LESS VULNERABLE CONFIDENCE AND SIGNIFICANCE PROCEDURES FOR LOCATION BASED ON A SINGLE SAMPLE : TRIMMING/WINSORIZATION 1

By JOHN W. TUKEY and DONALD H. McLAUGHLIN
Princeton University[1]

SUMMARY. The vulnerability of Student's t, insofar as efficiency and power are concerned, leads to consideration of substitutes. Among the most promising are ratios of trimmed means to square roots of suitable quadratic forms involving the same order statistics. Matching, across underlying distributions, of ratios of average of denominator to variance of numerator leads to selection of the Winsorized sum of squared deviations as the basis for a denominator. The resulting *trimmed* t should prove more useful when the amount of trimming is made to depend on the individual sample in a suitably prescribed manner. Exact critical values for the resulting *tailored* t seem to require Monte Carlo computation, but use of a simple modified denominator for trimmed t allows us to use the conventional t tables as a reasonable approximation.

1. Introduction

One of us (Tukey, 1962, p.16) has already summarized the advantages of the class of symmetric distributions as a natural first step in our progress from statistical techniques understood and known to be useful for Gaussian (= normal) distributions alone to techniques understood and useful in very much more general situations. Once we are firmly established with statistical techniques understood and known to be useful for symmetrical distributions, it will be time to take a further step. But the problem of adequate mastery of the symmetrical-distribution case is enough for the moment.

Indeed it seems enough for the present to confine ourselves to techniques which are not only symmetric in their action on the class of all samples, but are symmetric in their action on individual samples. Such a restriction is clearly both less important and more easily removed than the restriction to symmetrical distributions.

The present account confines itself further : (i) to the single-sample-for-location problem and (ii) to the first steps of a specific approach to that problem. It describes the results of certain easily accessible calculations, and outlines what appear to be the plausible next steps, as well as indicating a little of what is known, or believed to be true, beyond this scope.

2. The normal and the pathological : which is which ?

The Gaussian or Laplacian distribution, to the physicist the Maxwellian distribution, has long been known to the statistician as the normal distribution. However little noticed such a commonly-used name becomes, shades of its original meaning continue to cling—distributions that are not normal are, by at least slight implication, pathological. In the early stages of development or exploration of a particular aspect

[1] Research supported by U.S. Army Research Office (Durham). Reproduction in whole or part for any purpose of the U.S. Government permitted.

of statistics or data analysis, such an attitude may promote progress. But in a well-developed area such an attitude can only be respectable if it reflects the facts—if the *usual* is at least close to the "normal" in behaviour. This is not the case in almost all of the instances of data analysis which the writers—and, they believe, most practising statisticians, have come in contact. The typical distribution of errors and fluctuations has a shape whose tails are longer than that of a Gaussian distribution. (See Tukey (1960) for more extended discussion, and Mandelbrot (1960, 1961, 1962) for newer instances arising in economics.)

It is the Gaussian distribution that has to be regarded as somewhat pathological from the standpoint of practice. And distributions with shorter tails, while they do occur, are rather more pathological. Thus frequency of occurrence directs our attention to longer-tailed distributions.

There is another, quite distinct and independent, reason for emphasizing long-tailed distributions. The prevalence of minimax-loss approaches to uncertainty is not an accident. We all tend to have more interest in avoiding a large loss than in obtaining a large gain. If procedures optimum for Gaussianity are used against long-tailed distributions they tend to behave poorly, both relatively and absolutely. Their quality is usually not only far from optimal (for the specific situation) but also of far lower quality than would have been the case if the underlying distribution were normal.

Against short-tailed distributions, on the other hand, procedures optimum for Gaussianity are not infrequently relatively poor but absolutely good, in the sense that, while the optimum procedure for the specific situation would do much better, the performance of the Gaussianly optimum procedure will be better for short-tailed distributions than for Gaussian ones.

The use of s^2 as an indicator of scale is, of course, an outstanding instance of the behaviour just discussed, both for short-tailed and long-tailed distributions.

Accordingly, it seems appropriate to begin by giving major attention to symmetric distributions with tails longer than the Gaussian.

3. Measures of quality and insensitivity

Statistical procedures are identified among the more general procedures of data analysis, which themselves may or may not be based on a probability model, by the fact that they take explicit account of uncertainty. From a narrow viewpoint, the most important aspect of such a procedure is its *validity*, the extent to which any associated statements of probability are correct, or at least conservative. Does the nominal significance level really apply? Does the formal confidence interval have (at least) the asserted probability of covering the true value? When such questions are asked about the behaviour for other underlying distributions of a procedure calibrated for Gaussianity, the conventional term is "robustness." If we wish to be clear and specific, we should—and shall—speak of "robustness of validity."

But only a slightly broader view causes us to ask of a procedure not only "Is it valid?" but "Is it efficient?". While controlling its rate of error, does it do as well, say as much, extract as much from the data, etc., as it can? And when we ask such questions about the behaviour for other underlying distribution of a procedure first developed for (near) Gaussianity we are asking about "robustness of efficiency."

So long as we are to continue to use Gaussian underlying distributions as the standard of calibration and the natural starting point, thus assuring validity in the Gaussian situation,—and this seems likely to be a long, long, time—these arguments suggest that we should give major attention for the present to

(1) efficiency for Gaussian distributions,

(2) robustness of validity for long-tailed symmetrical distributions,

(3) robustness of efficiency for long-tailed symmetrical distributions, where the first probably deserves by far the least attention of the three.

4. Location from a single sample : competitors and challenges

If we are given a single (random) sample, $y_1, y_2, \ldots, y_n$ of observations from $dF(y-\mu)$ where $F(v)+F(-v) = 1$ (so that the distribution of y is symmetric around μ), a statistical technique which permits tests of significance also provides confidence statements, and vice versa.

The most classical technique for this problem is one of many sorts of uses of Student's t. Its rather moderate robustness of validity has been studied by a number of workers (Pearson, 1929; Rietz, 1939; Gayen, 1949; Bradley, 1952). (In two-sample and simple analysis-of-variance situations Student's t is much more robust.) What we know about its behaviour can be summarized as follows :

(a1) The average value of the square of its denominator is in fixed ratio to the variance of its numerator, independent of the underlying distribution.

(a2) Its robustness of validity for symmetric underlying distributions is moderate, being quite high for significance or diffidence levels of 30-40% (Gayen, 1949) but not as satisfactory for the usual 5%, 1%, etc. levels. (Its behaviour for unsymmetric underlying distributions is much less satisfactory.)

(a3) Contrary to most naive intuitions, confidence and significance will be over-estimated by Student's t, not when the underlying distribution is longer-tailed (as Gayen, 1949; Bradley, 1952; and Wonnacott, 1963 all agree) but rather when the underlying distribution is shorter-tailed (as Rider, 1929; Perlo, 1933; Laderman, 1939; and Gayen, 1949 all agree).

