
	3	73	73	73	54
	4	21	21	−2	−2
	5	31	32	32	15
	6	6	6	27	22
		Stratum 2			
143	4	73	46	−15	19
144	1	62	62	159	89
147	3	7	7	60	−6
148	3	82	88	123	72
152	6	51	51	88	14
153	8	11	83	−19	51
154	1	18	14	18	4
155	10	112	112	84	74
156	4	4	0	15	−15
157	2	71	71	3	88
		Stratum 3			
143	8	30	30	−4	28
	9	22	31	13	5
145	8	1	29	29	−20
147	5	7	16	7	41
	7	20	43	−43	53
	8	59	59	59	21
148	1	8	12	12	10
149	3	26	26	26	−33
	7	9	9	8	9
	9	16	15	−12	41
	10	95	159	76	98
150	3	23	27	−2	36
	5	3	3	3	10
	9	11	47	6	18
151	6	107	146	163	101
	7	70	91	1	86
	9	54	14	−9	26
152	2	69	71	7	40
153	2	25	13	8	22
	3	38	38	38	40
	5	8	8	−11	30
	6	43	34	4	31
	7	19	4	27	−8
154	7	17	11	11	19
	9	21	23	20	38
		Stratum 4			
142	10	50	50	36	21
146	7	12	12	12	−26
147	6	13	27	13	2
150	7	22	19	19	22
		Stratum 5			
142	6	95	103	−52	45
143	10	20	61	61	76
144	5	24	26	26	40
145	1	18	−2	−2	0
147	1	67	67	114	60
148	6	68	68	64	17
149	1	75	84	50	41
162	4	20	24	16	27
		Stratum 6			
100	1	48	48	−117	66
100	2	5	−4	14	−78
		Stratum 7			
101	1	33	33	33	23
101	2	21	21	−6	−58
		Stratum 8			
143	6	87	104	24	93
144	7	66	66	85	50
	9	133	118	116	96
	10	79	97	3	82
148	2	2	2	−6	50
152	7	61	65	65	76
159	1	155	164	70	79
160	9	73	73	76	52
161	4	36	59	−33	25
162	8	62	66	120	88
		Stratum 9			
142	1	45	45	51	70
	3	35	46	26	56
	7	65	66	111	66
143	1	67	67	67	49
	7	64	64	62	50
144	2	70	63	36	59
	3	42	55	42	43
	6	78	75	78	56
145	10	55	55	−86	8
146	1	54	54	59	48
	3	31	31	18	25
	4	17	0	−14	−8
147	2	83	83	60	75
	9	153	153	142	115
148	5	72	72	72	57
		Stratum 10			
120	1	1	1	9	2
	2	152	152	152	100

TABLE V (Continued)

p	c	U	T	D	W	p	c	U	T	D	W
		Stratum 11						Stratum 12			
120	3	171	178	14	109	130	1	64	64	16	44
	4	59	59	29	27		2	164	183	−80	94
	5	207	226	237	126		3	23	−2	−4	39
	6	92	92	72	69						

table I of [10], it is possible to compute the actual variance of W, and hence to find that $e(W) = .9649$. The efficiency estimated from the samples is within one standard error.

It will be recalled that the estimate W was chosen in an attempt to find a linear estimate that would be highly correlated with T and U. If we write $W = \overline{X} + \Gamma$, then the pair (Δ, Γ) is independent of $\overline{X}$, and $E(\Delta) = E(\Gamma) = 0$, so that

$$\text{Cov}\,(T, W) = \frac{1}{n} + E(\Delta\Gamma) = \frac{1}{n} + \sum P_s E(\Delta_s \Gamma_s). \tag{7.9}$$

This formula permits us to use the sampling results to estimate the correlation between T and W, and similarly for U and W. The computed results are

$$\hat{\rho}(T, W) = .994, \qquad \hat{\rho}(U, W) = .995. \tag{7.10}$$

These high correlations indicate that W does indeed behave very similarly to T and U, which lends relevance to the success of the estimate for $e(W)$. (Since the estimate for $e(W)$ is high by .004, one may wish to lower the estimates for $e(T)$ and $e(U)$ by this amount, which does not change the picture appreciably.)

The sampling results suggest that U is somewhat more efficient than T. (Since table V makes it clear that $T - \overline{X}$ and $U - \overline{X}$ are highly correlated, the estimated value of $e(U) - e(T)$ is more precise than the separate standard errors indicate.) This agrees with intuition. The estimate T differs from U by including, in the set of means whose median is taken, the identity means $M_{ii} = Y_i$. This inclusion may be said to move T from U in the direction of the median $\tilde{X}$, whose efficiency is only .686. This should, however, also mean that T has slightly less sensitivity than U to extreme values, as the investigation of section 11 shows to be the case.

The effectiveness of our indirect method of estimating the efficiencies is made clear by noting how many samples would be required to obtain estimates of similar precision by directly observing k values of the estimates themselves. For illustration, suppose V is normally distributed about zero and has efficiency $e(V)$. If $V_1, \cdots, V_k$ are observations on V, this efficiency may be estimated by $k/n \sum V_i^2$. For this estimate to have standard error τ, we would need $k = 2e^2(V)/\tau^2$. For the values $e(U) = .956$ and $\tau = .006$, corresponding to our estimate for $e(U)$, this gives $k = 50{,}000$, compared with the 100 samples actually used.

