ANALYSIS OF EXTREME VALUES

By W. J. Dixon[1]

University of Oregon

1. Introduction. It is well recognized by those who collect or analyze data that values occur in a sample of n observations which are so far removed from the remaining values that the analyst is not willing to believe that these values have come from the same population. Many times values occur which are "dubious" in the eyes of the analyst and he feels that he should make a decision as to whether to accept or reject these values as part of his sample. On the other hand he may not be looking for an error, but may wish to recognize a situation when an occasional observation occurs which is from a different population. He may wish to discover whether a significant analysis of variance indicates an extreme value significantly different from the remainder. Also, of course, the extreme value may differ significantly without causing a significant analysis of variance and he may wish to discover this. It is reasonable to suppose that a criterion for rejecting observations would be useful here also. The choice of a suitable criterion for rejecting observations introduces a number of questions.

1. Should any observations be removed if we wish a representative sample including whatever contamination arises naturally? In other words, it may be desirable to describe the population including *all* observations, for only in that way do we describe what is actually happening.

2. If the analyst wishes to sample the population unaffected by contamination he must either remove the contaminating items or employ statistical procedures which reduce to a minimum the effect of the contamination on the estimates of the population. That is, he may wish to describe only 95% of his population if the description is altered radically by the remaining 5% of the observations. He may have external reasons which are good and sufficient for wishing to describe only 95% of his observations. Suppose he wishes to use the sample for a statistical inference; the inclusion of all the data may sufficiently violate the assumptions underlying the inference to exclude the possibility of making a valid inference.

This paper will concern itself only with those problems which arise from Question 2.

If we wish to follow some procedure which attempts to remove contamination we must consider the performance of any proposed criterion with respect to the proportion of contamination the criterion will discover and, of course, the proportion of the "good" observations which are removed by the use of the criterion. But, perhaps more important, we must consider what sort of bias will result when the standard statistical procedures are applied to samples of observations which have been processed in this manner.

[1] This paper was prepared under a contract with the Office of Naval Research.

If we wish to follow a procedure which will not search for particular values to be excluded but will minimize their effect if present, we must investigate the sampling distributions of these modified statistics and estimate the loss in information resulting from their use when all observations are "good." We must also investigate the expected bias which will result when "bad" items are present even though essentially excluded. Perhaps most disturbing about the avoidance of "bad" items is the fact that a decision must still be made as to whether a "bad" item was present or not in order to know in which way our estimates may be biased. For example, a sample mean computed by avoiding the two end observations will not be a biased estimate of the mean of a symmetric population if both end items should actually be included or if both end items should not be included. However, if only one of the two should not be included this estimate of the mean will be biased.

2. Models of contamination. The performance of the various criteria for discovery of one or more contaminators will be measured with reference to contaminations of the following two types entering into samples of observations from a normal population with mean μ and variance σ^2, $N(\mu, \sigma^2)$

A. One or more observations from $N(\mu + \lambda\sigma, \sigma^2)$,

B. One or more observations from $N(\mu, \lambda^2\sigma^2)$.

A represents the occurrence of an "error" in mean value such as will occur in dial readings when errors are made in reading incorrectly digits other than the last one or two digits. Errors of this sort may result from momentary shifts in line voltage or from the inclusion among a group of objects of one or two items of completely different origin. This type of contamination will be referred to as "location error." B represents the occurrence of an "error" from a population with the same mean but with a greater variance than the remainder of the sample. This type of error will be referred to as a "scalar error." It is likely that many errors could be better described as a combination of A and B, but a study of these two errors separately should throw considerable light on the question of "gross errors" or "blunders."

Many authors have written on the subject of the rejection of outlying observations. Apparently none have been successful in obtaining a general solution to the problem. Nor has there been success in the development of a criterion for discovery of outliers by means of a general statistical theory; e.g., maximum likelihood. A large number of criteria have been advanced on more or less intuitive grounds as appropriate criteria for this purpose. In no case was investigation made of the performance of these criteria except for a few illustrative examples.

References for the criteria discussed in the next section are given at the end of this paper. Indications are given as to the significance values available in those papers.

3. Criteria to be considered. The performance of two types of criteria has been investigated for samples contaminated with location or scalar errors.

a) σ known or estimated independently,

b) σ unknown.

The n observations are ordered $x_1 < x_2 < \cdots < x_n$. The criteria involving external knowledge of σ are:

A. χ^2 test,

$$\chi^2 = \frac{\Sigma(x - \bar{x})^2}{\sigma^2}.$$

B. Extreme deviation,

$$B_1 = \frac{x_n - \bar{x}}{\sigma} \left(\text{or } \frac{\bar{x} - x_1}{\sigma}\right),$$

$$B_2 = \frac{x_n - x_{n-1}}{\sigma} \left(\text{or } \frac{x_2 - x_1}{\sigma}\right).$$

C. Range,

$$C_1 = \frac{w}{\sigma}, \qquad w = x_n - x_1,$$

$$C_2 = \frac{w}{s}, \qquad s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} \qquad (s \text{ independently estimated}).$$

The criteria involving only the information of a single sample of n observations are:

D. Modified F test.

1. For single outlier x_1,

$$D_1 = \frac{S_1^2}{S^2}, \quad \text{where} \quad S_1^2 = \sum_2^n (x - \bar{x}_1)^2, \quad \bar{x}_1 = \sum_2^n x/(n - 1),$$

$$S^2 = \sum_1^n (x - \bar{x})^2, \quad \bar{x} = \sum_1^n x/n$$

$$\left(\text{or for } x_n,\ D_1 = \frac{S_n^2}{S^2}\right).$$

2. For double outliers x_1, x_2,

$$D_2 = \frac{S_{1,2}^2}{S^2}, \quad \text{where} \quad S_{1,2}^2 = \sum_3^n (x - \bar{x}_{1,2})^2, \quad \bar{x}_{1,2} = \sum_3^n x/(n-2)$$

$$\left(\text{or for } x_n,\ x_{n-1},\ D_2 = \frac{S_{n,n-1}^2}{S^2}\right).$$

E. Ratios of ranges and subranges.

1. For single outlier x_1,

$$r_{10} = \frac{x_2 - x_1}{x_n - x_1}$$

$$\left(\text{or for } x_n,\ r_{10} = \frac{x_n - x_{n-1}}{x_n - x_1}\right).$$

2. For single outlier x_1 avoiding x_n,

$$r_{11} = \frac{x_2 - x_1}{x_{n-1} - x_1}$$

$$\left(\text{or for } x_n \text{ avoiding } x_1,\ r_{11} = \frac{x_n - x_{n-1}}{x_n - x_2}\right).$$

3. For single outlier x_1, avoiding x_n, x_{n-1},

$$r_{12} = \frac{x_2 - x_1}{x_{n-2} - x_1}$$

$$\left(\text{or for } x_n \text{ avoiding } x_1,\ x_2,\ r_{12} = \frac{x_n - x_{n-1}}{x_n - x_3}\right).$$

4. For outlier x_1 avoiding x_2,

$$r_{20} = \frac{x_3 - x_1}{x_n - x_1}$$

$$\left(\text{or for } x_n \text{ avoiding } x_{n-1},\ r_{20} = \frac{x_n - x_{n-2}}{x_n - x_1}\right).$$