(a4) Its robustness of efficiency is subject to serious question, especially since a single wild-appearing observation can seriously affect both $\bar{y}$ and s.

(a5) The method of its calculation can easily be extended (or analogized) to a very wide variety of situations without requiring changes in (Gaussian-theory) critical values for this reason.

(a6) It provides confidence limits with little more effort than significance tests.

(a7) If the underlying distribution should be Gaussian, these procedures will be optimal according to almost every criterion.

Toward the other extreme we find techniques which *can* be based on ordering deviations (of y's from a contemplated central value M) according to magnitude and basing the test (or confidence interval) upon the pattern of signs of deviations, which we will call the *sign-configuration.* This is most frequently and simply done by sorting the ranks of one sign in some simple way.

Scoring each rank with its rank number is called the one-sample Wilcoxon or signed-rank procedure and was introduced by Wilcoxon (1945, 1946, 1947, 1949). (For more available expositions see Moses, 1953; or Siegel, 1956, where the procedure is applied to differences of paired observations.) In his thesis, Walsh (1947, 1949) demonstrated a result equivalent to the fact that the probability of obtaining any configuration of signs of deviations (when the deviations are ranked by magnitude) is the same for all symmetric underlying distributions. (For a clarification of the relationship of his results to Wilcoxon's see Walsh, 1959.)

The most important aspects of our knowledge of such sign-configuration procedures can be summarized as follows :

(b1) Their robustness of validity is perfect for symmetric distributions.

(b2) Their robustness of efficiency has not been adequately studied.

(b3) The method of their calculation does not seem to be trivially extendable to more general situations (such as regression coefficients); new tables of critical values seem almost certain to be needed in any such extension.

(b4) The calculation of confidence intervals requires appreciably more effort than significance testing, although trial-and-error is not required (Moses, 1953), a simple graphical approach sufficing. (In more general situations, trial-and-error seems likely to be necessary.)

(b5) If the underlying distribution should be Gaussian, the efficiency of various of these procedures will be very high (Klotz, 1962), the loss of efficiency in comparison with the "optimum" t-procedure being almost negligible in this situation.

If we were only concerned with the single-sample problem, and were able to settle the question of robustness of efficiency for some sign-configuration procedure favourably, it would be reasonable to argue that such a procedure was a reasonable choice for routine work. Its only major defect would be its tendency, because of difference in labour, to cause its users to stop with significance tests in many instances where it would be profitable for them to push on to confidence intervals.

If we are to seek a new procedure, it should be one combining many of the relative advantages of t-procedures and sign-configuration procedures. As an ideal, possibly utopian, we might seek a procedure with these properties for symmetric underlying distributions :

(c1) Its robustness of validity is high.

(c2) Its robustness of efficiency is satisfactorily large.

(c3) Its method of calculation can be rather easily extended to a wide variety of situations with little change in critical values.

(c4) It provides confidence intervals almost as easily as significance tests.

(c5) If the underlying distribution should be Gaussian, its efficiency is high.

If we could find a procedure which met, or came close to meeting, these specifications, it would be an outstanding candidate for adoption as the method of choice for routine use.

5. A desirable direction of exploration

The criteria toward which we hope to make progress are diverse in kind and character—it would be unrealistic to expect any formal optimization procedure to actually lead us toward our goal. Accordingly, we have a choice between trying to modify the things we understand, or seeking to be struck by the lightning of a purely new approach. While waiting for lightning, we may as well proceed with modification, which appears to be more easily accomplished starting with Student's t.

When reached through a formal optimization procedure, Student's t arises as a single, integrated creation—and thus offers little guidance for modification. But Student's t did not first arise in such a way. Its numerator and denominator have very different conceptual origins, namely :

(d1) The numerator came first, as an intuitively effective point estimate of deviation from contemplated value.

(d2) The denominator followed, as something which reflected (indeed, was an estimate of) the variability of the previously chosen numerator.

One road—to the writers the currently most promising road—toward the goal set out in (c1) to (c5) is thus to begin by choosing a modified numerator, and then to seek a matching denominator.

In the case of Student's t, where attention was concentrated on an underlying Gaussian distribution — as was wholly appropriate when breaking new ground— "matching" needed only to refer to this situation. The fact that the denominator continued to estimate the variance of the numerator for all distributions was only a bonus, albeit one that proved to be very important. In the present approach, "matching" must refer to at least a modestly wide variety of symmetric distributions.

Granted that both frequency of occurrence and intensity of danger should cause us to give particular attention to longer-tailed distributions, our modification of the numerator must lie in the direction of attaching less weight to extreme—more precisely extreme-appearing—observations.

6. Trimming and Winsorizing

There are two sorts of simple modifications of an arithmetic mean which especially deserve consideration in this context, both for reasons of simplicity and for reasons arising from analyses of mathematical models to be reported elsewhere.

Given n *ordered* observations

$$y_1 \leqslant y_2 \leqslant \cdots \leqslant y_n \qquad \ldots \quad (1)$$

the (unweighted or equally weighted) arithmetic mean $\bar{y}$ or $y_{\bullet}$ is given by

$$y_{\bullet} = \frac{1}{n}\,(y_1+y_2+\ldots+y_n) = \Sigma\, y_j/\Sigma 1. \qquad \ldots \quad (2)$$

If $n = g+h+g$ (this mode of expression is chosen instead of $n = 2g+h$ to stress ordering), the g-times (symmetrically) trimmed mean y_{Tg} is given by

$$y_{Tg} = \frac{1}{n-2g}\,(y_{g+1}+y_{g+2}+\cdots+y_{n-g}) = \sum_{(Tg)} y_j \Big/ \sum_{(Tg)} 1 \qquad \ldots \quad (3)$$

and is the arithmetic mean of the set of h numbers obtained by dropping both the g lowest and the g highest values from the y_j. Clearly y_{Tg} pays less attention to extreme values than does $y_{\bullet}$, which may be regarded as the 0-fold trimmed (= untrimmed) mean.

A less intuitive contender is the g-times (symmetrically) Winsorized mean, y_{Wg}, given by

$$y_{Wg} = \frac{1}{n}\,(g{\cdot}y_{g+1}+y_{g+1}+y_{g+2}+\cdots+y_{n-g}+g{\cdot}y_{n-g}) = \sum_{(Wg)} y \Big/ \sum_{(Wg)} 1 \qquad \ldots \quad (4)$$

which is the arithmetic mean of the n values obtained by replacing (i) each of the g lowest y's by the value of the nearest other y, namely y_{g+1} and (ii) each of the g highest y's by the value of the nearest other y, namely y_{n-g}. Again we have paid less attention to individual extreme y's, but we have managed not to divert our attention from the "tails" of the sample so thoroughly. Instead of replacing each deleted y_j by y_{Tg}, which is one reasonable interpretation of the calculation of y_{Tg}, since

$$y_{Tg} = \frac{1}{n}\,(g{\cdot}y_{Tg}+y_{g+1}+y_{g+2}+\cdots+y_{n-g}+g{\cdot}y_{Tg}), \qquad \ldots \quad (5)$$

we have only replaced each deleted y by the nearest retained y. (As noted by Dixon (1960), this procedure has been called Winsorization in honour of Charles P. Winsor, who actively sponsored its use in actual data analysis.)