8. Normality

Asymptotically, as $n \to \infty$, the estimates T and U are known to become normally distributed. It is the approximate normality of the estimates which, to a large extent, justifies our use of variance ratio as a measure of efficiency. If the estimate T based on n observations has (approximately) the distribution $\mathfrak{N}(0, \text{Var}(T))$, and if the efficiency of T relative to the arithmetic mean is defined to be $e = 1/n \text{Var}(T)$, then T will have (approximately) the same distribution as the mean of ne observations. It is accordingly desirable to know, when using $1/n\hat{\sigma}^2(T)$ as an indication of the efficiency of T, the extent to which the actual distribution of T agrees with the normal approximation $\mathfrak{N}(0, \hat{\sigma}^2(T))$. Questions of this kind are usually difficult to answer with precision, but the special features of the normal population again make it possible to use our samples to throw a good deal of light on it for normal samples of size 18.

Because T is symmetrically distributed about zero, it suffices to compare $P(|T| > c)$ with the corresponding probability for the fitted normal $\mathfrak{N}(0, \hat{\sigma}^2(T))$. Let

$$\rho(\delta, c) = P(|\overline{X} + \delta| > c) = \Phi(-\sqrt{n}(c + \delta)) + \Phi(-\sqrt{n}(c - \delta)). \tag{8.1}$$

Then, because of the independence of $\overline{X}$ and $\Delta = T - \overline{X}$, integration with respect to the distribution of Δ gives

$$P(|T| > c) = E\rho(\Delta, c). \tag{8.2}$$

Conditioning on the strata gives

$$E\rho(\Delta, c) = \sum_s P_s E\rho(\Delta_s, c). \tag{8.3}$$

Using the observed values Δ_{sj} of our stratified sample, we therefore have in

$$\sum_s P_s[\sum_j \rho(\Delta_{sj}, c)/k_s] \tag{8.4}$$

an unbiased estimate for $P(|T| > c)$. Clearly,

$$\sum P_s^2\{k_s \sum \rho^2(\Delta_{sj}, c) - [\sum \rho(\Delta_{sj}, c)]^2\}/k_s^2(k_s - 1) \tag{8.5}$$

is an unbiased estimate for the variance of (8.4).

The comparison between the actual distribution of T and the fitted normal is especially worthwhile in the tails of the distribution, since experience indicates that the approach to normality is usually slowest there. The calculations were made for both estimates, for $c = .7$ and $c = .9$, corresponding to about 2.9 and 3.7 standard deviations for T and U. The results are shown below.

TABLE VI

PROBABILITY OUTSIDE $(-c, c)$

c	Estimate	From (8.4)	From Fitted Normal
.7	T	.00386 ± .00013	.00382
	U	.00371 ± .00010	.00368
.9	T	.000205 ± .000012	.000200
	U	.000192 ± .000009	.000188

This excellent agreement, even in the extreme tails, indicates that the asymptotic normality has taken hold at $n = 18$, and also supports the relevance of the estimated efficiencies given in section 7.

The power of the methods we are using is even more striking here than in the estimation of efficiency. To illustrate this point, consider the direct estimation of $P(|U| > .9)$ by observing the frequency with which $|U| > .9$. To produce an estimate as good as that reported above would require something like 2,400,000 such samples, compared with the 100 samples we have used.

9. A family of estimates

While T and U have excellent distributional properties, they are laborious to compute when n is large. This disadvantage may be mitigated by the development of computer programs, but not every potential user will have easy access to a computer. In this section we present a general class of estimates of which T and U are special cases, and indicate other members of the class which are simpler to compute and which also seem to have good distributional properties.

Corresponding to any nonempty subset $\mathcal{S}$ of $\mathcal{A}$, we may define an estimate S by

$$S = \text{med}\, \{M_{ij}\colon (i, j) \in \mathcal{S}\}. \tag{9.1}$$

We shall assume throughout that $\mathcal{S}$ enjoys the symmetry property:

$$(i, j) \in \mathcal{S} \quad \text{implies} \quad (n + 1 - j, n + 1 - i) \in \mathcal{S}; \tag{9.2}$$

in view of the symmetry of the population about its center μ, this assures that S is symmetrically distributed about μ. Note that T and U are special cases of (9.1) corresponding to $\mathcal{S} = \mathcal{A}$ and $\mathcal{S} = \mathcal{B}$.

The labor of computing S depends primarily on the number $\#(\mathcal{S})$ of means whose median is sought, and rises somewhat faster than $\#(\mathcal{S})$. It is the fact that $\#(\mathcal{A})$ and $\#(\mathcal{B})$ both increase at speed n^2 which makes T and U difficult to compute when n is large. Roughly speaking, the work required by these estimates is proportional to n^3. As indicated in section 6, it is not difficult to compute T and U by hand at $n = 18$. The labor would be discouragingly heavy at $n = 50$.

Can we choose $\mathcal{S}$ so that $\#(\mathcal{S})$ is small, permitting S to be computed easily, and still have S share the good distributional properties of T and U? We suggest two approaches to this problem.

