American Society for Quality

Procedures for Detecting Outlying Observations in Samples

FRANK E. GRUBBS*

U. S. Army Aberdeen Research and Development Center
Aberdeen Proving Ground, Maryland 21005

Procedures are given for determining statistically whether the highest observation, the lowest observation, the highest and lowest observations, the two highest observations, the two lowest observations, or more of the observations in the sample are statistical outliers. Both the statistical formulae and the application of the procedures to examples are given, thus representing a rather complete treatment of tests for outliers in single samples. This paper has been prepared primarily as an *expository* and *tutorial* article on the problem of detecting outlying observations in much experimental work. We cover only tests of significance in this paper.

1. SCOPE OF PAPER

1.1 This is an expository and tutorial type of paper which deals with the problem of outlying observations in samples and how to test the statistical significance of them. An outlying observation, or "outlier," is one that appears to deviate markedly from other members of the sample in which it occurs. In this connection, the following two alternatives are of interest:

1.1.1 An outlying observation may be merely an extreme manifestation of the random variability inherent in the data. If this is true, the values should be retained and processed in the same manner as the other observations in the sample.

1.1.2 On the other hand, an outlying observation may be the result of gross deviation from prescribed experimental procedure or an error in calculating or recording the numerical value. In such cases, it may be desirable to institute an investigation to ascertain the reason for the aberrant value. The observation may even eventually be rejected as a result of the investigation, though not necessarily so. At any rate, in subsequent data analysis the outlier or outliers will be recognized as probably being from a different population than that of the sample values.

1.2 It is our purpose here to provide statistical rules that will lead the experi-

menter almost unerringly to look for causes of outliers when they really exist, and hence to decide whether alternative 1.1.1 above is not the more plausible hypothesis to accept as compared to alternative 1.1.2 in order that the most appropriate action in further data analysis may be taken. The procedures covered herein apply primarily to the simplest kind of experimental data, i.e., replicate measurements of some property of a given material, or observations in a supposedly single random sample. Nevertheless, the tests suggested do cover a wide enough range of cases in practice to have rather broad utility.

2. GENERAL

2.1 When the skilled experimenter is clearly aware that a gross deviation from prescribed experimental procedure has taken place, the resultant observations should be discarded, whether or not it agrees with the rest of the data and without recourse to statistical tests for outliers. If a reliable correction procedure, for example, for temperature, is available, the observation may sometimes be corrected and retained.

2.2 In many cases evidence for deviation from prescribed procedure will consist primarily of the discordant value itself. In such cases it is advisable to adopt a cautious attitude. Use of one of the criteria discussed below will sometimes permit a clear-cut decision to be made. In doubtful cases the experimenter's judgment will have considerable influence. When the experimenter cannot identify abnormal conditions, he should at least report the discordant values and indicate to what extent they have been used in the analysis of the data.

2.3 Thus, for purposes of orientation relative to the overall problem of experimentation, our position on the matter of screening samples for outlying observations is precisely the following:

Physical Reason Known or Discovered for Outlier(s)
  (i)   Reject observation(s)
  (ii) Correct observation(s) on physical grounds
  (iii) Reject it (them) and possibly take additional observation(s)

Physical Reason Unknown—Use Statistical Test
  (i)   Reject observation(s)
  (ii) Correct observation(s) statistically
  (iii) Reject it (them) and possibly take additional observation(s)
  (iv) Employ truncated sample theory for censored observations

2.4 The statistical test may always be used to lend support to a judgment that a physical reason does actually exist for an outlier, or the statistical criterion may be used routinely as a basis to initiate action to find a physical cause.