5. For outlier x_1 avoiding x_2 and x_n,

$$r_{21} = \frac{x_3 - x_1}{x_{n-1} - x^1}$$

$$\left(\text{or for } x_n \text{ avoiding } x_{n-1},\ x_1,\ r_{21} = \frac{x_n - x_{n-2}}{x_n - x_2}\right).$$

6. For outlier x_1 avoiding x_2 and x_n, x_{n-1},

$$r_{22} = \frac{x_3 - x_1}{x_{n-2} - x_1}$$

$$\left(\text{or for } x_n \text{ avoiding } x_{n-1},\ x_1,\ x_2,\ r_{22'} = \frac{x_n - x_{n-2}}{x_n - x_3}\right).$$

F. Extreme deviation and standard deviation.
For single outlier x_n,

$$F = \frac{x_n - \bar{x}}{s} \quad \left(\text{or for } x_1,\ F = \frac{\bar{x} - x_1}{s}\right).$$

The performance of the large number of criteria listed here will be assessed with respect to discovery of contamination of the type given in Section 2.

4. Performance of criteria (estimate of σ available). The χ^2 test will of course give an indication of a large dispersion and since the extreme values are chief contributors to the sum of squares, it is possible to use this test as a criterion for rejecting a value or values which are at the greatest distance from the mean. It might be supposed the B_1 and B_2 would give better results since particular attention is paid to the end item. The same argument would influence one in favor of C_1 or C_2. The performance of C_2 can, of course, be expected to vary with the degrees of freedom in the independent estimate of σ. For this study the degrees of freedom for this estimate were held to the single value 9 d.f.

χ^2 may be used since if the value of χ^2 is too large (greater than some upper percentage point for χ^2) we might reject the value most distant from the mean. χ^2 tables may be used for percentage points. Percentage points for the other statistics considered here are given in the references at the end of this paper.

The criteria A, B_1, B_2, C_1, C_2 were investigated for $\alpha = 1\%$, 5% and 10% for $\lambda = 2, 3, 5, 7$, where one or more items are selected from a population $N(\mu + \lambda\sigma, \sigma^2)$ and the remainder from $N(\mu, \sigma^2)$. Investigations were also made for one item from $N(\mu, \lambda^2\sigma^2)$ for $\lambda = 2, 4, 8, 12$. The investigation was carried out by sampling methods. The performances of different criteria were assessed for the same group of samples in order to obtain more precision in the comparison of the different tests. All of the points appearing on the graphs in the subsequent sections of this paper were based on from 66 to 200 determinations.

The performance of the above criteria is measured by computing the proportion of the time the contaminating distribution provides an extreme value and the test discovers this value. Of course, performance could be measured by the proportion of the time the test gives a significant value when a member of the contaminating population is present in the sample, even though not at an extreme. However, since it is assumed that discovery of an outlier will frequently be followed by the rejection of an extreme we shall consider discovery a success only when the extreme value is from the contaminating distribution.

The performance was judged by applying the criteria to each sample, always suspecting an outlier in the direction of the shifted mean for location error. Since the location errors were inserted by adding a fixed value to one or more of the observations, the largest value was tested as an outlier. The measure of performance was the percentage of location errors identified. When the location error was not an outlier, no test was performed and a failure for the test recorded.

In the case of the model of contamination involving the scalar error, the value was suspected which was farthest from the mean. This of course, alters somewhat the level of significance, but this procedure was followed alike for all criteria investigated. The performance was measured in the same fashion as for location errors.

Considering first, location errors, a study of the performance curves showing the per cent discovery of contaminators plotted against λ (the number of standard deviation units the population of contaminators is removed from the remainder), shows that the level of performance for σ known is considerably above the level of performance when σ is not known. The difference is greater for $n = 5$ than for $n = 15$ and, of course, the difference will diminish as the sample size increases. Figure 1 shows the performance curves for $\alpha = 5\%$ (5% significance level for the test for an outlier) of $B_1 = (x_n - \bar{x})/\sigma$ for $n = 5$ and $n = 15$ and of $r_{10} = \dfrac{x_n - x_{n-1}}{x_n - x_1}$ for $n = 5$ and $n = 15$.

The graphs for $\alpha = 1\%$ and 10% would be similar in appearance. Figure 2 indicates the change in performance for $\alpha = 1\%$, 5%, and 10%. The curves plotted are for $B_1 = (x_n - \bar{x})/\sigma$. The curves for A, B_2, C_1, C_2 show very similar results.

The curve for test B_1 was used in Figures 1 and 2 since it gives the best performance of all criteria which are considered here if a single location error is present. The curves showing the comparative performance of these criteria as well as one to be considered later (r_{10}) are given in Figure 3 for $\alpha = 5\%$ and for $n = 5$ and $n = 15$.

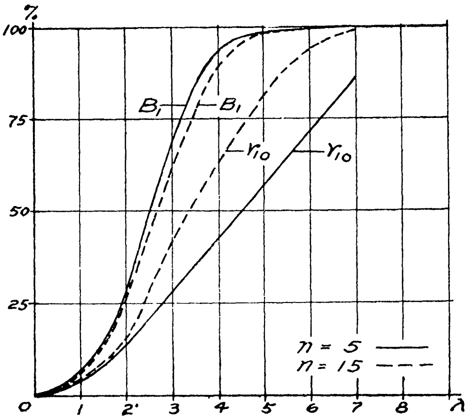


FIG. 1. Improvement in performance obtained with knowledge of σ, $\alpha = 5\%$, $n = 5$, 15.

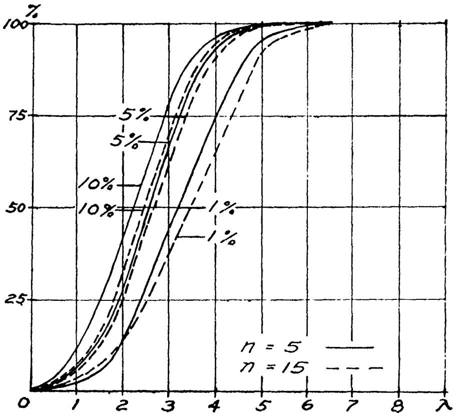


FIG. 2. The effect of the level of significance on the performance of B_1 ; $\alpha = 1\%$, 5%, 10%; $n = 5$, 15.

The following statements can be made from inspection of Figure 3:

a) The differences among A, B_1, B_2, and C_1 are not great.

b) The knowledge of σ is less important in larger samples.

c) The curve for C_2 lies above that of r_{10} for $n = 5$ and below that of r_{10} for $n = 15$. This is consistent with the use of 9 d.f. in the independent estimate of σ.