(i) *Representative order statistics.* A large sample drawn from a smooth population can be adequately represented by a modest number of its order statistics. The basic reasons for this are as follows. If we are given the values of two order statistics, say $Y_i = u$ and $Y_{i+a+1} = v$, then the intermediate order statistics $Y_{i+1}, \cdots, Y_{i+a}$ are conditionally distributed like an ordered random sample drawn from that portion of the population in the interval (u, v). If the population density is smooth and $v - u$ is small, this conditional distribution is nearly rectangular, so that the intermediate observations contain little information about μ. If a/n is small, $Y_{i+a+1} - Y_i$ is likely to be small, even if a itself is large. If we choose a subset $\mathcal{I}$ of the sequence $1, 2, \cdots, n$, about equally spaced throughout the sequence, we may therefore expect $\{Y_i: i \in \mathcal{J}\}$ to carry nearly all of the information in the sample, even if $\#(\mathcal{J})$ is modest.

Now let $\mathcal{A}(\mathcal{J})$ denote the set of pairs $(i, j) \in \mathcal{A}$ for which $i \in \mathcal{J}$ and $j \in \mathcal{J}$, and let $T(\mathcal{J})$ denote the estimate (9.1) with $\mathcal{S}$ replaced by $\mathcal{A}(\mathcal{J})$. If $\#(\mathcal{J})$ is modest, the hand computation of $T(\mathcal{J})$ will be feasible even if n is quite large. Suppose for example $n = 1000$ and $\#(\mathcal{J}) = 18$. Once the representative order statistics $\{Y_i: i \in \mathcal{J}\}$ have been noted, the work of finding $T(\mathcal{J})$ is the same as that for finding T with $n = 18$, as illustrated in section 6. To the extent to which the 18 selected order statistics represent the sample, we may expect $T(\mathcal{J})$ to have distributional properties similar to those of T computed from the entire set of 1000 observations.

In an entirely analogous way we may define the estimate $U(\mathcal{J})$. For two special cases, this estimate coincides with estimates studied by Mosteller [9]. He proposed as an estimate for μ the arithmetic mean, say $\bar{Y}_k$, of k selected order statistics. From table I we see that when $\#(\mathcal{J}) = 2$, $U(\mathcal{J}) = \bar{Y}_2$, and when $\#(\mathcal{J}) = 4$, $U(\mathcal{J}) = \bar{Y}_4$. Mosteller investigated the asymptotic efficiencies of his estimates $\bar{Y}_k$ for the normal population, for several different methods of spacing the selected order statistics. The asymptotic efficiency of $\bar{Y}_2$ was found to range from .793 to .810, while that of $\bar{Y}_4$ ranged from .896 to .914, depending on the spacing. These figures, and other figures given in his table II, suggest that in normal samples one may suffer an efficiency loss of about one or two percent as a result of replacing a large sample by about 20 equally spaced order statistics.

(ii) *Central means.* We may expect a mean M_{ij} to be near μ if and only if Y_i and Y_j are nearly symmetric in the sample, that is, if $i + j$ is near $n + 1$. This suggests that the estimates T and U would be little affected if we eliminated from $\mathcal{A}$ and $\mathcal{B}$ those pairs (i, j) for which $|i + j - n - 1|$ is large. In general, S may be a reasonable estimate if we use in (9.1) a set of central means M_{ij} where (i, j) is near the diagonal $i + j = n + 1$.

As a preliminary, we note that $\mathcal{A}$ and $\mathcal{B}$ may be reduced substantially without affecting T and U at all. The discussion will be given in terms of $\mathcal{B}$, with analogous remarks holding for $\mathcal{A}$.

It is clear that $M_{ij} = \frac{1}{2}(Y_i + Y_j)$ is an increasing function of both i and j. This means that, if i and j are small enough, we may be sure that M_{ij} falls below the central values in the set $\{M_{ij}: (i, j) \in \mathcal{B}\}$, and hence below U, whatever the values $\{Y_i\}$ may be. The number of means M_{ab} in $\mathcal{B}$ with $a \geq i$, $b \geq j$ is easily seen to be $N_{ij} = (n + 1 - j)[\frac{1}{2}(n + j) - i]$. Let $\mathcal{E}$ denote the set of pairs (i, j) with $N_{ij} \geq \frac{1}{2}[\frac{1}{2}n(n - 1) + 1]$, and let $\mathcal{E}'$ be the set symmetric to $\mathcal{E}$ about the line $i + j = n + 1$. Then it can be seen that any M_{ij} with $(i, j) \in \mathcal{E}$ is below U, while any M_{ij} with $(i, j) \in \mathcal{E}'$ is above U, and hence that

$$U = \text{med}\ \{M_{ij}: (i, j) \in \mathcal{B}^*\} \tag{9.3}$$

where $\mathcal{B}^* = \mathcal{B} - \mathcal{E} - \mathcal{E}'$.

The reduction from $\mathcal{B}$ to $\mathcal{B}^*$ is substantial, and it can be shown that

$$\#(\mathcal{B}^*)/\#(\mathcal{B}) \to \{1 - \sqrt{2} - \log(\sqrt{2} - 1)\} = .4672. \tag{9.4}$$

In a computer program which finds U by comparing each pair of means M_{ij}, the use of $\mathcal{B}^*$ instead of $\mathcal{B}$ would reduce this work by 78%.