We shall find it convenient to continue to use the notation just illustrated throughout the discussion that follows, namely

(e1) Replacement of a subscript by a "$\bullet$" indicates a simple arithmetic mean.

(e2) Replacement of a subscript by "Tg" indicates a g-times trimmed mean.

(e3) Replacement of a subscript by "Wg" indicates a g-fold Winsorized mean.

(e4) The indication "(Tg)" on a summation sign indicates (g-times) trimmed summation = summation over the subscripts remaining after the sample of y's is trimmed g times on each tail.

(e5) The indication "(Wg)" on a summation indicates (g-times) Winsorized summation = summation over the subscripts for which the y's were not deleted, repeating each extreme undeleted subscript $g+1$ times and each other subscript once.

If we were concerned with but one amount of trimming or one amount of Winsorization we could well make use of simpler notations. (One such has been suggested at the end of (b2) on page 12 of Tukey, 1960.) But when concerned with several values of g it seems advisable to use a more explicit notation.

7. Choice of numerator

For underlying distributions whose shapes are very close to Gaussian, the Winsorized means are less variable than trimmed means. While the efficiency for Gaussianity of trimmed means is quite high, the fractional loss being crudely $2g/3n$ (corresponding to efficiency of about 2/3 for the median), that of the corresponding Winsorized means is much higher. At the other extreme, where very long-tailed distributions are involved, trimmed means are clearly more efficient than Winsorized means. Where does the transition take place ?

At the time when a previous account was written, preliminary analysis suggested a moderately broad scope for the Winsorized mean (cf. Tukey, 1962, p. 18). Now that further analysis (to be reported elsewhere) has been carried out, it would appear that the trimmed mean is likely to be more widely useful than had been supposed. As we shall see, this change in interpretation makes the present programme more attractive.

8. Criteria of matching

In striving to "match" denominators to a given numerator, we must choose a criterion of matching. The natural, and we believe reasonable choice, is to begin by following the example of Student's t and ask that

average value of denominator squared

and variance of numerator

should be in constant proportion over as broad a spectrum of symmetrical distributions as is reasonably convenient. This is, again, *only a first step*. Once we find a denominator which matches well in this sense, we are ready to calculate critical values of the ratio for various symmetrical distributions, and learn whether its validity is really robust.

The numerators we consider are linear combinations of order statistics; their variances will be linear combinations of variances and covariances of order statistics. And the squared denominators are likely to be quadratic functions of order statistics; their average values will depend on averages, variances, and covariances of order statistics. Accordingly we naturally begin by turning to those symmetrical distributions for which low moments of order statistics are available.

Three distributions are outstanding in this regard :

(f1) the rectangular distribution, for which low moments are available in closed form,

(f2) the Gaussian distribution, for which 1st and 2nd moments are available for sample sizes $\leqslant 20$ (Teichroew, 1956; Sarhan and Greenberg, 1956; see alternatively Teichroew, 1962),

(f3) one long-tailed distribution—the lambda distribution with $\lambda = -0.1$ —called the "special distribution" by Hastings, Mosteller, Tukey, and Winsor (1947), for which they provided 1st and 2nd moments for sample sizes $\leqslant 10$.

In addition to these distributions, published tables of low moments of order statistics from symmetric distributions seem restricted to

(f4) the isosceles triangular distribution (Sarhan, 1954, p. 320) for sample sizes up to 5, and

(f5) the double exponential distribution (Sarhan, 1954, p. 320) for sample sizes up to 5.

In view of the central importance of the Gaussian distribution, and the ease of handling order statistic moments for the rectangular distribution, the natural course is to begin by trying to "match" for these two distributions and, once reasonable success is obtained, to check the match for the other distributions, giving special emphasis to the match for distributions longer-tailed than the Gaussian.

Work on the preparation of extensive tables of order statistic moments for lambda distributions is in progress. When these are available, somewhat better checks on matching will be possible. However, since the ultimate check is in terms of % points of the ratio rather than in terms of comparing individual values of numerator and denominator, the need for such further checks is not great.

9. Denominators matched to the trimmed mean : early trials

The denominator most naively associated with the trimmed mean, y_{Tg}, is (some multiple of) the formal standard deviation of the trimmed sample, whose square is proportional to the (g-times) trimmed sum of squared deviations (SSD)

$$\text{SSD}_{Tg} = \sum_{(Tg)} (y_i - y_{Tg})^2. \qquad \cdots \quad (6)$$

The most convenient of the ratios that must be nearly constant if we are to have matching is

$$\frac{\text{ave SSD}_{Tg}}{\text{var } y_{Tg}} = \text{divisor}_1(g+h+g) \qquad \cdots \quad (7)$$

where the name "divisor" is justified by the fact that

$$\sqrt{\frac{\text{SSD}_{Tg}}{\text{divisor}_1\,(g+h+g)}} \qquad \cdots \quad (8)$$

is the natural normalization to an actual denominator for use with y_{Tg}. When we investigate the behaviour of

$$\frac{\text{Gaussian divisor}_1\,(g+h+g)}{\text{rectangular divisor}_1\,(g+h+g)} \qquad \dots \quad (9)$$

where "Gaussian" and "rectangular" specify the distributions for which "ave" and "var" are calculated = the underlying distributions for which the appropriate forms of (8) yield denominators the average of whose square equals the variance of y_{Tg}, we find that (9) is moderately, but probably not satisfactorily close to unity, as Table 1 shows. For more detail, see McLaughlin and Tukey (1961).

TABLE 1. VALUES OF THE RATIO (9) FOR SELECTED g AND h

h=size of sample after trimming	g=number of observations trimmed from each end					
	1	2	3	5	8	9
2	1.009	1.007	1.005	1.003	1.002	1.001
3	1.002	1.002	1.002	1.001	1.001	
4	.995	.996	.997	.998	.999	
5	.987	.989	.991	.995		
9	.966	.967	.971	.979		
10	.963	.962	.967	.976		
15	.951	.947				
18	.948					

Reflection upon these results showed that, for better matching, the denominator should be modified in such a way as to give more attention to the outlying portions of the sample. This led to trials of various alternatives such as using SSD_{Tg^*}, where $g^* < g$, with y_{Tg} so that a less-trimmed sample provided the denominator. Trial and consideration eventually led to investigation of

$$\text{SSD}_{Wg} = \sum_{(Wg)} (y_i - y_{Wg})^2 \qquad \dots \quad (10)$$

the equally-many-times *Winsorized* sum of squares of deviations as a basis for a denominator to be used with the *trimmed* mean y_{Tg}.