3. BASIS OF STATISTICAL CRITERIA FOR OUTLIERS

3.1 There are a number of criteria for testing outliers. In all of these the doubtful observation is included in the calculation of the numerical value of a

sample criterion (or statistic), which is then compared with a critical value based on the theory of random sampling to determine whether the doubtful observation is to be retained or rejected. The critical value is that value of the sample criterion which would be exceeded by chance with some specified (small) probability on the assumption that all the observations did indeed constitute a random sample from a common system of causes, a single parent population, distribution or universe. The specified small probability is called the "significance levels" or "percentage point" and can be thought of as the risk of erroneously rejecting a good observation. It becomes clear, therefore, that if there exists a real shift or change in the value of an observation that arises from non-random causes (human error, loss of calibration of instrument, change of measuring instrument, or even change of time of measurements, etc.), then the observed value of the sample criterion used would exceed the "critical value" based on random sampling theory. Tables of critical values are usually given for several different significance levels, for example, 5%, 1%. For statistical tests of outlying observations, it is generally recommended that a low significance level, such as 1%, be used and that significance levels greater than 5% should not be common practice. (Note 1).

3.2 It should be pointed out that almost all criteria for outliers are based on an assumed underlying normal (Gaussian) population or distribution. When the data are not normally or approximately normally distributed, the probabilities associated with these tests will be different. Until such time as criteria not sensitive to the normality assumption are developed, the experimenter is cautioned against interpreting the probabilities too literally when normality of the data is not assured.

3.3 Although our primary interest here is that of detecting outlying observations, we remark that the statistical criteria used also test the hypothesis that the random sample taken did indeed come from a normal or Gaussian population. The end result is for all practical purposes the same, i.e., we really want to know once and for all whether we have in hand a sample of homogeneous observations.

4. RECOMMENDED CRITERIA FOR SINGLE SAMPLES

4.1 Let the sample of n observations be denoted in order of increasing magnitude by $x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_n$. Let x_n be the doubtful value, i.e. the largest value. The test criterion, T_n, recommended here for a single outlier is as follows:

$$T_n = (x_n - \bar{x})/s$$

where

$\bar{x}$ = arithmetic average of all n values, and

Note 1: In this paper, we will usually illustrate the use of the 5% significance level. Proper choice of level in probability depends on the particular problem and just what may be involved, along with the risk that one is willing to take in rejecting a good observation, i.e., if the null-hypothesis stating "all observations in the sample come from the same normal population" may be assumed.

TABLE 1

Table of Critical Values for T (One-sided Test) When Standard Deviation is Calculated from the Same Sample

Number of Observations n	5% Significance Level	2.5% Significance Level	1% Significance Level
3	1.15	1.15	1.15
4	1.46	1.48	1.49
5	1.67	1.71	1.75
6	1.82	1.89	1.94
7	1.94	2.02	2.10
8	2.03	2.13	2.22
9	2.11	2.21	2.32
10	2.18	2.29	2.41
11	2.23	2.36	2.48
12	2.29	2.41	2.55
13	2.33	2.46	2.61
14	2.37	2.51	2.66
15	2.41	2.55	2.71
16	2.44	2.59	2.75
17	2.47	2.62	2.79
18	2.50	2.65	2.82
19	2.53	2.68	2.85
20	2.56	2.71	2.88
21	2.58	2.73	2.91
22	2.60	2.76	2.94
23	2.62	2.78	2.96
24	2.64	2.80	2.99
25	2.66	2.82	3.01
30	2.75	2.91	
35	2.82	2.98	
40	2.87	3.04	
45	2.92	3.09	
50	2.96	3.13	
60	3.03	3.20	
70	3.09	3.26	
80	3.14	3.31	
90	3.18	3.35	
100	3.21	3.38	

$$T_n = \frac{x_n - \bar{x}}{s} \qquad s = \left\{ \frac{\sum (x_i - \bar{x})^2}{n-1} \right\}^{\frac{1}{2}} = \left\{ \frac{n \sum x_i^2 - (\sum x_i)^2}{n(n-1)} \right\}^{\frac{1}{2}}$$

$$T_1 = \frac{\bar{x} - x_1}{s} \qquad x_1 \le x_2 \le \cdots \le x_n$$

Note: Values of T for $n \le 25$ are based on those given in Reference [8]. For $n > 25$, the values of T are approximated. All values have been adjusted for division by $n - 1$ instead of n in calculating s.