If the question of ease in computation or application is important, it may be desirable to use B_2 or C_1 in place of B_1 for they are slightly easier to compute and it is not necessary to measure all observations to obtain the value of these statistics. From Figure 3 it will be noted that the performances of these criteria are nearly as good as for B_1. If two outliers may be expected in a single sample,

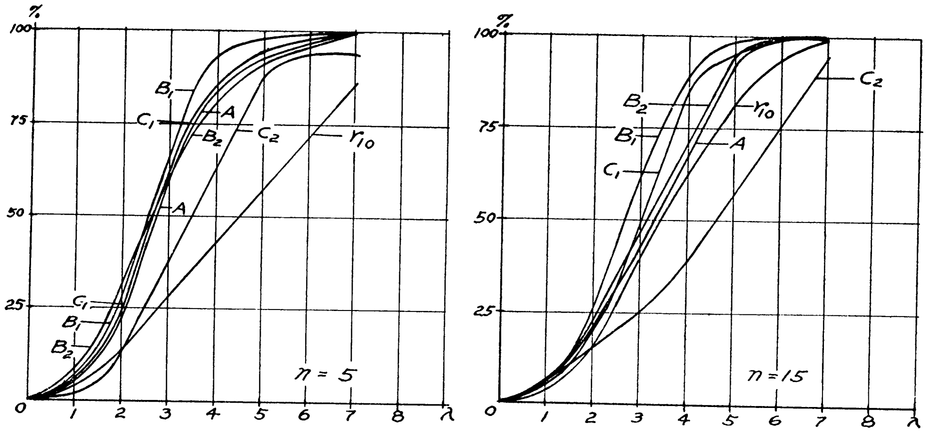


FIG. 3. Comparison of the performance of criteria using σ known (or using external estimates of σ) and r_{10} for samples of size 5 and 15, $\alpha = 5\%$.

the performance of B_2 will be lowered and the performance of B_1 and C_1 will be improved. Any differences between the performance of B_1 and the performance of C_1 when two outliers are present was not discernable for $n = 5$ or 15. Figure 4 illustrates the improvement in performance for B_1 for $\alpha = 5\%$ and $n = 15$.

The performance curves of these criteria if a scalar error is present are very similar to those above except that:

1. A high level of performance is approached very slowly. For example, see Figure 5 showing the performance of B_1 and r_{10} for $n = 5$ and $n = 15$ and $\alpha = 5\%$.
2. There is a smaller difference in the performance between the criteria with σ known and σ unknown (see Figure 5).

The performance of B_1 and C_1 are noticeably increased by the introduction of more contaminators while that of B_2 decreases. No difference in the perform-

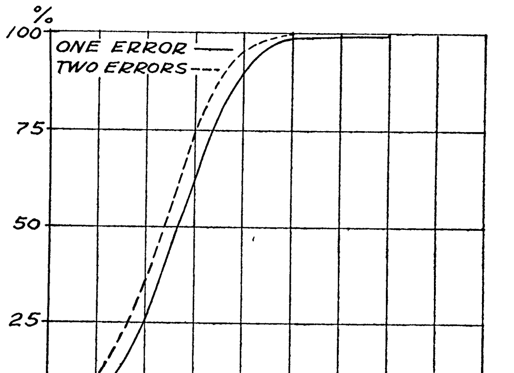


FIG. 4. Comparison of the performance of B_1 for one and two location errors in samples of size 15, $\alpha = 5\%$.

ance of B_1 and C_1 were noted for either $n = 5$ or $n = 15$. Figure 6 shows the increase in performance of two contaminators for B_1 for $n = 15$, $\alpha = 5\%$.

The general recommendations for possibilities of either type of contamination, location or scalar errors, would lead one to the use of B_1 or C_1 if σ is known.

Criterion C_1 is recommended since:

1. Its performance is almost as good as the performance of B_1 for a single outlier. Their performances are about equal for two outliers and C_1 affords protection for outliers either above or below the mean.
2. It is simple to compute.

If ease of computation is not essential and maximum performance is desired, the criterion B_1 should be used. The performance of C_2 will approach that of B_1 as the number of degrees of freedom in the denominator increases.

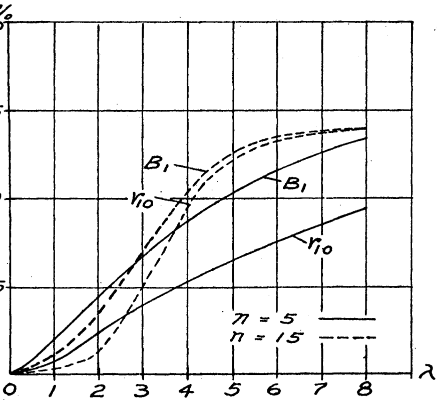


FIG. 5. Comparison of the performance of B_1 and r_{10} for one scalar error for samples of size 5 and 15, $\alpha = 5\%$.

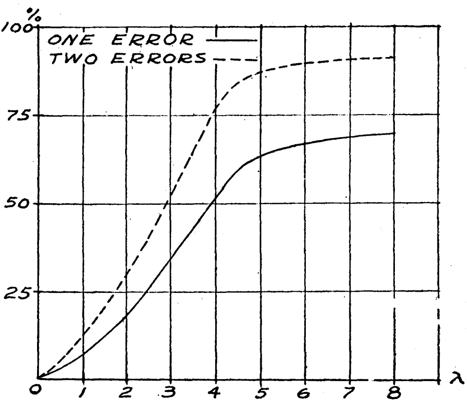


FIG. 6. Comparison of the perfo mance B_1 for one and two scalar errors in samples of size 15, $\alpha = 5\%$.

5. Performance of criteria (no external estimate of σ). Criteria D_1 and D_2 have strong intuitive reasons for their use since the dispersion is estimated by s^2. The r ratios are attractive because of their simplicity and their preoccupation with the extreme values. Test F is the "studentized" ratio corresponding to B_1, and is equivalent to D_1 since $D_1 = 1 - F^2/(n - 1)$. There is no apparent difference in the performance of D_1 and r_{10} when one outlier is present and no apparent difference in D_2 and r_{20} when two outliers are present. This is true for both models of contamination and for the three levels of significance investigated. However the comparison of D_2 and r_{20} was made only for $n = 5$ since critical values are not available[2] for D_2 for $n = 15$. (Critical values are available for $n \leq 12$.)

The performance of D_1 and r_{10} under the two models of contamination can be obtained by reference to the curve for r_{10} in Figure 1 and Figure 5. The curve for D_1 is practically identical with the curve for r_{10}.

[2] After this paper was submitted, the critical values of D_2 have been extended to $n \leq 20$ (see references).

There is no question that r_{10} is simpler to use, so that if this condition of contamination (scalar errors) exists, r_{10} would probably be chosen. However, as before, we should investigate what happens when more than one error is present. D_2 is designed for this case as is r_{20}. Since the performance of these two criteria is approximately the same, r_{20} would probably be chosen because of its simplicity. Critical values for this statistic are available for $n \leq 30$.

r_{11}, r_{12}, r_{20}, r_{21}, r_{22} were designed for use in situations where additional outliers may occur and we wish to minimize the effect of these outliers on the investigation of the particular value being tested.