The matching of the corresponding divisor for different distributions

$$\text{divisor}_2\,(g+h+g) = \frac{\text{ave SSD}_{Wg}}{\text{var } y_{Tg}} \qquad \dots \quad (11)$$

has to be examined. The first check is again of

$$\frac{\text{Gaussian divisor}_2\,(g+h+g)}{\text{rectangular divisor}_2\,(g+h+g)} \qquad \dots \quad (12)$$

with the results shown in Table 2.

TABLE 2. VALUES OF THE RATIO (12) FOR SELECTED g AND h

h=size of sample after trimming	g=number of observations trimmed from each end					
	1	2	3	5	8	9
2	1.0092	1.0071	1.0053	1.0031	1.0017	1.0014
3	1.0052	1.0045	1.0036	1.0023	1.0013	
4	1.0037	1.0033	1.0027	1.0018	1.0011	
5	1.0031	1.0023	1.0021	1.0015		
9	1.0026	1.0021	1.0015	1.0009		
10	1.0026	1.0020	1.0015	1.0009		
15	1.0024	1.0020				
18	1.0023					

The line for $h = 2$, where only the 2 central observations are retained in either the trimmed or Winsorized samples, must be the same in Tables 1 and 2. Elsewhere in Table 2, only one entry rises as much as 0.5% above unity. (As compared with a fall of more than 5% for the t-denominator.) And since divisor$_2(g+h+g)$ will appear under a square root, this corresponds to a suggested difference in critical value between rectangular and Gaussian of 1 part in 400. (As much as almost 1 in 200 when only the two central values survive trimming.) For the direction of easy computation but of lesser importance, the direction of shorter tails than Gaussian, the suggested behaviour of divisor$_2$ $(g+h+g)$ is close to excellent. What of the other side?

Table 3 presents the available values for

$$\frac{\text{special divisor}_2\,(g+h+g)}{\text{rectangular divisor}_2\,(g+h+g)} \qquad \ldots \quad (13)$$

where "special" refers to the lambda distribution with $\lambda = -0.1$, (cf. Hastings, Mosteller Tukey, and Winsor (1947)), which can be roughly thought of as a t distribution with 5 degrees of freedom. Most of the values are close to 1.01, a value which suggests a 0.5% difference in critical values between the rectangular and the special.

TABLE 3. VALUES OF (13) FOR AVAILABLE g AND h

h=size of sample after trimming	g=number of observations trimmed from each end			
	1	2	3	4
2	1.025	1.018	1.013	1.010
3	1.016	1.012	1.009	
4	1.013	1.009	1.007	
5	1.011	1.008		
6	1.010	1.007		
7	1.010			
8	1.010			

The remaining easy comparisons are made in Table 4, which offers no reason to change the conclusions already reached.

TABLE 4. OTHER RATIOS OF VALUES OF divisor$_2$ $(1+h+1)$

distribution compared with the rectangular	h=size of sample after trimming	distribution divisor$_2$ / rectangular divisor$_2$
Isoceles triangular	2	1.007
	3	1.014
double exponential	2	1.041
	3	1.037

The "suggestion" of a somewhat larger divisor for longer-tailed distributions requires some consideration at this point. What we are finding is that the divisor, defined as

$$\frac{\text{ave (denominator)}^2}{\text{var numerator}} \qquad \ldots \quad (14)$$

increases slightly as we pass from a very short-tailed distribution to a quite long-tailed one. In the case of Student's t, this ratio does not change at all, but Gayen's results suggest that, for modest numbers of degrees of freedom, the corresponding critical values change (decrease) by several times the fractions with which we are concerned. Since we would not expect this effect to be so prominent for trimmed and Winsorized statistics, it seems likely that the matching behaviour of SSD_{Wg}, as a denominator, at least so far as divisor$_2$ $(g+h+g)$ is concerned, is all that we can ask at this stage.

11. Approximation to the denominator—trimmed t

Depending upon the numerical behaviour of the divisor, we have a number of choices in putting the resulting procedure into approximate practice. If its behaviour is simple enough, we may calculate a "trimmed t" using a convenient approximation to the divisor, and then compare the results with, as an approximation, the (normal-theory) critical values of Student's t, or, more precisely, with modified critical values appropriate to the precise distribution (say on normal theory) of "trimmed t". Whether or not the behaviour of the divisor is simple, we can always choose a convenient working divisor and then tabulate the appropriate critical values. Clearly the first of these possibilities is the more desirable. Does divisor$_2$ behave simply?

Table 5 shows the ratio of the normal-theory values of divisor$_2$ to $h(h-1)$, all values falling between 1.00 and 1.02. (For rectangular-theory values, see Section 19.) Clearly we will do quite well to use $h(h-1)$ as the working divisor, especially when we recall that using a slightly undersized divisor corresponds to using slightly longer-confidence intervals, and is thus slightly conservative.

TABLE 5. RATIO OF NORMAL THEORY divisor_2 TO $h(h-1)$

h=size of sample after trimming	g=number of observations trimmed from each end					
	1	2	3	5	8	9
2	1.009	1.007	1.005	1.003	1.002	1.001
3	1.016	1.015	1.013	1.009	1.006	
4	1.016	1.016	1.015	1.011	1.008	
5	1.015	1.016	1.015	1.012		
9	1.010	1.012	1.012	1.011		
10	1.010	1.011	1.011	1.010		
15	1.007	1.008				
18	1.006					

The next question, of course, has to do with the approximate distribution of the result, particular in the vicinity of the conventional tail areas. To help us with this problem there are two pieces of information :

(g1) Dixon and Tukey (1963) have studied the approximate distribution of Winsorized t, where y_{Wg} (rather than y_{Tg}) is combined with s_{Wg}. The results of this study indicate a quite Student's-t-like distribution, with the best fit obtainable for a number of degrees of freedom typically somewhat less than $h-1$(= the number corresponding to a sample of size equal to the trimmed sample).

(g2) R. A. Jensen has made some preliminary experimental sampling and Monte Carlo investigations into the critical values of t_T. These suggest, rather than indicate, that these critical values may be rather close to those of Student's t for $h-1$ degrees of freedom.

Accordingly, if we wished to use trimmed t on an approximate basis (but see Sections 14 to 16), we would calculate

$$t_{Tg} = \frac{y_{Tg}-M}{\sqrt{\mathrm{SSD}_{Wg}/h(h-1)}} \qquad \dots \quad (15)$$

where M is a contemplated value for the center of the distribution sampled, and refer the result to Student's t on $h-1$ degrees of freedom.

And we could, as a next step, proceed to a more precise determination of actual critical values for (15).