s = estimate of the population standard deviation based on the sample data, calculated with $n - 1$ degrees of freedom as follows:

$$s = \left\{ \frac{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2}{n - 1} \right\}^{\frac{1}{2}} = \left\{ \frac{n \sum x_i^2 - (\sum x_i)^2}{n(n - 1)} \right\}^{\frac{1}{2}} = \sqrt{A_{xx}/n(n - 1)}$$

If x_1 rather than x_n is the doubtful value, the criterion is as follows:

$$T_1 = (\bar{x} - x_1)/s$$

The critical values for either case, for the 1 per cent and 5 per cent levels of significance, are given in Table 1. Table 1 and the following tables give the "one-sided" significance levels. (In a previous ASTM tentative recommended practice (1961), the tables listed values of significance levels double those in the present practice, since it was considered that the experimenter would test either the lowest or the highest observation (or both) for statistical significance. However, to be consistent with actual practice and in an attempt to avoid further misunderstanding, single-sided significance levels are tabulated here so that both viewpoints can be represented.)

4.2 The hypothesis that we are testing in every case is that all observations in the sample come from the same normal population. Let us adopt, for example, a significance level of 0.05. If we are interested *only* in outliers that occur on the *high side*, we should always use the statistic $T_n = (x_n - \bar{x})/s$ and take as critical value the 0.05 point of Table 1. On the other hand, if we are interested *only* in outliers occurring on the *low side*, we would always use the statistic $T_1 = (\bar{x} - x_1)/s$ and again take as a critical value the 0.05 point of Table 1. Suppose, however, that we are interested in outliers occurring on *either side*, but do not believe that outliers can occur on both sides simultaneously. We might, for example, believe that at some time during the experiment something possibly happened to cause an extraneous variation on the high side or on the low side, but that it was very unlikely that two or more such events could have occurred, one being an extraneous variation on the high side *and* the other an extraneous variation on the low side. With this point of view we should use the statistic $T_n = (x_n - \bar{x})/s$ or the statistic $T_1 = (\bar{x} - x_1)/s$ which ever is larger. If in this instance we use the 0.05 point of Table 1 as our critical value, the true significance level would be twice 0.05 or 0.10. If we wish a significance level of 0.05 and not 0.10, we must in this case use as a critical value the 0.025 point of Table 1. Similar considerations apply to the other tests given below.

Example 1

As an illustration of the use of T_n and Table 1, consider the following ten observations on breaking strength (in pounds) of 0.104-in. hard-drawn copper wire: 568, 570, 570, 570, 572, 572, 572, 578, 584, 596. The doubtful observation is the high value, $x_{10} = 596$. Is the value of 596 significantly high? The mean is $\bar{x} = 575.2$ and the estimated standard deviation is $s = 8.70$. We compute

$$T_{10} = (596 - 575.2)/8.70 = 2.39$$

From Table 1, for $n = 10$, note that a T_{10} as large as 2.39 would occur by chance

with probability less than 0.05. In fact, so large a value would occur by chance not much oftener than 1% of the time. Thus, the weight of the evidence is against the doubtful value having come from the same population as the others (assuming the population is normally distributed). Investigation of the doubtful value is therefore indicated.

4.3 An alternative system, the Dixon criteria, based entirely on ratios of differences between the observations is described in the literature [5] and may be used in cases where it is desirable to avoid calculation of s or where quick judgment is called for. For the Dixon test, the sample criterion or statistic changes with sample size. Table 2 gives the appropriate statistic to calculate and also gives the critical values of the statistic for the 1%, 5% and 10% levels of significance.