It has been suggested that D_1 could be used repeatedly to remove more than one outlier from a sample. This procedure cannot be recommended since the presence of additional outliers handicaps the performance of both D_1 and r_{10} for small sample sizes and therefore the process of rejection might never get started. For larger sample sizes the performance of D_1 is affected much less by the presence of two errors than is the performance of r_{10}. The repetitive use of D_1 is not recommended in this case either since r_{20} performs in a superior manner to D_1 in such situations. This difference in performance of D_1 and r_{10} depends markedly on the level of significance used as well as the sample size. For small samples there is little difference in performance for any of the levels of significance one might use. For the larger sample sizes there is no appreciable difference for very high levels of significance. The difference is however very great for lower levels of significance. In fact as λ increases for two errors of the location type, the level of significance which divides the region of approach to zero performance from the region of approach to perfect performance of D_1 is given by the level of significance corresponding to a significance value of $\frac{1}{2}\left(\frac{n}{n-1}\right)$ for D_1. Thus, for example, in samples of size 15, $\frac{1}{2}\left(\frac{n}{n-1}\right) = \frac{15}{28} = .536$. This value lies between the values for the 2.5% and 5% level of significance. These values are .503 and .556 respectively. Therefore the use of the 1% or 2.5% levels will give poorer and poorer performance as λ increases, and the use of the 5% or 10% levels will give better and better performance as λ increases when two errors are present. The dividing point is such that for samples of size 11 or less the use of any of the given levels of significance will cause the performance to decrease as λ increases. For samples of size $n \leq 14$ the 1%, 2.5% and 5% levels have the same effect, and for samples of size $n \leq 16$ the 1% and 2.5%, for samples of size $n \leq 19$ just the 1% level. For three such errors the limit approached by D_1 as λ increases is $\frac{2}{3}\left(\frac{n}{n-1}\right)$. Therefore, the performance of D_1 will approach zero for all levels of significance and for all sample sizes for which critical values are known except the 10% level of significance for sample sizes larger than 21. An indication of these limiting values $\frac{k-1}{k}\,\frac{n}{n-1}$ for k contaminations present can be obtained by considering these k values to

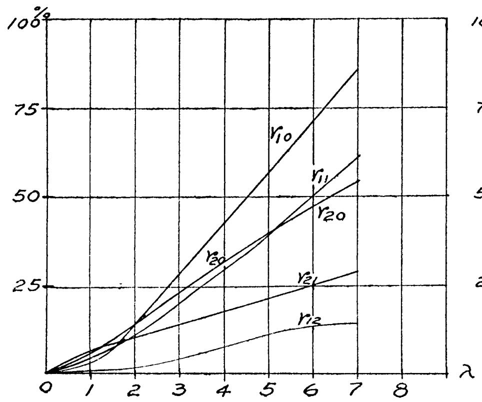


FIG. 7. Comparison of the performance of the r criteria for one location error in samples of size 5, $\alpha = 5\%$.

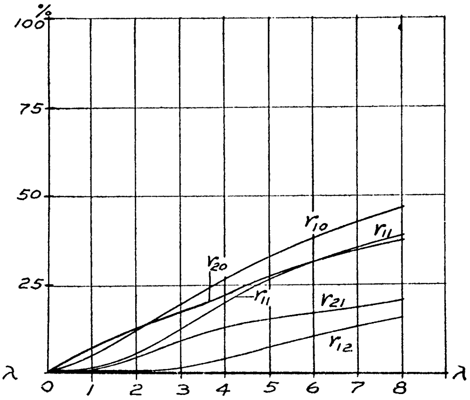


FIG. 8. Comparison of the performance of the r criteria for one scalar error in samples of size 5, $\alpha = 5\%$.

be at a distance k from the population mean, computing D_1 and allowing λ to increase indefinitely.

The comparative performance of the r criteria, $\alpha = 5\%$, in samples of size 5 for the two models of contamination (one contaminator present) are given in Figures 7 and 8. For samples of size 15 the curves are given in Figures 9 and 10. A single curve suffices here since there is no discernable difference in the curves for the different r criteria. There is considerable difference in the performance curves if more than one outlier is present. However, the performances of r_{10}, r_{11}, r_{12} are essentially the same when two location outliers are present as are the performances of r_{20}, r_{21}, r_{22}. Figures 11 and 12 show the comparative performance of r_{10}, r_{11}, r_{12} for one and two contaminators for $\alpha = 5\%$ and $n = 5$. Figures 13 and 14 are for $n = 15$. Figures 15 and 16 show the comparative per-

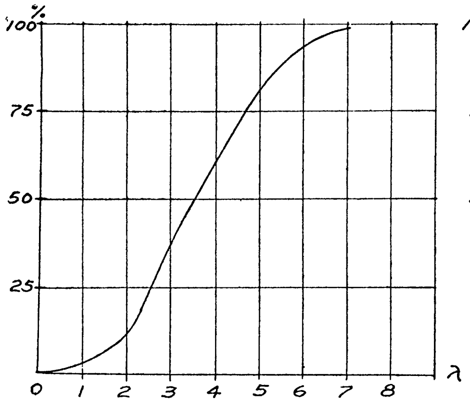


FIG. 9. Performance of the r criteria for one location error in samples of size 15, $\alpha = 5\%$.

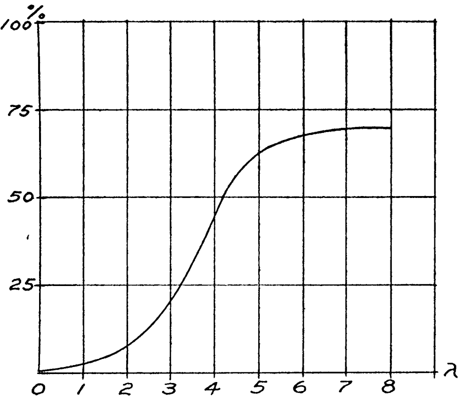


FIG. 10. Performance of the r criteria for one scalar error in samples of size 15, $\alpha = 5\%$.

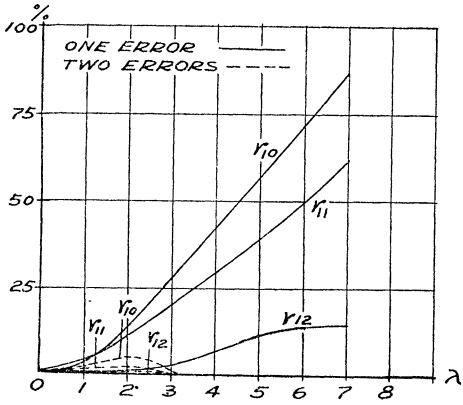


FIG. 11. Comparison of the performance of the $r_{1\cdot}$ criteria for one and two location errors in samples of size 5, $\alpha = 5\%$.

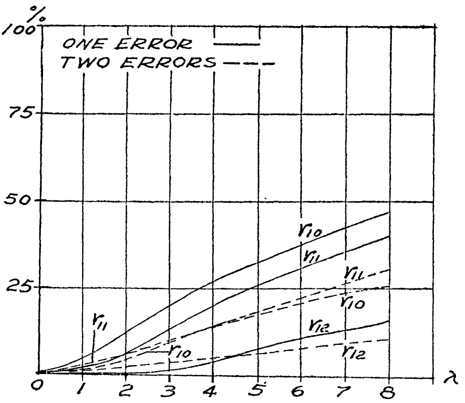


FIG. 12. Comparison of the performance of the $r_{1\cdot}$ criteria for one and two scalar errors in samples of size 5, $\alpha = 5\%$.

formance for r_{20}, r_{21}, (r_{22} is not a test for $n = 5$) for one and two contaminators for $\alpha = 5\%$ and $n = 5$. Figures 17 and 18 are for r_{20}, r_{21}, r_{22} for $n = 15$. The six curves represented by the single curve of Figure 17 lie within 5% of the curve shown. The same is true of the three curves represented by each of the two curves of Figure 18.