12. Matching to the Winsorized mean : a query

Turning now to the Winsorized mean, analogy with the results just described leads us to begin with denominators which pay more attention to the tails of the sample than does the numerator. Working with the values of the trimmed sample alone, an attempt to match drives us rapidly to the (g-times) inner range.

$$W_{Tg} = y_{n-g} - y_{g+1} = W_{Wg} \qquad \dots \quad (16)$$

The corresponding divisor

$$\text{divisor}_3\,(g+h+g) = \frac{\text{ave}(W_{Tg}^2)}{\text{var}\; y_{Wg}} \qquad \dots \quad (17)$$

does not appear to be so satisfactorily constant as we change the underlying distribution, as Table 6 shows.

TABLE 6. VALUES OF $\dfrac{\text{Gaussian divisor}_3\,(g+h+g)}{\text{rectangular divisor}_3\,(g+h+g)}$

h=size of sample after trimming	1	2	3	5	8	9
2	1.009	1.007	1.005	1.003	1.002	1.001
3	.962	.985	.989	.994	.997	
4	.961	.963	.971	.982	.990	
5	.950	.945	.953	.968		
9	.961	.914	.900	.922		
10	.970	.914	.903	.913		
15	1.022					
18	1.066					

The most natural ways to seek improvement are (i) the use of observations *not in* y_{Wg} to help assess its variability, (ii) the use of a denominator in which differences among central observations are *subtracted* from W_{Tg} in order to further emphasize the behaviour of the ends of the trimmed sample. Both of these have strong heuristic disadvantages : the first because we may lose the advantage of getting wholly free of the observations excluded from the trimmed sample, with the consequence that, in very long-tailed distributions, we may improve our numerator rather more than we improve our estimate of its performance; the second because such subtractions, if effective, must tend to decrease the relative stability of the denominator (measured, if you will, in "effective degrees of freedom").

Whether either of these approaches, or some other, such as using

$$W_{Tg} + A \cdot |y_{Wg} - y_{Tg}| \qquad \dots \quad (18)$$

for a suitable value of A, will prove satisfactory, seems likely to be better investigated by working with actual distributions of ratios and the corresponding critical values rather than with ratios of moments. If this be so, it will be well to gain experience first with the distributional behaviour of t_{Tg}.

13. Power of trimmed t

In connection with a study of Monte Carlo methods adapted to statistical problems of sampling from non-normal distributions, Wonnacott (1963) has investigated some selected aspects of the power of trimmed t both when the underlying distribution is Gaussian and when it is a symmetrical Johnson (1949) distribution with low moments corresponding to a t-distribution with 6 or 4.7 degrees of freedom. In general his results are as would have been expected, though the comparison of Student's t and Wilcoxon-Walsh procedures is, surprisingly, somewhat unfavourable to the latter.

A few highlights of Wonnacott's (1963) comparison of powers for non-normal underlying distributions are these :

(h1) Singly trimmed t for $n = 10$ is more powerful than Student's t for the Johnson distribution with moments matching t_6.

(h2) Five-times trimmed t for $n = 20$ is very close to Student's t for the same underlying distribution.

(h3) Three-times trimmed t for $n = 10$ is almost as powerful as the nearest Wilcoxon-Walsh procedure for the Johnson distribution with moments matching $t_{4\cdot 7}$.

In general, the prospects for the effectiveness of t_{Tg} seem very good.

14. What should we expect to be used ?

Once we have obtained all desired critical values of t_{Tg}, what are we to do in practice? Let us suppose the error rate (here = significance or diffidence level) at which we are going to work has been fixed, once for all. This assumption is of course unrealistic, but it enables us to look more clearly at the issues which concern us most immediately. We may as well also fix $n = g+h+g$.

Let, then,

$$a_g = \text{critical value of } t_{Tg} \qquad \dots \quad (19)$$

and consider a man with a single sample of y's. He has a wide variety of choices. He may take any of :

point estimate	*interval estimate*
$y_\bullet$	$y_\bullet \pm a_0\sqrt{\text{SSD}/n(n-1)}$
y_{T1}	$y_{T1} \pm a_1\sqrt{\text{SSD}_{W1}/(n-2)(n-3)}$
y_{T2}	$y_{T2} \pm a_2\sqrt{\text{SSD}_{W2}/(n-4)(n-5)}$
y_{T3}	$y_{T3} \pm a_3\sqrt{\text{SSD}_{W3}/(n-6)n-7)}$

$$\dots \quad (20)$$

We know that, if his underlying (symmetrical) distribution behaves "reasonably" (in quotes because we do not yet know what is "reasonable" and what is not), systematic use *without exception* of any *single* one of the interval estimates will offer him validity, those with larger g probably being more robust in this regard. But our motivation in entering upon this whole question was to improve efficiency while maintaining validity. And we are almost certain that the qualitative behaviour of relative efficiency will appear as in Figure 1.

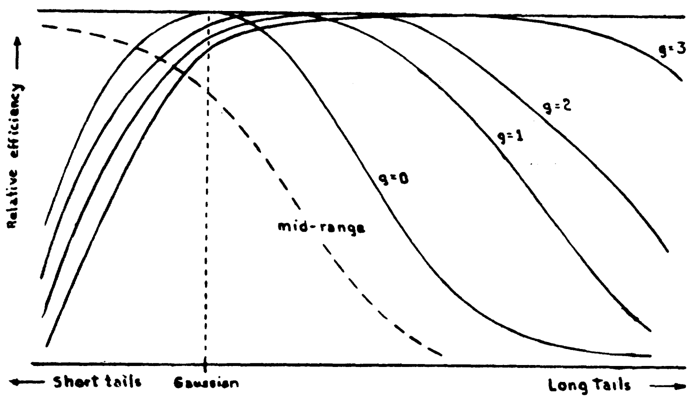

Fig. 1. Anticipated qualitative behaviour of relative efficiency.

Knowing this, it is most unlikely that the user will be content to use any single value of g *without exception*. When his samples are long-tailed, thus appearing to come from long-tailed distributions, he will want to use values of $g > 0$. When they look "Gaussian" he will wish to use $g = 0$ (=Student's t). (And it is conceivable that some will feel that they sometimes really do have an underlying distribution with shorter tails, and will wish to try to take advantage of this by occasional use of a numerator more of the character of the mid-range. But this, for the reasons discussed in Section 5, is well treated as a second-order perturbation.)

15. Is this wise ?

At first glance it looks as though such a user has started upon a dangerous course. All those who have handled much data know, often largely instinctively, how dangerous it would be to trust a single small sample to tell us with any precision about the shape of the distribution from which it came. And it tends to appear that this is what a man is doing when he picks a g in the light of the specific sample to which he is to apply the chosen procedure.