Example 2

As an illustration of the use of Dixon's test, consider again the observations on breaking strength given in Example 1, and suppose that a large number of such samples had to be screened quickly for outliers and it was judged too time-consuming to compute s. Table 2 indicates use of

$$r_{11} = \frac{x_n - x_{n-1}}{x_n - x_2} \quad \text{for a sample size of ten. Thus, for } n = 10,$$

$$r_{11} = \frac{x_{10} - x_9}{x_{10} - x_2}.$$

For the measurements of breaking strength above,

$$r_{11} = \frac{596 - 584}{596 - 570} = .462$$

which is a little less than .477, the 5% critical value for $n = 10$. Under the Dixon criterion, we should therefore *not* consider this observation as an outlier at the 5% level of significance. This illustrates how border-line cases may be accepted under one test but rejected under another. It should be remembered, however, that the T-statistic discussed above is the best one to use for the single-outlier case, and final statistical judgment should be based on it. See Ferguson, References [6], [7].

Further examination of the sample observations on breaking strength of hard-drawn copper wire indicates that none of the other values needs testing. (Note 2.)

4.4 A test equivalent to T_n (or T_1) based on the sample sum of squared deviations from the mean for all the observations and the sum of squared deviations omitting the "outlier" is given by Grubbs in [8]

4.5 The next type of problem to consider is the case where we have the possibility of two outlying observations, the least and the greatest observation, in a

---

Note 2: With experience we may usually just look at the sample values to observe if an outlier is present. However, strictly speaking the statistical test should be applied to all samples to guarantee the significance levels used. Concerning "multiple" tests on a single sample, we comment on this below.

sample. (The problem of testing the two highest or the two lowest observations is considered below.) In testing the least and the greatest observations simultaneously as probable outliers in a sample, we use the ratio of sample range to sample standard deviation test of David, Hartley and Pearson [4]. The significance levels for this sample criterion are given in Table 3. An example in astronomy follows.

Example 3

There is one rather famous set of observations that a number of writers on the subject of outlying observations have referred to in applying their various tests for "outliers". This classic set consists of a sample of 15 observations of the

TABLE 2

*Dixon Criteria for Testing of Extreme Observation (Single Sample)**

n	Criterion		Significance Level		
			10%	5%	1%
3	$r_{10} = \dfrac{x_2 - x_1}{x_n - x_1}$	if smallest value	.886	.941	.988
4		is suspected;	.679	.765	.889
5		if largest value	.557	.642	.780
6	$= \dfrac{x_n - x_{n-1}}{x_n - x_1}$	is suspected.	.482	.560	.698
7			.434	.507	.736
8	$r_{11} = \dfrac{x_2 - x_1}{x_{n-1} - x_1}$	if smallest value	.479	.554	.683
9		is suspected;	.441	.512	.635
10			.409	.477	.597
	$\dfrac{x_n - x_{n-1}}{x_n - x_2}$	if largest value is suspected.			
11	$r_{21} = \dfrac{x_3 - x_1}{x_{n-1} - x_1}$	if smallest value is suspected;	.517	.576	.679
12			.490	.546	.642
13			.467	.521	.615
	$= \dfrac{x_n - x_{n-2}}{x_n - x_2}$	if largest value is suspected.			
14	$r_{22} = \dfrac{x_3 - x_1}{x_{n-2} - x_1}$	if smallest value is suspected.	.492	.546	.641
15			.472	.525	.616
16			.454	.507	.595
17	$= \dfrac{x_n - x_{n-2}}{x_n - x_3}$	if largest value is suspected;	.438	.490	.577
18			.424	.475	.561
19			.412	.462	.547
20			.401	.450	.535
21			.391	.440	.524
22			.382	.430	.514
23			.374	.421	.505
24			.367	.413	.497
25			.360	.406	.489

* From W. J. Dixon, "Processing Data for Outliers", *Biometrics*, March 1953, Vol. 9, No. 1, Appendix, Page 89. (Reference [5]) $x_1 \leq x_2 \leq \cdots \leq x_n$

"vertical semi-diameters of Venus made by Lieutenant Herndon in 1846 and given in William Chauvenet's *A Manual of Spherical and Practical Astronomy*, Vol. II (5th ed., 1876). In the reduction of the observations, Prof. Pierce assumed two unknown quantities and found the following residuals which have been arranged in ascending order of magnitude:

−1.40″	−0.24	−0.05	0.18	0.48
−0.44	−0.22	0.06	0.20	0.63
−0.30	−0.13	0.10	0.39	1.01

TABLE 3

Critical Values for w/s (Ratio of Range to Sample Standard Deviation)*

Number of Observations n	5% Significance Level	1% Significance Level	0.5% Significance Level
3	2.00	2.00	2.00
4	2.43	2.44	2.45
5	2.75	2.80	2.81
6	3.01	3.10	3.12
7	3.22	3.34	3.37
8	3.40	3.54	3.58
9	3.55	3.72	3.77
10	3.68	3.88	3.94
11	3.80	4.01	4.08
12	3.91	4.13	4.21
13	4.00	4.24	4.32
14	4.09	4.34	4.43
15	4.17	4.43	4.53
16	4.24	4.51	4.62
17	4.31	4.59	4.69
18	4.38	4.66	4.77
19	4.43	4.73	4.84
20	4.49	4.79	4.91
30	4.89	5.25	5.39
40	5.15	5.54	5.69
50	5.35	5.77	5.91
60	5.50	5.93	6.09
80	5.73	6.18	6.35
100	5.90	6.36	6.54
150	6.18	6.64	6.84
200	6.38	6.85	7.03
500	6.94	7.42	7.60
1000	7.33	7.80	7.99

* Taken from H. A. David, H. O. Hartley and E. S. Pearson, "The Distribution of the Ratio in a Single Sample of Range to Standard Deviation," Biometrika, Vol. 41 (1954), pp. 482–493. (Reference [4])

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

$$w = x_n - x_1$$
$$x_1 \leq x_2 \leq \cdots \leq x_n$$

The deviations -1.40 and 1.01 appear to be outliers. Here the suspected observations lie at each end of the sample. Much less work has been accomplished for the case of outliers at both ends of the sample than for the case of one or more outliers at only one end of the sample. This is not necessarily because the "one-sided" case occurs more frequently in practice but because "two-sided" tests are more difficult to deal with. For a high and a low outlier in a single sample, the procedure below may possess near optimum properties. For optimum procedures when there is at hand an independent estimate, s^2 of σ^2, see "Some Tests for Outliers" by C. P. Quesenberry and H. A. David, Technical Report No. 47, OOR (ARO) project No. 1166, Virginia Polytechnic Institute, Blacksburg, Virginia.

4.6 For the observations on the semi-diameters of Venus given above, all the information on the measurement error is contained in the sample of 15 residuals. In cases like this, where no independent estimate of variance is available (i.e. we still have the single sample case), a useful statistic is the ratio of the range of the observations to the sample standard deviation:

$$\frac{w}{s} = \frac{x_n - x_1}{s} \quad \text{where} \quad s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$$

If x_n is about as far above the mean, $\bar{x}$, as x_1 is below $\bar{x}$, and if w/s exceeds some chosen critical value, then one would conclude that *both* the doubtful values are outliers. If, however, x_1 and x_n are displaced from the mean by different amounts, some further test would have to be made to decide whether to reject as outlying only the lowest value or only the highest value or both the lowest and highest values.

4.7 For this example the mean of the deviations is $\bar{x} = .018$, $s = .551$, and

$$w/s = \frac{1.01 - (-1.40)}{.551} = \frac{2.41}{.551} = 4.374$$

From Table 3 for $n = 15$, we see that the value of $w/s = 4.374$ falls between the critical values for the 1% and 5% levels, so if the test were being run at the 5% level of significance, we would conclude that this sample contains one or more outliers. The lowest measurement, $-1.40''$, is $1.418''$ below the sample mean, and the highest measurement, $1.01''$, is $.992''$ above the mean. Since these extremes are not symmetric about the mean, either *both* extremes are outliers or else only -1.40 is an outlier. That -1.40 is an outlier can be verified by use of the T_1 statistic. We have

$$T_1 = (\bar{x} - x_1)/s = \frac{.018 - (-1.40)}{.551} = 2.574 \quad \text{and from}$$