Since no loss in performance results for larger samples from the use of r_{20}, r_{21}, r_{22} in place of r_{10}, r_{11}, r_{12}, and further, these criteria are not appreciably affected by the presence of another outlier it would seem unwise to recommend the use of r_{10}, r_{11}, r_{12}. However, note that for small samples (see Figures 11 and 12) the performances of r_{10} and r_{11} and r_{12} are considerably better when a single

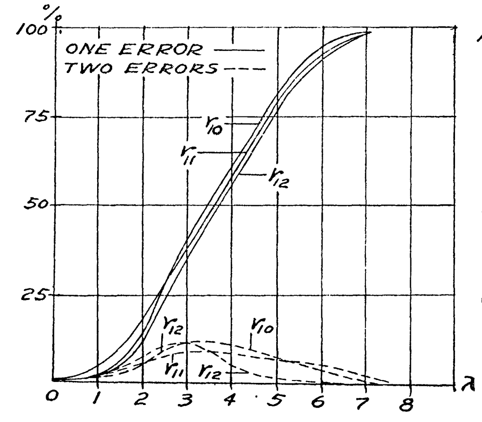


FIG. 13. Comparison of the performance of the $r_{1\cdot}$ criteria for one and two location errors in samples of size 15, $\alpha = 5\%$.

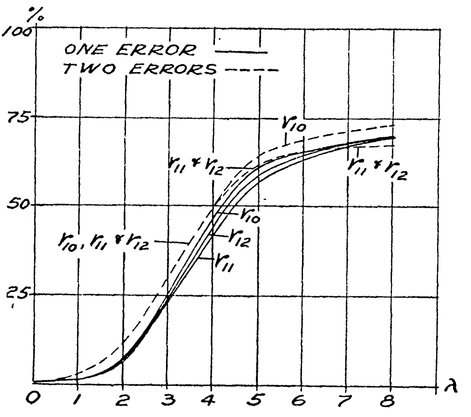


FIG. 14. Comparison of the performance of the $r_{1\cdot}$ criteria for one and two scalar errors in samples of size 15, $\alpha = 5\%$.

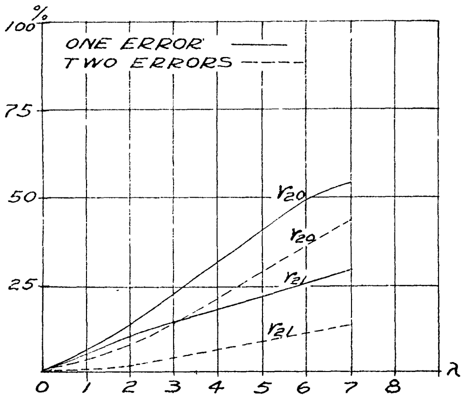


FIG. 15. Comparison of the performance of the $r_{2\cdot}$ criteria for one and two location errors in samples of size 5, $\alpha = 5\%$.

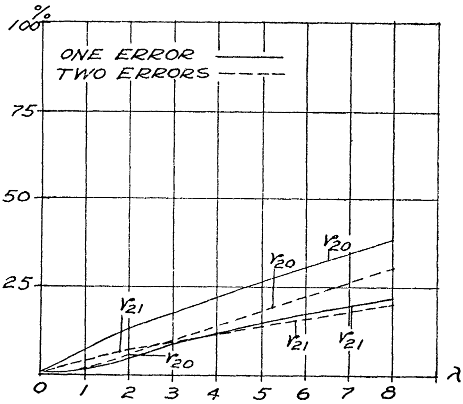


FIG. 16. Comparison of the performance of the $r_{2\cdot}$ criteria for one and two scalar errors in samples of size 5, $\alpha = 5\%$.

outlier is present. Therefore in larger ($n > 10$) samples r_{20} or r_{21} would appear to be the best criteria. In samples of size 10 or less, r_{10} or r_{20} should be used; r_{21} if the extreme value at the opposite end should be avoided.

It should be noted in the comparisons that no model of contamination was investigated which would cause one or more errors at both extremes in the sample. It is obvious that the performance of D_1 and D_2 would be considerably decreased while the performance of r_{11}, r_{12}, and r_{21}, r_{22} would not be materially affected since these criteria avoid values at the opposite extreme. Their repeated use might discover most of such outliers, while D_1 or D_2 might fail on the first trial.

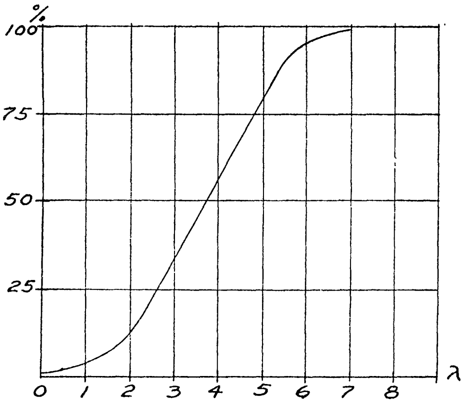


FIG. 17. Comparison of the performance of the $r_{2\cdot}$ criteria for one and two location errors in samples of size 15, $\alpha = 5\%$.

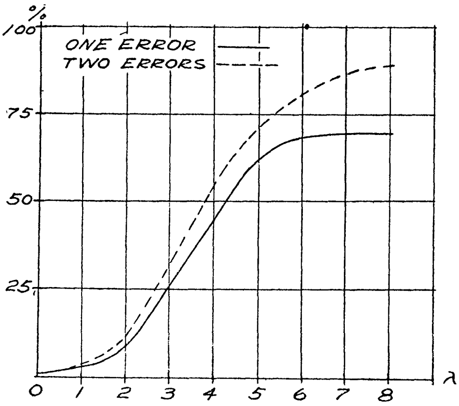


FIG. 18. Comparison of the performance of the $r_{2\cdot}$ criteria for one and two scalar errors in samples of size 15, $\alpha = 5\%$.

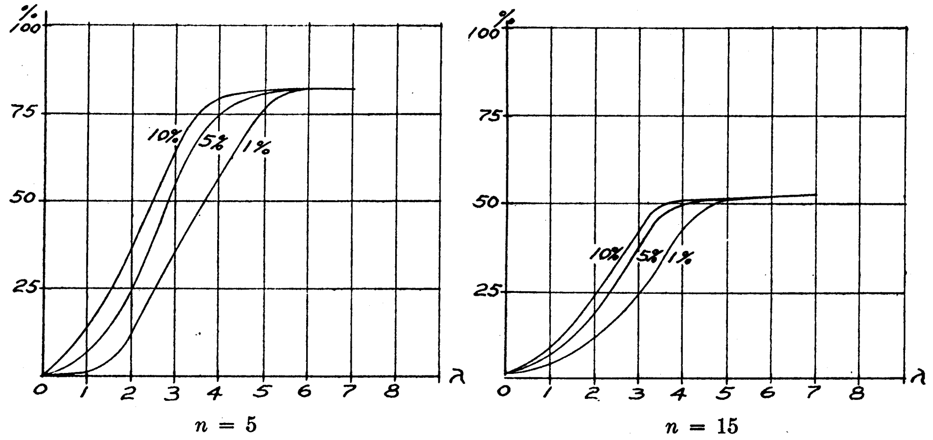


FIG. 19. Performance of B_1 for various levels of significance when the population is 10% contaminated with location errors.