Let us turn to a simpler, classical situation. What of the man with a sample of, say, size 3 from a distribution known to be Gaussian? He has an estimate of σ^2 based on 2 degrees of freedom. If he believes that he "knows" σ^2 from other evidence, and acts accordingly, he is often in bad trouble. If he makes a significance test or states a confidence interval as if σ^2 were known, he will indeed be incurring much larger risks than he claims. It has been 55 years since Student (1908) showed us the way out of this dilemma. We have only to plan to use an estimate, s^2, of σ^2 based on these observations, to admit that this choice will be fallible, and to ask how must we readjust the critical values to allow for this fallibility. Doing this for the use of s^2 for σ^2, which takes us to the distribution of Student's t, is a familiar operation, now regarded as logically simple. The only possible difficulties are computational ones associated with the calculation of appropriate critical values, once the rule for choosing the estimate of σ^2 is fixed. (Thus, for example, the answers for Lord's t (Lord, 1947, 1950) where this is based upon range, are somewhat different from those for Student's t.)

The situation for allowance for long tails is a similar one. Once we fix a way for the user to choose a given value of g—and decide what the *form* of the adjustment to the critical value will be—no logical difficulty remains. There is a system of critical values which will enable a man who behaves as specified, *without exception,* to make statements which will be valid (for appropriate underlying distributions). The only difficulty is a purely computational one—find the *amount* of the needed change in critical values.

16. Individually trimmed, t = tailored t

The selection of the exact procedure for choosing a value of g, so long as this procedure is sensible, is a matter of little importance. At first glance there are at least two plausible alternatives, namely :

$$\text{Choose that } g \text{ which minimizes } \mathrm{SSD}_{Wg}/h(h-1) \qquad \dots \quad (21)$$

and

$$\text{Choose that } g \text{ which minimizes } a_g \cdot \sqrt{\mathrm{SSD}_{Wg}/h(h-1)}. \qquad \dots \quad (22)$$

However, it is only for distributions with the very longest tails that it will be sensible to choose values of g that discard almost all the sample values. (The cost of discarding so many sample values will not be paid in terms of the stability of y_{Tg}, which cannot be made worse than the stability of the median—which is excellent in sampling from long-tailed distributions—but rather in terms of the stability of the denominator—stability of the estimate of variability of y_{Tg}.) The simplest means of avoiding difficulty in this connection will be to introduce a fixed function $G(n)$ and choose g according to :

$$\text{Choose that } g \leqslant G(n) \text{ which minimizes } \mathrm{SSD}_{Wg}/h(h-1) \qquad \dots \quad (23)$$

or

$$\text{Choose that } g \leqslant G(n) \text{ which minimizes } a_g \cdot \sqrt{\mathrm{SSD}_{Wg}/h(h-1)}. \qquad \dots \quad (24)$$

The choice between (23) and (24), *for a given* $G(n)$, is not likely to have a substantial effect on the adjustment required for the critical values. Nor is it likely to have any substantial effect upon the way in which relative efficiency depends on length of tail. The choice between (23) and (24) will almost surely be based upon considerations pertinent to the user.

There are three such which seem likely to be of major importance:

(j1) It is slightly easier to use (23), since the square-rootings and multiplications by a_g required in the criterion of (24) are avoided.

(j2) The use of (24) may be more palatable, since one begins by seeking out the g for which the naive limits of (20) are closest together, and one retains this best-seeming value of g.

(j3) The use of (24) makes it very easy to make a conservative correction to the confidence limits set by a man who has used the naive limits of (20), perhaps without revealing how he chose g.

The strength of (j3) is of course greatest, when the change in critical values takes the form

$$a_g \rightarrow b_{n,G} \cdot a_g \qquad \ldots \quad (25)$$

so that the resulting confidence limits are

$$y_{Tg} \pm b_{n,G} \cdot a_g \cdot \sqrt{\mathrm{SSD}_{Wg}/h(h-1)} \qquad \ldots \quad (26)$$

where g is chosen by (24).

17. Purely a matter of selection ?

One attitude toward the problems we have just been discussing would be that they are purely matters of allowance for selection of procedure, another routine instance of a broad problem which routinely faces us in the analysis of data : There is a single set of data, and several alternative procedure by which it could be analyzed. While it is possible to retain validity by pledging that one will always use a single procedure, it is clear that failure to do something to adapt the procedure to the data leads to a loss of effectiveness. But if we naively select the procedure that appears to work best in each specific instance, we are exposing ourselves to a certain loss of validity.

There are almost always two ways out of such a bind, the one we propose to adopt, namely calculating adjusted critical values which allow for the choice, by a prescribed rule, of the apparently most appropriate procedure, and one which is rather in the spirit of Robbins' empirical Bayes techniques (Robbins, 1951, 1955; see also Neyman, 1962). In the second approach, one would regard an individual batch of data as one of a family of such batches, and plan to borrow information from the other batches to determine how we are to proceed with the first. (This general procedure is of course adopted daily, often at a very general level, by every working statistician whose familiarity with data of a certain class or classes help him to choose among procedures for its analysis.)

If we wish to appear "whiter than driven snow" insofar as selection is concerned, we may decide to choose the procedure to be used upon a given batch of data solely upon the evidence offered by the *other* batches of the family. If (i) the batches are independent, and (ii) their assignment to a "family" was without regard to their behaviour, even the most vehement and inquisitive seeker for selection bias could not ascribe any to the way in which the procedure was selected.

Why did we not propose such an approach to the selection of g? Mainly, we believe, because of a feeling about what will eventually be discovered to be the case. Specifically, suppose that the shape of the underlying distribution is fixed and that we are applying some procedure, or mixture of procedures, to samples drawn from distributions with this fixed shape. On the basis of available evidence and insights it is reasonable to suppose that one can do better, even after making due allowance for the $b_{n,G}$ factor that will then be required, by using different values of g for samples of different apparent long-tailedness, than by using any one fixed value of g.

For a specific shape of distribution, it is a factual question whether "mixing the g's" is advantageous or not. In time we should know the answer for certain specific distributions. In the meantime, we can do no better than follow our best judgement.

One could, if one wished, repeat an analogous argument about $G(n)$, which, in contrast to g, we have implicitly proposed to determine from a whole large family of batches of data. We have chosen to avoid making such an argument, and feel quite happy, on our present knowledge and insight, in making a sharp distinction between :

(k1) $G(n)$ to be picked on the evidence of a *family* of batches, and

(k2) g to be picked, subject to $g \leqslant G(n)$, on the evidence of *the single* batch in question.

To what extent this distinction is made because of a deeper understanding of the role of g, and to what extent it reflects a real distinction between what the user can gain from individualized choice of g in comparison with the individualized choice of $G(n)$ is hard to say.

18. Required next steps

What then are the next steps to be taken in making tailored t properly available for use? Some of them, surely, are these :

(11) Determine the values of a_g for an adequate net of values of g and n, and suitable error rates.

(12) Choose a reasonable function $G(n)$ and determine the values of b_{Gn} for a suitable pattern of values of g, n, and error rate. (A considerably sparser pattern may suffice for adequate interpolation.)