Any reduction below $\mathcal{B}^*$ will change the estimate, but in the interest of simplicity this may be worthwhile. The most extreme reduction would consist in using only the diagonal means. The resulting estimate will be considered in the next section.

Each of the estimates (9.1) corresponds to a test according to the method expounded in section 2 of [6]. Without loss of generality we may consider tests for the hypothesis $\mu = 0$. For any $\mathcal{S}$ satisfying (9.2), let Σ denote the number of positive means M_{ij}, where $(i, j) \in \mathcal{S}$. It is obvious that Σ is nondecreasing when all sample values are increased by the same amount, thus meeting condition C of the cited section. By the assumed symmetry of the population, we may under the null hypothesis $\mu = 0$ associate with each sample $X_i, \cdots, X_n$ an equally likely sample $X'_i = -X_i$. In terms of order statistics $Y'_i = -Y_{n+1-i}$ so that $M'_{ij} = -M_{n+1-j,n+1-i}$. In view of (9.2), this means that $\Sigma + \Sigma' = \#(\mathcal{S})$; and since Σ and Σ' are equally likely, this equation implies that Σ is symmetrically distributed about $\frac{1}{2}\#(\mathcal{S})$, as required by condition D. The estimate S corresponds to the test statistic Σ, in the sense of [6].

10. The estimate D

We now consider the estimate (9.1) with $\mathcal{S}$ consisting only of diagonal points. Let

$$D = \text{med}\ \{M_{ij}\colon (i, j) \in \mathcal{D}\} \tag{10.1}$$

where $\mathcal{D}$ consists of the diagonal pairs (i, j) with $1 \leq i < j \leq n$ and $i + j = n + 1$. This estimate, the median of the means of symmetric order statistics, is much easier to compute than T or U.

Computations analogous to those of sections 7 and 8 were carried out for D, using the values of $D - \overline{X}$ recorded in table V. The estimated efficiency of D relative to $\overline{X}$, in normal samples of size $n = 18$, is .954 ± .007. The results comparable to those in table VI are the following.

TABLE VII

PROBABILITY OUTSIDE $(-c, c)$

c	From (7.1)	From Fitted Normal
.7	.00375 ± .00013	.00372
.9	.000196 ± .000012	.000192

It must be emphasized that, unlike T and U, the asymptotic properties of D are not known. In particular, it may be questioned whether D is asymptotically normal. However, the sampling experiment suggests strongly that, at least with samples of moderate size drawn from a normal population, the behavior of D

is very close to that of T and U, both with regard to variance and shape of distribution. Because D is substantially easier to compute than are T and U, these sampling results suggest that the estimate D is worth further study. In particular, it would be of interest to find its asymptotic properties.

The test corresponding to D is of some interest. For the case $S = D$, let us denote the statistic Σ of section 9 by G; that is, G is the number of positive symmetric means. To test the hypothesis $\mu = 0$ against the alternative $\mu > 0$, we should reject when G is large. This test may be viewed as a one-sample analogue of the two-sample Galton test discussed in [4], with the upper and lower halves of the single sample playing roles analogous to those of the two equal samples in that test.

To derive the null distribution of G, consider a sample of n drawn from a continuous population symmetric about 0. If $a_1 > a_2 > \cdots > a_n > 0$ are the absolute values in this sample, we may reconstitute a sample distributionally equivalent to that drawn, by assigning independently and at random to these absolute values the signs $+$ and $-$. The sequence of n signs may now be used to determine an n-step path in the manner discussed in chapter III of Feller [3]. The 2^n possible paths are equally likely, and the null distribution of G may be found by relating G to properties of the path.

Let us establish a correspondence between the order statistics of the reconstituted sample and the steps of the path by letting the j-th step of the path correspond to that order statistic which equals $\pm a_j$, and denote by S_i the step path that thus corresponds to Y_i. Then the steps $S_1, S_2, \cdots$ are the successive down-steps of the path, read from left to right, while $S_n, S_{n-1}, \cdots$ are the successive up-steps.

Let us characterize the condition $M_{i,n+1-i} > 0$, for $1 \leq i \leq \frac{1}{2}n$, in terms of the steps. Clearly, $M_{i,n+1-i} > 0$ means that $|Y_i| < Y_{n+1-i}$, which in turn assures that S_{n+1-i} is an up-step; that S_{n+1-i} comes before S_i; and that S_i is preceded by at most $i - 1$ down-steps. From these facts it is clear that both S_i and S_{n+1-i} are positive steps, in the sense that both lie above the horizontal axis drawn through the start of the path. Similarly, one can argue that $M_{i,n+1-i} < 0$ implies that both S_i and S_{n+1-i} are negative steps.

When n is even, say $n = 2k$, we now see that $2G$ is just the number of positive steps in the path. To make this simple relation true also when $n = 2k + 1$, we must modify the definition of G slightly, by adding to its value as previously defined the number $\frac{1}{2}$ in case $Y_{k+1} > 0$, which occurs if and only if S_{k+1} is the $(k + 1)$-st up-step and hence a positive step.