Table 1 this value is greater than the critical value for the 5% level, so we reject -1.40. Since we have decided that -1.40 should be rejected, we use the remaining 14 observations and test the upper extreme 1.01, either with the criterion

$$T_n = \frac{x_n - \bar{x}}{s}$$

or with Dixon's r_{22} . Omitting $-1.40''$ and renumbering the observations, we compute $\bar{x} = 1.67/14 = .119$, $s = .401$, and

$$T_{14} = \frac{1.01 - .119}{.401} = 2.22$$

From Table 1, for $n = 14$, we find that a value as large as 2.22 would occur by chance more than 5% of the time, so we should retain the value 1.01 in further calculations. We next calculate Dixon's sample criterion:

$$r_{22} = \frac{x_{14} - x_{12}}{x_{14} - x_3} = \frac{1.01 - .48}{1.01 + .24} = \frac{.53}{1.25}$$

or

$$r_{22} = .424$$

From Table 2 for $n = 14$, we see that the 5% critical value for r_{22} is .546. Since our calculated value (.424) is less than the critical value, we also retain 1.01 by Dixon's test, and no further values would be tested in this sample. (Note 3.)

4.8 We next turn to the case where we may have the two largest or the two smallest observations as probable outliers. Here, we employ a test provided by Grubbs [8] which is based on the ratio of the sample sum of squares when the two doubtful values are omitted to the sample sum of squares when the two doubtful values are included. If simplicity in calculation is the prime require-ment, then the Dixon type of test (actually omitting one observation in the sample) might be used for this case. In illustrating the test procedure, we give the following Examples 4 and 5.

Example 4

In a comparison of strength of various plastic materials, one characteristic studied was the per cent elongation at break. Before comparison of the average elongation of the several materials, it was desirable to isolate for further study any pieces of a given material which gave very small elongation at breakage com-pared with the rest of the pieces in the sample. In this example, one might have primary interest only in outliers to the left of the mean for study, since very high readings indicate exceeding plasticity, a desirable characteristic.

Following are ten measurements of per cent elongation at break made on material No. 23: 3.73, 3.59, 3.94, 4.13, 3.04, 2.22, 3.23, 4.05, 4.11, 2.02. Arranged in ascending order of magnitude, these measurements are: 2.02, 2.22, 3.04, 3.23, 3.59, 3.73, 3.94, 4.05, 4.11, 4.13. The questionable readings are the two lowest, 2.02 and 2.22. We can test these two low readings simultaneously by using the criterion $S_{1,2}^2/S^2$ of Table 4. For the above measurements:

$$S^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{n \sum x_i^2 - (\sum x_i)^2}{n} = \frac{10(121.3594) - (34.06)^2}{10}$$

$$S^2 = 5.351$$

Note 3: It should be noted that in a multiplicity of tests of this kind, the final overall significance level will be less than that used in the individual tests, as we are offering more than one chance of accepting the sample as one produced by a random operation. It is not our purpose here to cover the theory of multiple tests.

and

$$S^2_{1,2} = \sum_{i=3}^{n} (x_i - \bar{x}_{1,2})^2 = \frac{(n-2)\sum_{i=3}^{n} x_i^2 - \left(\sum_{i=3}^{n} x_i\right)^2}{(n-2)}$$

$$\left(\text{where} \quad \bar{x}_{1,2} = \sum_{i=3}^{n} x_i/(n-2)\right)$$

$$= \frac{8(112.3506) - (29.82)^2}{8}$$

$$S^2_{1,2} = \frac{9.5724}{8} = 1.197$$

TABLE 4

*Critical Values for $S^2_{n-1,n}/S^2$ or $S^2_{1,2}/S^2$ for Simultaneously Testing the Two Largest or Two Smallest Observations**