(13) Investigate the power of the resulting procedure, both for a Gaussian underlying distribution and for longer-tailed underlying distributions. Comparison with both Student's t and sign-configuration procedures will be in order.

(14) Make a start on extending the ideas and insights gained in the single-sample situation to more general situations.

Once reasonable progress has been made with (11) and (12), tailored t will be fully useable. We would expect to recommend it for routine use at that time.

So far as one can see, all the steps just mentioned demand Monte Carlo techniques for their solution, although we would not exclude the possibility of an analytic attack on some of the simpler ones. It should be emphasized that by Monte Carlo we do not mean naive experimental sampling, where one merely draws samples from the postulated underlying distribution and calculates a trimmed-t value from each, thus building up an empirical distribution approaching that of trimmed t like $n^{-\frac{1}{2}}$. Such naive procedures waste computational effort, and drastically reduce the accuracy of results that can be reached with plausible amounts of effort. Instead one must plan to use as much as possible of one's knowledge and insight in designing a modified sampling scheme whose results estimate a number known to be the same as the number estimated by the naive procedure. (See Kahn, 1956 for general discussion, and Arnold, Bucher, Trotter, and Tukey, 1956, or Wonnacott, 1963 for specific examples.)

While we plan to work on these problems at Princeton, we would welcome activity by others.

19. Algebra for the rectangular case

In closing we should set down the algebra which shows that

$$\text{rectangular divisor}_2(g+h+g) \doteq h(h-1). \qquad \dots \quad (27)$$

For a rectangular distribution with ends at 0 and 1 we have the well-known results for the low moments of the order statistics $y_1 \leqslant y_2 \leqslant y_3 \leqslant \dots \leqslant y_n$

$$\left.\begin{aligned} \text{ave } y_i &= \frac{i}{n+1} \text{ for all } i \\ \text{cov}(y_i, y_j) &= \frac{i(n+1-j)}{(n+1)^2(n+2)} \text{ for } i<j \end{aligned}\right\} \qquad \dots \quad (28)$$

whence

$$\text{var}(y_j-y_i) = \frac{(j-i)(n+1)-(j-i)^2}{(n+1)^2(n+2)} \qquad \dots \quad (29)$$

and

$$\text{ave}(y_j-y_i)^2 = \frac{(j-i+1)(j-i)}{(n+1)(n+2)} \qquad \dots \quad (30)$$

which depends upon $j-i$ and n alone as would be expected from the symmetric distribution of equivalent blocks.

In view of the general identity

$$2\cdot\Sigma 1 \cdot \Sigma(u_i-u.)^2 = \Sigma\Sigma(u_j-u_i)^2 \qquad \dots \quad (31)$$

we have, using the Winsorized observations as the u_i

$$\begin{aligned} 2n\cdot \text{SSD}_{Wg} = 0+2g \sum_{g+1}^{n-g}(y_j-y_{g+1})^2 + \sum_{g+1}^{n+g}\ \sum_{g+1}^{n-g}(y_j-y_i)^2 \\ +2g\Sigma(y_{n-g}-y_i)^2+2g^2(y_{n-g}-y_{g+1})^2+0, \end{aligned} \qquad \dots \quad (32)$$

whence, setting $i = g+r$, $j = g+s$, and using (30)

$$2n(n+1)(n+2)\cdot \text{ave SSD}_{Wg} = 4g\sum_1^h (r-1)(r) + \sum\sum_1^h (r-s)(r-s+1)+2g^2\cdot(h-1)h \ \dots \quad (33)$$

Now $\Sigma\, r(r-1) = (h+1)(h)(h-1)/3$. And the double sum can be easily evaluated by considering a sample of size h, from the same rectangular distribution, and the corresponding untrimmed SSD, say SSD(z), for which

$$2h \cdot \text{SSD}(z) = \Sigma\Sigma(z_r - z_j)^2 \qquad \ldots \quad (34)$$

$$\text{ave SSD}(z) = (h-1)\sigma^2 = \frac{h-1}{12} \qquad \ldots \quad (35)$$

Now write (33) for $n = h$ and $g = 0$, finding

$$2 \cdot h(h+1)(h+2) \text{ ave SSD}(z) = \sum_1^h \Sigma(r-s)(r-s+1) \qquad \ldots \quad (36)$$

but the left-hand side equals

$$2h(h+1)(h+2)\frac{(h-1)}{12} = \frac{1}{6}(h-1)\, h(h+1)(h+2) \qquad \ldots \quad (37)$$

so that

$$\frac{1}{6}(h-1)\, h(h+1)(h+2) = 2(h)(h+1)(h+2) \text{ ave SSD}(z) = \Sigma\Sigma(r-s)(r-s+1). \quad \ldots \quad (38)$$

$$2n(n+1)(n+2) \text{ ave SSD}_{W_g} = 4g\,\frac{(h-1)\, h(h+1)}{3} + \frac{1}{6}(h-1)\, h(h+1)(h+2) + 2g^2(h-1)h \qquad \ldots \quad (39)$$

and
$$\text{ave SSD}_{W_g} = \frac{(h-1)h}{12n(n+1)(n+2)}[(h+1)(h+2)+8g(h+1)+12g^2]$$

$$= \frac{(h-1)h}{12n(n+1)(n+2)}[3n^2-(h-2)(2n+1)] \qquad \ldots \quad (40)$$

We turn now to
$$h^2 \text{ var } y_{T_g} = \sum_{(T_g)}\Sigma \text{ cov } (y_i, y_j) \qquad \ldots \quad (41)$$

and notice that, if $r < s$ $\quad (n+1)^2(n+2) \text{ cov } (y_{g+r}, y_{g+s})$

$$\begin{aligned} &= (g+r)(n+1-g-s) = (g+r)(g+h+1-s) \\ &= g^2+g(r-s)+g(h+1)+(h+1)^2(h+2) \text{ cov } (z_r, z_s) \\ &= g^2-g \cdot |r-s|+g(h+1)+(h+1)^2(h+2) \text{ cov } (z_r, z_s) \qquad \ldots \quad (42) \end{aligned}$$

where the requirement that $r < s$ can be lifted for the last form, so that, using (41)

$$h^2(n+1)^2(n+2) \text{ var } y_{T_g} = h^2 \cdot g^2 - g\sum_1^r \Sigma\, |r-s| + h^2(h+1)g + (h+1)^2(h+2)(h^2 \text{ var } z_\bullet) \quad (43)$$

and, since $\quad h^2 \cdot \text{var } z_\bullet = h/12$ and $\Sigma\Sigma\,|r-s| = h(h^2-1)/3$

we have
$$\text{var } y_{T_g} = \frac{(h+1)^2(h+2)+12h(h+1)g-4(h^2-1)g+12hg^2}{12h(n+1)^2(n+2)}$$

$$= \frac{h(3n-2h+3)+2}{12h(n+1)(n+2)}, \qquad \ldots \quad (44)$$

Now

$$\frac{\text{ave SSD}_{W_g}}{h(h-1)\text{ var } y_{T_g}} = \frac{h(3n^2-2nh+4n-h+2)}{n(h(3n-2h+3)+2)}$$

$$= 1 + \frac{(n-h)(h-2)}{n(3(n-h)h+(h+1)(h+2))} \quad \dots \quad (45)$$

which will usually be close to 1.