The distribution of the number of positive steps has been investigated. For the case $n = 2k$, it is given in ([3], p. 77) in a simple closed form, which is shown (p. 80) to have the arc sine limit law. The case $n = 2k + 1$ is discussed in [1], and from the remarks made there it is clear that the same limit law holds (as would indeed be the case had we not modified the definition of G). The nonnormality of the limit law of G suggests that the estimate D may also have a nonnormal limit distribution.

While the test G is interesting because of its relations to D, to the Galton test, and to random walks, it suffers from a practical disadvantage: the sample size must be large before the customary small significance levels become available. This fact reflects the remarkable tendency, emphasized by Feller, for random paths to remain always on one side of the axis.

11. Tolerance of extreme values

A principal motive behind the search in recent years for estimates of location alternative to classical $\overline{X}$ has been the realization that $\overline{X}$ is sensitive to extreme values. This motive is explicit in the proposal of trimmed and Winsorized means, and it was also impelling in the development of estimates based on rank tests, such as T. It is the purpose of this section to introduce a simple numerical measure of the degree to which an estimate of location is able to tolerate extreme values, and to use this measure to compare several estimates. No pretense is made that the proposed measure exhausts the complex extreme-value problem; however, it provides in some cases an easily computed solution to one aspect of that problem.

As an example to motivate the definition, consider the trimmed mean

$$R = (Y_3 + Y_4 + Y_5 + Y_6 + Y_7 + Y_8)/6 \tag{11.1}$$

where $n = 9$. It is obvious that $Y_3 \leq R \leq Y_8$, whatever be the sample values. This implies that Y_1, Y_2, and Y_9 may be as extreme as desired without causing R to fall outside the range of the remaining six values. The estimate R is therefore able to withstand two extreme values on the left, or one extreme value on the right, or both, however extreme they may be. It cannot however handle more than this. For example, if there are two extreme values on the right, R will be affected. Indeed, if Y_8 (and hence Y_9) is made to tend to ∞, while the values $Y_1, \cdots, Y_7$ remain fixed at arbitrary values, then $R \to \infty$. Similarly $R \to -\infty$ if $Y_3 \to -\infty$, whatever be the fixed values of $Y_4, \cdots, Y_9$. In these circumstances, it seems natural to say that R can "tolerate" just two extremes on the left and one extreme on the right.

Let us now formalize this idea in a definition. Let V be an estimate based on a sample of size n. Suppose there exist integers $\alpha \geq 0$, $\beta \geq 0$ such that

$$Y_{\alpha+1} \leq V \leq Y_{n-\beta} \qquad \text{for all} \quad Y_1, \cdots, Y_n; \tag{11.2}$$

(11.3) whatever be the fixed values of $Y_{\alpha+2}, \cdots, Y_n$,

$$Y_{\alpha+1} \to -\infty \quad \text{implies} \quad V \to -\infty;$$

(11.4) whatever be the fixed values of $Y_1, \cdots, Y_{n-\beta-1}$,

$$Y_{n-\beta} \to \infty \quad \text{implies} \quad V \to \infty.$$

We shall then say that V can tolerate α extreme values on the left and β extreme values on the right. If, as often happens, $\alpha = \beta$, we shall denote their common value by γ. If it is desired to make explicit the dependence on n and V, we may write $\alpha_n(V)$ for α, and so on.

Our quantities α and β do not exist for all estimates. In particular, they are undefined for any estimate which can fall outside the sample range, since in this case no choice of α and β will satisfy (11.2). Examples of such estimates are given in theorem 1 below. Other examples are provided by certain Bayes estimates, which can be arbitrarily far from the sample, if the sample should fall sufficiently far from the region in which the population was believed, a priori, to lie. The measures α and β do however exist for two classes of estimates, as we shall now show.

First, let us consider the class of linear combinations of order statistics, say

$$L = w_1 Y_1 + \cdots + w_n Y_n. \tag{11.5}$$

The behavior of these estimates depends on the signs of the cumulative weights, say

$$\begin{aligned} A_i &= w_1 + w_2 + \cdots + w_i, \\ B_i &= w_{n-i+1} + w_{n-i+2} + \cdots + w_n, \end{aligned} \tag{11.6}$$

where $i = 1, 2, \cdots, n$.

THEOREM 1. *If $A_n \neq 1$, or if any A_i or B_i is negative, α and β do not exist. If $A_n = 1$ and all $A_i \geq 0$ and all $B_i \geq 0$, then $\alpha_n(L)$ is the smallest integer a for which $A_{a+1} > 0$, and $\beta_n(L)$ is the smallest integer b for which $B_{b+1} > 0$.*

PROOF. If $A_n \neq 1$, then $Y_1 = \cdots = Y_n = y \neq 0$ makes $L = A_n y$ fall outside the range. If $A_i < 0$ for some positive integer i, let $Y_1 = \cdots = Y_i = y < 0$ and $Y_{i+1} = \cdots = Y_n = 0$. Then $L = A_i y > 0 = Y_n$ so that L again falls outside the range. A similar argument covers $B_i < 0$, and the first statement of the theorem is proved.