6. Sampling from a contaminated population. In the previous sections the performance of the various criteria were assessed for samples where a certain number of contaminators were present. One might well ask why a test is needed is it is known that contaminators are present. It would seem more realistic to state that a certain per cent of contamination will occur in the long run and that one will not know in any particular case whether 0, 1, 2, · · · contaminators will be present. One would then wish a criterion to indicate the presence of contamination in a particular sample.

The performances of these criteria will be investigated for the same two models of contamination and their performances will be reported as per cent of

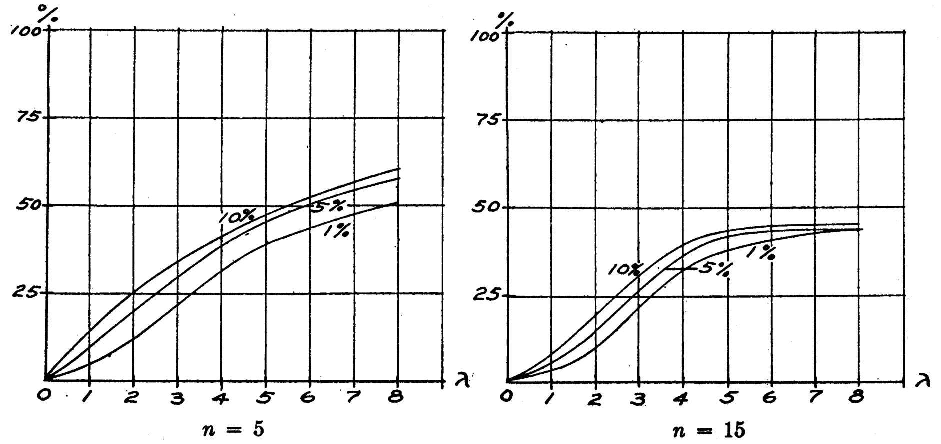
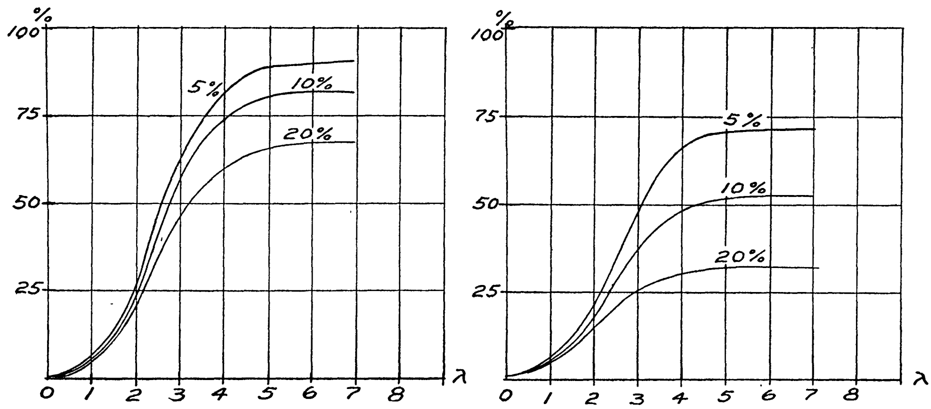


FIG. 20. Performance of B_1 for various levels of significance when the population is 10% contaminated with scalar errors.

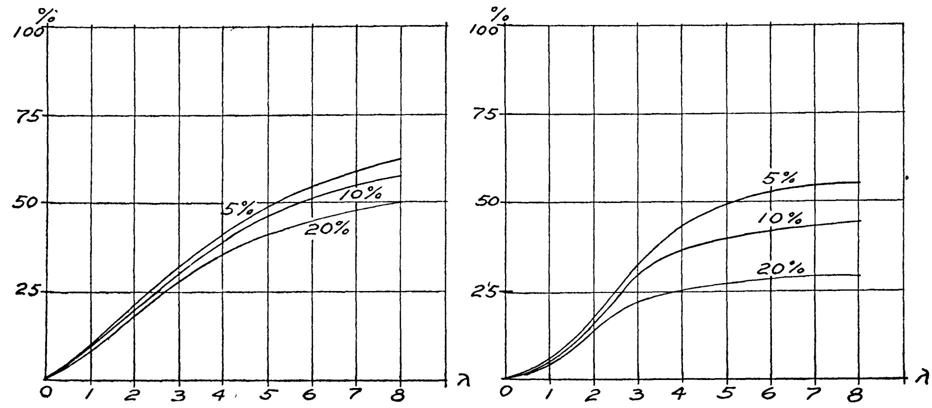


$n = 5$ $n = 15$

FIG. 21. Performance of B_1 for various levels of contamination for location errors and using the 5% level of significance.

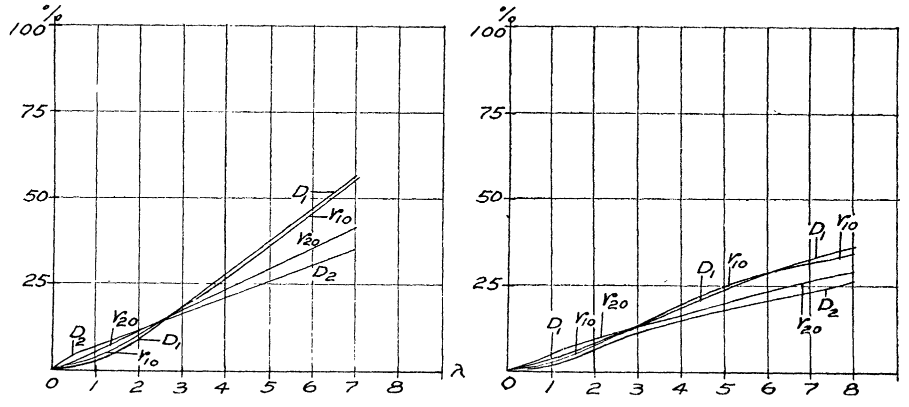
total contamination discovered. The tests will be applied only once to each sample. Repeated use of the criterion would in many cases increase the per cent of total contamination discovered. It is not known what effect such a procedure would have on the level of significance.

Investigation has been made for 5, 10, and 20% contamination. For example, in samples of size 5 which have 10% contamination, on the average, 59.0% of the samples will contain no "errors", 32.8% will contain one, 7.3% two, 0.8% three, 0.1% four, and 0.0% five. Thus in 100 samples of 5 which are 10% contaminated with location errors having mean $\mu + 5\sigma$, about 59 contain no errors. If the r_{10} criteria is used with a 5% level of significance one value will be "dis-



$n = 5$ $n = 15$

FIG. 22. Performance of B_1 for various levels of contamination for scalar errors and using the 5% level of significance.