References

Arnold, H. J., Bucher, B. D., Trotter, H. F. and Tukey, J. W. (1956): Monte Carlo techniques in a complex problem about normal samples. *Symposium on Monte Carlo Methods.* (Edited by H. A. Meyer), John Wiley and Sons, Paper 9, 80–88.

Bradley, R. A. (1952): The distribution of the t and F statistics for a class of non-normal populations. *Virginia J. Science*, **3**, 1–32.

Dixon, W. J. (1960): Simplified estimation from censored normal samples. *Ann. Math. Stat.*, **31**, 385–391.

Dixon, W. J. and Tukey, J. W. (1963): Approximate behaviour of the distribution of Winsorized t (Trimming/Winsorization 2). In manuscript.

Gayen, A. K. (1949): The distribution of 'Student's' t in random samples of any size drawn from non-normal universes. *Biometrika*, **36**, 353–369.

Hastings, C. Jr., Mosteller, F., Tukey, J. W. and Winsor, C. P. (1947): Low moments for small samples: a comparative study of order statistics. *Ann. Math. Stat.*, **18**, 413–426.

Johnson, N. L. (1949): Systems of frequency curves generated by methods of translation. *Biometrika*, **36**, 149–176.

Kahn, H. (1956): Use of different Monte Carlo sampling procedures. *Symposium on Monte Carlo Methods.* (Edited by H. A. Meyer), John Wiley and Sons, Paper 15, 146–190.

Klotz, J. (1962): Small sample power and efficiency for the one sample Wilcoxon and normal scores tests. Paper read at meeting of the Institute of Mathematical Statistics, Minneapolis, 7 Sept. 1962, 17 pp.

Laderman, J. (1939): The distribution of "Student's" ratio for samples of two items drawn from non-normal universes. *Ann. Math. Stat.*, **10**, 376–379.

Lord, E. (1947): The use of range in place of standard deviation in the t-test. *Biometrika*, **34**, 41-67. (Correction, **39**, 442).

——— (1950): Power of the modified t-test (u-test) based upon range. *Biometrika*, **37**, 64–77.

Mandelbrot, B. (1960): The Pareto-Levy law and the distribution of income. *Int. Econ. Rev.*, **1**, 79–106.

——— (1961): Stable Paretian random functions and the multiplicative variation of income. *Econometrica*, **29**, 517–543.

——— (1962): The variation of certain speculative prices. *IBM Research Note NC*-**87**, 111.

McLaughlin, D. H. and Tukey, J. W. (1961): The variance of symmetrically trimmed samples from normal populations, and its estimation from such trimmed samples. (Trimming/Winsorization 1). *Stat. Tech. Res. Group*, (Princeton University) *Tech. Rept.*, **42**, 24.

Moses, L. (1953): Non-parametric methods. H. M. Walker and J. Lev's *Statistical Inference*, Henry Holt, Chapter 18, 426–450.

Neyman, J. (1962): Two breakthroughs in the theory of statistical decision making. *Rev. Internat. Statist. Inst.*, **30**, 11–27.

Pearson, E. S. assisted by Adyanthaya, N. K. (1929): The distribution of frequency constants in small samples from non-normal symmetrical and skew populations. *Biometrika*, **21**, 251–289.

PERLO, V. (1953) : On the distribution of Student's ratio for samples of three drawn from a rectangular distribution. *Biometrika*, **25**, 203–204.

RIDER, P. R. (1929) : On the distribution of the ratio of mean to standard deviation in small samples from non-normal universes. *Biometrika*, **21**, 124–143.

RIETZ, H. L. (1939) : On the distribution of the "Student" ratio for small samples from certain non-normal populations. *Ann. Math. Stat.*, **10**, 265–274.

ROBBINS, H. (1951) : Asymptotically subminimax solutions of compound statistical decision problems. *Proc. 2nd Berkeley Symp. Math. Stat. Prob.*, 131–148.

——— (1955) : An empirical Bayes' approach to statistics. *Proc. 3rd Berkeley Symp. Math. Stat. and Prob.*, **1**, 157–164.

SARHAN, A. E. (1954) : Estimation of the mean and standard deviation by order statistics. *Ann. Math. Stat.*, **25**, 317–328.

SARHAN, A. E. and GREENBERG, B. G. (1956) : Estimation of location and scale parameters by order statistics from singly and doubly censored samples. Part I, the normal distribution up to samples of size 10. *Ann. Math. Stat.*, **27**, 427–451.

SIEGEL, S. (1956) : *Non-Parametric Statistics for the Behavioral Sciences*, McGraw-Hill, 75–83.

STUDENT (1908). On the probable error of a mean. *Biometrika*, **6**, 1–25. Reprinted in "*Student's*" *Collected Papers*, Biometrika Office, Paper 2, 11–34.

TEICHROEW, D. (1956) : Tables of expected values of order statistics for samples of size 20 and less from the normal distribution. *Ann. Math. Stat.*, **27**, 410–426.

——— (1962) : Tables of lower moments of order statistics for samples from the normal distribution. *Contributions to Order Statistics* (Edited by A. E. Sarhan and B. G. Greenberg), Chapter 10B, 190–205, John Wiley.

TUKEY, J. W. (1960) : A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics: Essays in honour of Harold Hotelling* (Edited by I. Olkin *et al.*), 448–485.

——— (1962) : The future of data analysis. *Ann. Math. Stat.*, **33**, 1–67 and 812.

WALSH, J. E. (1947) : Some significance tests for the median which are valid under very general conditions. Ph.D. Thesis, Princeton University.

——— (1949) : Some significance tests for the median which are valid under very general conditions. *Ann. Math. Stat.*, **20**, 64–81.

——— (1959) : Comments on "The simplest signed-rank tests". *J. Amer. Stat. Ass.*, **54**, 213–224.

WILCOXON, F. (1945) : Individual comparisons of grouped data by ranking methods. *Biometrics Bulletin* (later *Biometrics*) **1**, 80–83.

——— (1946) : Individual comparisons of grouped data by ranking methods. *J. Econ. Entomology*, **39**, 269–270.

——— (1947) : Probability tables for individual comparisons by ranking methods. *Biometrics*, **3**, 119–122.

——— (1949 and later editions) : *Some Rapid Approximate Statistical Methods*, American Cyanamid Company, Stamford Research Laboratories.

WONNACOTT, T. H. (1963) : A Monte-Carlo method of obtaining the power of certain tests of location when sampling from non-normal distributions. Ph.D. Thesis, Princeton University.

Paper received : February, 1963.