Now suppose $A_n = 1 = B_n$, and $A_i \geq 0$ and $B_i \geq 0$ for all i. Since $A_n > 0$, there must be an integer a, $1 \leq a \leq n$, such that $A_1 = \cdots = A_a = 0$, $A_{a+1} > 0$, and hence $w_1 = \cdots = w_a = 0$ and $w_{a+1} > 0$. Similarly there exists b such that $w_n = \cdots = w_{n-b+1} = 0$ and $w_{n-b} > 0$. We see that $A_{n-b} = 1 = B_{n-a}$. We must verify (11.2)–(11.4) with a, b and L replacing α, β and V.

Since $A_i \geq 0$ and $Y_i \leq Y_{i+1}$, we must have

$$A_i Y_i + w_{i+1} Y_{i+1} \leq A_{i+1} Y_{i+1} \qquad \text{for} \quad i = 1, 2, \cdots, n-1. \tag{11.7}$$

Applying (11.7) inductively on i, we find

$$\begin{aligned} L &\leq A_{a+1} Y_{a+1} + w_{a+2} Y_{a+2} + \cdots + w_n Y_n \\ &\leq A_{n-b} Y_{n-b} + w_{n-b+1} Y_{n-b+1} + \cdots + w_n Y_n. \end{aligned} \tag{11.8}$$

The final term of (11.8) equals $A_{n-b} Y_{n-b} = Y_{n-b}$, hence $L \leq Y_{n-b}$. The analogous argument shows $Y_{a+1} \leq L$, so that (11.2) holds.

To verify (11.3), consider the first inequality of (11.8). If $Y_{a+2}, \cdots, Y_n$ are fixed arbitrarily and $Y_{a+1} \to -\infty$, the middle member of (11.8) will tend to $-\infty$, and therefore so must L. The analogous argument checks (11.4).

Several of the following corollaries are obvious, so proofs are omitted or only sketched.

Tukey has discussed the use of trimmed and Winsorized means for the location problem (see section 14 of [11] and the references there given). The (a, b)-trimmed mean is the arithmetic mean of the $n - a - b$ observations that remain after the a smallest and b largest observations have been removed. For example, the estimate R above is a (2, 1)-trimmed mean.

COROLLARY 1.1. *If R is an (a, b)-trimmed mean, then $\alpha_n(R) = a$ and $\beta_n(R) = b$.*

The (a, b)-Winsorized mean is the arithmetic mean of the n values after each of $Y_1, \cdots, Y_a$ has been replaced by Y_{a+1}, and each of $Y_{n-b+1}, \cdots, Y_n$ has been replaced by Y_{n-b}.

COROLLARY 1.2. *If Z is an (a, b)-Winsorized mean, then $\alpha_n(Z) = a$ and $\beta_n(Z) = b$.*

COROLLARY 1.3. *The mean $\overline{X}$ has zero tolerance for every n.*

PROOF. The mean $\overline{X}$ is the special case of L with each $w_i = 1/n$. As remarked above, $\overline{X}$ cannot tolerate even one extreme value.

COROLLARY 1.4. *The median $\tilde{X}$ has tolerance $\gamma_n(\tilde{X}) = [\frac{1}{2}(n - 1)]$.*

PROOF. We are using the customary notation, where $[u]$ means the greatest integer not greater than u. If n is even, say $n = 2k$, then $\tilde{X} = \frac{1}{2}Y_k + \frac{1}{2}Y_{k+1}$ so that by theorem 1, $\alpha = \beta = k - 1 = [\frac{1}{2}(2k - 1)]$. If n is odd, say $n = 2k + 1$, then $\tilde{X} = Y_{k+1}$ so that $\alpha = \beta = k = [\frac{1}{2}(2k + 1 - 1)]$.

It is clear from (11.2) that, whenever α and β exist, $\alpha + \beta \leq n - 1$. Therefore, γ, whenever it exists, must satisfy $2\gamma \leq n - 1$ or $\gamma \leq [\frac{1}{2}(n - 1)]$. Thus the median $\tilde{X}$ has the maximum possible value of γ, corresponding to the intuitive idea that it tolerates extreme values as well as is possible for any estimate that treats the two extremes symmetrically. (Of course other order statistics may be more tolerant of extremes on one side only.)

COROLLARY 1.5. *The tolerance $\alpha_n(Y_i) = i - 1$ and $\beta_n(Y_i) = n - i$.*

COROLLARY 1.6. *The sample midrange has $\gamma_n = 0$.*

COROLLARY 1.7. *The estimates discussed in section 5, where the weight of Y_i is proportional to $f(EY_i)$, have $\gamma = 0$. This includes the estimate W of section 5.*

As a second class of estimates for which α and β exist, consider the estimates S defined in (9.1), where for the moment we do not impose (9.2). Denote by $\mathcal{K}_a$ the subset of $\mathcal{S}$ consisting of those pairs $(i, j) \in \mathcal{S}$ for which $i \geq a$. Similarly, let $\mathcal{L}_b$ be the subset of $\mathcal{S}$ consisting of those pairs $(i, j) \in \mathcal{S}$ for which $j \leq b$.