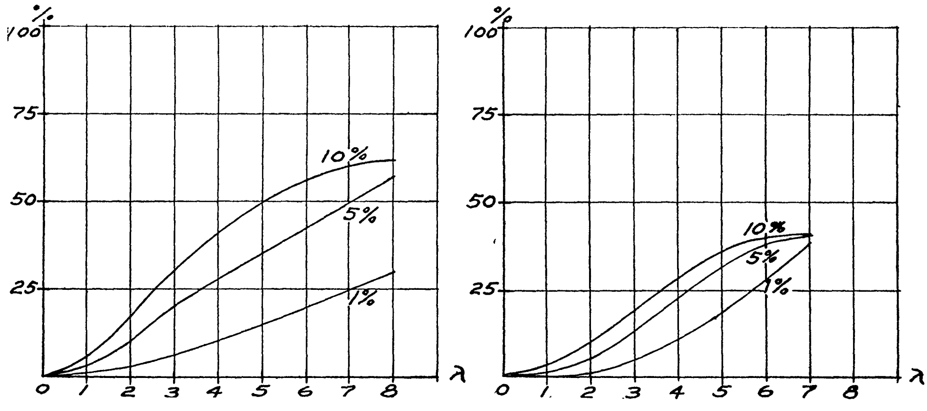


(Location) (Scalar)

FIG. 23. Performance of r_{10} , D_1 , r_{20} , D_2 in samples of size 5 using the 5% level of significance and sampling from a population which is 10% contaminated.

covered" in 3.0 of the samples containing no errors. Of the 33 samples containing one "error" the "error" would by discovered in 18 of these samples. This criteria would discover none of the "errors" in samples containing more than one "error". We would have obtained 18 of the 50 contaminating values and 3 which were members of the original population.

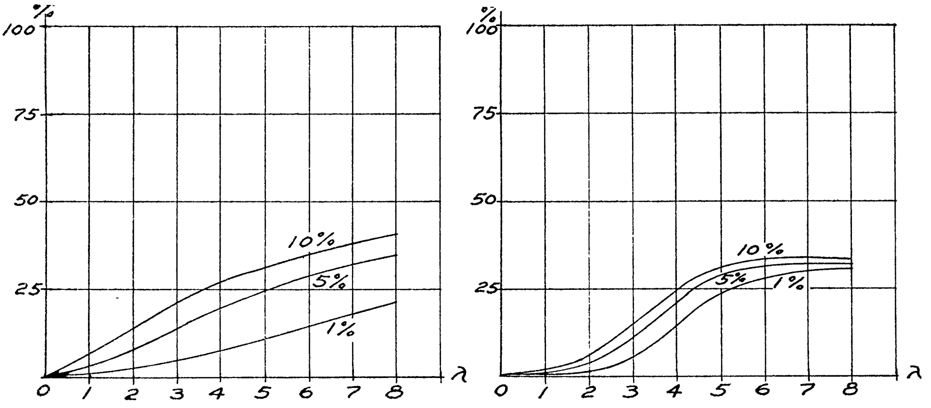
When σ is known the performance will increase when more contaminators are present. Performance however has been measured in terms of finding a single contaminator; i.e., the test has been used only once. Therefore even with increasing percent contamination the level of performance will decrease with increasing contamination. Repeated use of the test criteria has not been investigated.



$r_{10}(D_1)$
$n = 5$

$r_{22}(D_1 , r_{20} , r_{21})$
$n = 15$

FIG. 24. Performance of r_{10}(D$_1$) and $r_{22}(D_1 , r_{20} , r_{21})$ for various levels of significance when the population is 10% contaminated with location errors.



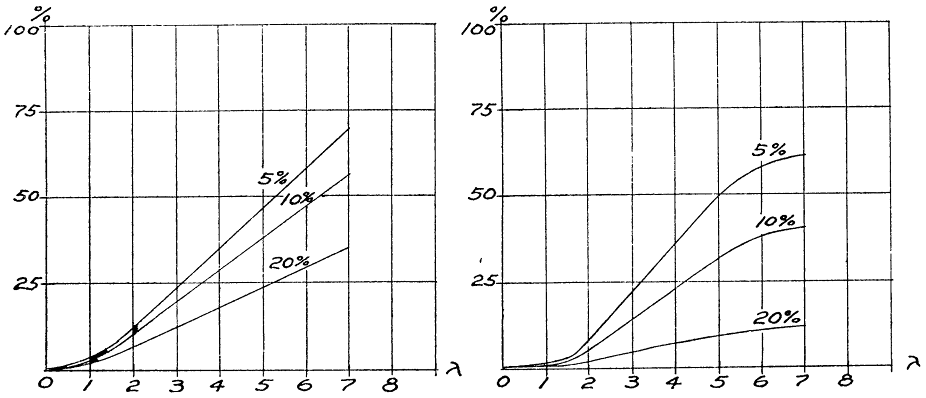
$r_{10}(D_1)$
$n = 5$

$r_{22}(D_1, r_{20}, r_{21})$
$n = 15$

FIG. 25. Performance of $r_{10}(D_1)$ and $r_{22}(D_1, r_{20}, r_{21})$ for various levels of significance when the population is 10% contaminated with scalar errors.

Criteria B_1 gives the best performance for both location and scalar errors for the levels of contamination and levels of significance considered. A and C_1 are only slightly inferior. B_2 is handicapped when more than one error is present thus its performance is poorer for heavier contamination. Figure 19 shows the performance of B_1 for the different levels of significance, 10% contamination, and the two sample sizes 5 and 15 for location errors. Figure 20 shows the results for scalar errors. Figures 21 and 22 show the performance of B_1 for the 5% level of significance for the different levels of contamination.

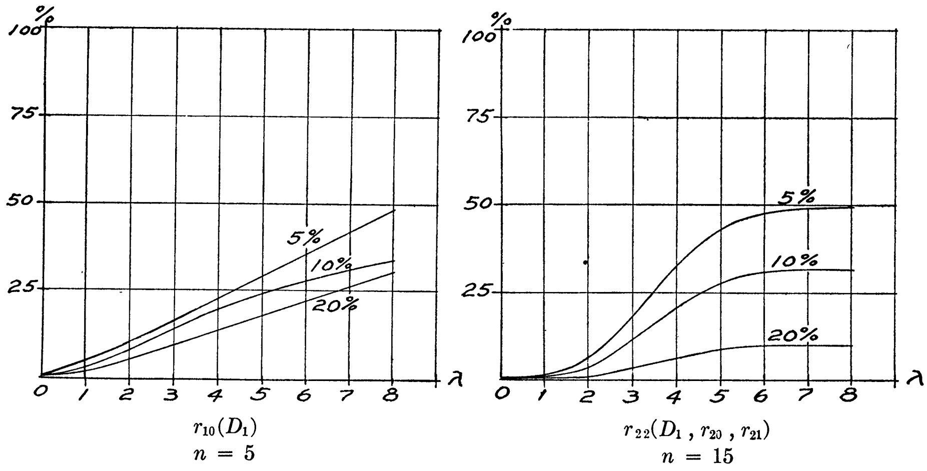
When σ is not known the performance of various criteria will eventually decrease as more and more contaminators are present in the sample even though



$r_{10}(D_1)$
$n = 5$

$r_{22}(D_1, r_{20}, r_{21})$
$n = 15$

FIG. 26. Performance of $r_{10}(D_1)$ and $r_{22}(D_1, r_{20}, r_{21})$ for various levels of contamination for location errors and using the 5% level of significance.