THEOREM 2. *The tolerance $\alpha_n(S)$ is the largest integer a such that*

$$\frac{1}{2}\{\#(\mathcal{S}) + 1\} \leq \#(\mathcal{K}_{a+1}). \tag{11.9}$$

We have that $\beta_n(S)$ is the largest integer b such that

$$\frac{1}{2}\{\#(\mathcal{S}) + 1\} \leq \#(\mathcal{L}_{b+1}). \tag{11.10}$$

PROOF. The argument depends slightly on the parity of $\#(\mathcal{S})$; we shall give it for the even case, say $\#(\mathcal{S}) = 2q$. Then, if a is the largest integer for which (11.9) holds, $\#(\mathcal{K}_{a+1}) \geq q + 1$ and $q \geq \#(\mathcal{K}_{a+2})$. It is clear that $Y_{a+1} \leq M_{i,j}$ for every $(i, j) \in \mathcal{K}_{a+1}$. Therefore, Y_{a+1} is not greater than at least $q + 1$ of the $2q$ means, the median of which is S. It follows that $Y_{a+1} \leq S$. Similarly (11.10) implies $S \leq Y_{n-b}$ and (11.2) holds.

Now fix $Y_{a+2}, \cdots, Y_n$ and let $Y_{a+1} \to -\infty$. This implies $M_{ij} \to -\infty$ for every $i \leq a + 1$. At most $\#(\mathcal{K}_{a+2})$ of the $2q + 1$ means whose median is S can avoid tending to $-\infty$, and hence at least $q + 1$ means tend to $-\infty$, implying that $S \to -\infty$. This checks (11.3), and the argument for (11.4) is analogous.

If $\mathcal{S}$ satisfies the symmetry condition (9.2), then clearly $\alpha_n(S) = \beta_n(S) = \gamma_n(S)$. This applies to T, U, and D.

COROLLARY 2.1. *The tolerance* $\gamma_n(D) = [\frac{1}{4}(n - 2)]$.

The argument depends on $n \bmod 4$; we give it for $n = 4k$. Then $\#(\mathfrak{D}) = 2k$ and $\#(\mathcal{K}_{a+1}) = 2k - a$. The largest integer a for which (11.9) holds is $k - 1 = [\frac{1}{4}(2k - 2)]$.

COROLLARY 2.2. *We have*

$$\gamma_n(U) = [n - \tfrac{1}{2} - \tfrac{1}{2}\sqrt{2n^2 - 2n + 5}]. \tag{11.11}$$

PROOF. Here $\#(\mathcal{B}) = \frac{1}{2}n(n - 1)$ and $\#(\mathcal{K}_a) = \frac{1}{2}(n - a)(n - a + 1)$. We seek the largest integer a such that

$$\tfrac{1}{4}n(n - 1) + \tfrac{1}{2} \leq \tfrac{1}{2}(n - a - 1)(n - a). \tag{11.12}$$

In the range considered, the right side of (11.12) is a decreasing function of a, treated as continuous. Therefore $\alpha_n(U)$ is $[\tilde{a}]$ where $\tilde{a}$ is the root of the quadratic equation obtained by inserting an equality sign in (11.12).

COROLLARY 2.3. *We have*

$$\gamma_n(T) = [n + \tfrac{1}{2} - \tfrac{1}{2}\sqrt{2n^2 + 2n + 5}]. \tag{11.13}$$

The proof is similar.

We remark that $\gamma_{n+1}(U) = \gamma_n(T) \geq \gamma_n(U) \geq \gamma_n(D)$.

Table VIII compares the values of $\gamma_n(V)$ for $n = 1(1)20$ and $V = \tilde{X}, T, U$ and D.

TABLE VIII

TOLERANCE OF SEVERAL ESTIMATES

n	$\tilde{X}$	T	U	D	n	$\tilde{X}$	T	U	D
1	0	0	0	0	11	5	3	3	2
2	0	0	0	0	12	5	3	3	2
3	1	0	0	0	13	6	3	3	2
4	1	1	0	0	14	6	4	3	3
5	2	1	1	0	15	7	4	4	3
6	2	1	1	1	16	7	4	4	3
7	3	2	1	1	17	8	5	4	3
8	3	2	2	1	18	8	5	5	4
9	4	2	2	1	19	9	5	5	4
10	4	3	2	2	20	9	5	5	4

If $\lambda(V) = \lim_{n\to\infty} (\gamma_n(V)/n)$ exists, this limit represents the fraction of extreme values which the estimate V can tolerate at each end of a large sample. For large n,

$$\gamma_n(T) = \left(1 - \frac{\sqrt{2}}{2}\right)\left(1 + \frac{1}{2n}\right) + 0\left(\frac{1}{n^2}\right) = \gamma_n(U). \tag{11.14}$$

Thus

$$\lambda(U) = \lambda(T) = 1 - \frac{\sqrt{2}}{2} = .293. \tag{11.15}$$

For comparison, $\lambda(\tilde{X}) = .5$ and $\lambda(D) = .25$.

12. Comparison of several estimates

We conclude by comparing several estimates with regard to three desiderata: (i) efficiency in estimating the center of a normal population; (ii) tolerance of extreme values in the sense of section 11; and (iii) ease of computation. These are of course not the only considerations, but seem important ones. The comparison will be made primarily for sample size $n = 18$, since most is known for that case, but one may suppose that similar results hold for n near 18.