$r_{10}(D_1)$
$n = 5$

$r_{22}(D_1, r_{20}, r_{21})$
$n = 15$

FIG. 27. Performance of $r_{10}(D_1)$ and $r_{22}(D_1, r_{20}, r_{21})$ for various levels of contamination for scalar errors and the 5% level of significance, $\alpha = 5\%$.

several of the criteria show improvement in discovering a single error if two are present. The performance of these criteria is greatly affected by the size of the sample. For samples of size 5, r_{10} and D_1 perform alike, r_{10} being superior to the other r's (r_{20} second best) for the levels of contamination considered, and D_2 is inferior to r_{20}. Figure 23 compares the performance of r_{10}, D_1, r_{20}, and D_2 for the 5% level of significance and 10% contamination. The results for other levels of significance and contamination are comparable.

For samples of size 15, r_{20}, r_{21} and r_{22} perform alike as do r_{10}, r_{11} and r_{12}. D_1 and r_{20}, r_{21}, r_{22} perform approximately the same and are superior to r_{10}, r_{11},

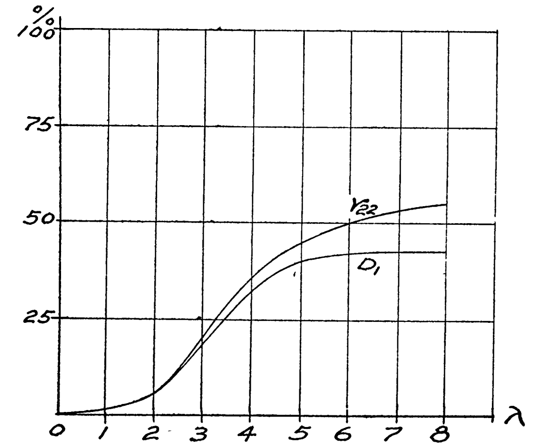


FIG. 28. A comparison of the performance of r_{22} and D_1 for two scalar contaminators when tests are made at one extreme only, $\alpha = 5\%$, $n = 15$.

and r_{12}. Critical values are not available for D_2 for $n > 12$. The performances of D_1, r_{20}, r_{21} and r_{22} are indicated by a single line in Figures 24, 25, 26, and 27 which show the effect of level of significance and level of contamination of the performance of D_1, r_{20}, r_{21} and r_{22} for samples of size 15 and for r_{10} (D_1) for samples of size 5.

7. Remarks and conclusions. Throughout the investigation of performance, location errors were placed only at one extreme and scalar errors at either extreme. The test for an error was made using as a suspected value the extreme value in the direction of the location error or in the case of the scalar error the value most distant from the mean. It can be expected then that if performance were assessed when location errors could occur in either direction, different results would be obtained. Also in the case of scalar errors if errors were always sought at one particular extreme or at both extremes different results would be obtained. If these changes were made in the models of contamination, those criteria designed to avoid errors at the other extreme would have an advantage over those which were not so designed for σ unknown. If σ is known the criteria which do not avoid the other extreme would have an advantage over those which do avoid the other extreme. These points just mentioned will be used to discriminate between those criteria which were judged to be equal in performance under the models used in the sampling study. For example, Figure 28 compares the performance of r_{22} and D_1 for two scalar contaminators when tests are made only at one extreme, $\alpha = 5\%$, $n = 15$.

1. For σ known:

B_1 or C_1 should be used, or in small samples A, B_1 or C_1 should be used.

2. For σ unknown:

r_{10} should be used for very small samples. r_{22} should be used for sample sizes over 15. Probably r_{21} would be best for sample sizes from about 8 to 13. If simplicity in computation is not important and "errors" are not expected at both extremes D_1 would do equally well. When critical values are available for larger n, D_2 should prove useful in the larger sample sizes.

LITERATURE REFERRING TO CRITERIA LISTED IN SECTION 3

(B_1) A. T. McKay, "The distribution of the difference between the extreme observation and the sample mean in samples of n from a normal universe," *Biometrika*, Vol. 27 (1935), pp. 466–471. Procedures for obtaining percentage values given.

(B_2) J. O. Irwin, "On a criterion for the rejection of outlying observations," *Biometrika*, Vol. 17 (1925), pp. 238–250. $Pr(B_2 > \lambda), \lambda = .1(.1)5.0$; $n = 2, 3, 10(10)100(100)1{,}000$. Tables concerning the second and thiŗd ordered observations are also given.

(C_1) E. S. Pearson and H. O. Hartley, "The probability integral of the range in samples of n observations from the normal population," *Biometrika*, Vol. 32 (1942), pp. 301–310. C.1%, 0.5%, 1.0%, 2.5%, 5%, 10%, $n = 2(1)12$, values to 20 available by interpolation.

(C_2) D. Newman, "The distribution of ranges in samples from a normal population, expressed in terms of an independent estimate of the standard deviation," *Biometrika*, Vol. 31 (1940), pp. 20–30. 1% and 5% points for C_2; for w, $n = 2(1)12, 20$; s, d.f. $= 5(1)20, 24, 30, 40, 60, \infty$.

(C_2) E. S. PEARSON AND H. O. HARTLEY, "Tables of the probability integral of the studentized range," *Biometrika*, Vol. 33 (1942), pp. 89–99. Upper and lower 5% and 1% points for C_2 ; for w, $n = 2(1)20$; for s, d.f. $= 10(1)20, 24, 30, 40, 60, 120, \infty$.

(C_2, B_1) K R. NAIR, "The distribution of the extreme deviate from the sample mean and its studentized forms," *Biometrika*, Vol. 35 (1948), pp. 118–144. B_1 upper and lower .1%, .5%, 1%, 2.5%, 5%, 10% points for $n = 3(1)9$.

(D_1 , D_2 , F, B_1) F. E. GRUBBS, "Sample criterion for testing outlying observations," *Annals of Math. Stat.*, Vol. 21 (1950), pp. 27–58. F, D_1 : 1%, 2.5%, 5%, 10%, $n \leq 25$; D_2: 1%, 2.5%, 5%, 10%, $n \leq 20$; B_1: 1%, 2.5%, 5%, 10%, $n \leq 25$.

(F) W. R. THOMPSON, "On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation," *Annals of Math. Stat.*, Vol. 6 (1935), pp. 214–219. 20%, 10%, 5%, $n = 3(1)22(10)42$, 102, 202, 502, 1002.

(F) E. S. PEARSON AND CHANDRA SEKAR give a further discussion of F in "The efficiency of statistical tools and a criterion for the rejection of outlying observations," *Biometrika*, Vol. 28 (1936), pp. 308–320. 10%, 5%, 2.5%, 1%, $n = 3(1)19$.

(r's) W. J. DIXON, "Ratios involving extreme values," *Annals of Math. Stat.*, to be published. r_{10} , r_{11} , r_{12} , r_{20} , r_{21} , r_{22} ; .5%, 1%, 2%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, $n \leq 30$.