If only (i) is considered, then $\overline{X}$ has a strong claim to be the best estimate, as it is the optimum estimate for the normal location problem according to several criteria. This classical estimate is also easy to compute. However $\overline{X}$ has no tolerance of extreme values, and cannot be considered if (ii) is important.

If only (ii) is considered, then $\tilde{X}$ is the estimate of choice, as it maximizes γ. It is again easy to compute. However, it is poor on criterion (i), with $e(\tilde{X}) = .68563$ at $n = 18$, and $ae(\tilde{X}) = 2/\pi = .63662$. In most situations one would not pay so great a price in efficiency for what may be an unnecessarily great protection against extreme values.

If we require good performance on both (i) and (ii), the estimates T, U and D are all satisfactory. At $n = 18$, all have efficiency near .95 according to the sampling experiments, and $\gamma(T) = \gamma(U) = 5$, $\gamma(D) = 4$. The slightly lower tolerance of D may be balanced against the fact that it is considerably easier to compute than T or U, and if all three desiderata are considered, D is perhaps preferable to $\overline{X}$, $\tilde{X}$, T and U.

Finally, let us compare D with the symmetrically trimmed and Winsorized means. According to corollaries 1.1 and 1.2, these estimates achieve tolerance γ if we trim or Winsorize γ observations at each end. For $n = 18$, table I of [10] permits us to find the variances, and hence the efficiencies, of the (γ, γ)-trimmed and (γ, γ)-Winsorized means for each γ. The results are given in table IX. When $\gamma = 8$, both estimates coincide with $\tilde{X}$; when $\gamma = 0$, both estimates coincide with $\overline{X}$.

It has been pointed out by Dixon [2] that, for $n \leq 20$, the (γ, γ)-Winsorized mean has efficiency that agrees to three figures with the optimum attainable among all weighted averages of the order statistics which assign weight zero to the γ smallest and γ largest. By theorem 1, this implies that, to three figures, the values of efficiency of the Winsorized mean in table IX may be taken to be the efficiency of the optimum linear estimate with tolerance γ.

We see that, at $n = 18$, the efficiency of D is approximately the same as that of the optimum linear estimate with $\gamma = 2$. As $\gamma(D) = 4$, the estimate D provides substantially better tolerance of extreme values, while giving the same normal efficiency. Put another way, if we desire tolerance $\gamma = 4$ the efficiency of the optimum linear estimate is only .889 compared with $e(D) = .955 \pm .007$. This indicates that the restriction to linear estimates entails a substantial cost according to criteria (i) and (ii), and again suggests that the estimate D is worth consideration in the symmetric location problem.

TABLE IX

EFFICIENCIES WITH NORMAL SAMPLES OF 18

γ	Trimmed mean	Winsorized mean
0	1.00000	1.00000
1	.97462	.98116
2	.94084	.95581
3	.90367	.92501
4	.86429	.88896
5	.82314	.84749
6	.78030	.80021
7	.73535	.74649
8	.68563	.68563

REFERENCES

[1] DAVID BLACKWELL and J. L. HODGES, JR., "Geometrical proofs of path counts," unpublished (1965).

[2] W. J. DIXON, "Simplified estimation from censored normal samples," *Ann. Math. Statist.*, Vol. 31 (1960), pp. 385–391.

[3] WILLIAM FELLER, *An Introduction to Probability Theory and Its Applications*, Vol. I, New York, Wiley, 1957 (2d ed.).

[4] J. L. HODGES, JR., "Galton's rank-order test," *Biometrika*, Vol. 42 (1955), pp. 261–262.

[5] J. L. HODGES, JR. and E. L. LEHMANN, "The efficiency of some nonparametric competitors of the *t*-test," *Ann. Math. Statist.*, Vol. 27 (1956), pp. 324–335.

[6] ———, "Estimates of location based on rank tests," *Ann. Math. Statist.*, Vol. 34 (1963), pp. 598–611.

[7] ARNLJOT HØYLAND, "Numerical evaluation of Hodges-Lehmann estimates," *Det Kongelige Norske Videnskabers Selskabs Forhandlinger*, Vol. 37 (1964), pp. 42–47.

[8] JEROME KLOTZ, "Small sample power and efficiency for the one sample Wilcoxon and normal scores tests," *Ann. Math. Statist.*, Vol. 34 (1963), pp. 624–632.

[9] FREDERICK MOSTELLER, "On some useful 'inefficient' statistics," *Ann. Math. Statist.*, Vol. 17 (1946), pp. 377–408.

[10] A. E. SARHAN and B. G. GREENBERG, "Estimation of location and scale parameters by order statistics from singly and doubly censored samples," *Ann. Math. Statist.*, Part I, Vol. 27 (1956), pp. 427–451; Part II, Vol. 29 (1958), pp. 79–105.

[11] JOHN W. TUKEY, "The future of data analysis," *Ann. Math. Statist.*, Vol. 33 (1962), pp. 1–67.
[12] The RAND Corporation, *A Million Random Digits with 100,000 Normal Deviates*, Glencoe, Free Press, 1955.