Number of Observations n	10% Significance Level	5% Significance Level	1% Significance Level
4	.0031	.0008	.0000
5	.0376	.0183	.0035
6	.0921	.0565	.0186
7	.1479	.1020	.0440
8	.1994	.1478	.0750
9	.2454	.1909	.1082
10	.2853	.2305	.1415
11	.3226	.2666	.1736
12	.3552	.2996	.2044
13	.3843	.3295	.2333
14	.4106	.3568	.2605
15	.4345	.3818	.2859
16	.4562	.4048	.3098
17	.4761	.4259	.3321
18	.4944	.4455	.3530
19	.5113	.4636	.3725
20	.5269	.4804	.3909

$$S^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad x_1 \le x_2 \le \cdots \le x_n$$

$$S^2_{1,2} = \sum_{i=3}^{n} (x_i - \bar{x}_{1,2})^2 \qquad \bar{x}_{1,2} = \frac{1}{n-2}\sum_{i=3}^{n} x_1$$

$$S^2_{n-1,n} = \sum_{i=1}^{n-2} (x_i - \bar{x}_{n-1,n})^2 \qquad \bar{x}_{n-1,n} = \frac{1}{n-2}\sum_{i=1}^{n-2} x_i$$

* These significance levels are taken from Table V of Grubbs, Reference [8]. An observed ratio less than the appropriate critical ratio in this table calls for rejection of the null hypothesis.

We find

$$\frac{S_{1,2}^2}{S^2} = \frac{1.197}{5.351} = .224$$

From Table 4 for $n = 10$, the 5% significance level for $S_{1,2}^2/S^2$ is .2305. Since the calculated value is less than the critical value, we should conclude that both 2.02 and 2.22 are outliers. In a situation such as the one described in this example, where the outliers are to be isolated for further analysis, a singificance level as high as perhaps even 10% would probably be used in order to get a reasonable size of sample for additional study.

Example 5

The following ranges (horizontal distances in yards from gun muzzle to point of impact of a projectile) were obtained in firings from a weapon at a constant angle of elevation and at the same weight of charge of propellant powder:

<div style="text-align:center">

Distances in Yards

4782	4420
4838	4803
4765	4730
4549	4833

</div>

It is desired to make a judgment on whether the projectiles exhibit uniformity in ballistic behavior or if some of the ranges are inconsistent with the others. The doubtful values are the two smallest ranges, 4420 and 4549. For testing these two suspected outliers, the statistic $S_{1,2}^2/S^2$ of Table 4 is probably the best to use. (Note 4.)

The distances arranged in increasing order of magnitude are:

<div style="text-align:center">

4420	4782
4549	4803
4730	4833
4765	4838

</div>

The value of S^2 is 158,592. Omission of the two shortest ranges, 4420 and 4549, and recalculation gives $S_{1,2}^2$ equal to 8590.8. Thus,

$$\frac{S_{1,2}^2}{S^2} = \frac{8590.8}{158,592} = .054$$

which is significant at the .01 level (See Table 4). It is thus highly unlikely that the two shortest ranges (occurring actually from excessive yaw) could have come from the same population as that represented by the other six ranges. It should be noted that the critical values in Table 4 for the 1% level of significance are smaller than those for the 5% level. So for this particular test, the calculated value is significant if it is *less* than the chosen critical value.

Note 4: Kudo [11] indicates that if the two outliers are due to a shift in location or level, as compared to the scale σ, then the optimum sample criterion for testing should be of the type: min. $(2\bar{x} - x_i - x_j)/s = (2\bar{x} - x_1 - x_2)/s$ in our Example 5.

4.9 If simplicity in calculation is very important, or if a large number of samples must be examined individually for outliers, the questionable observations may be tested with the application of Dixon's criteria. Disregarding the lowest range, 4420 we test if the next lowest range 4549 is outlying. With $n = 7$, we see from Table 2 that r_{10} is the appropriate statistic. Renumbering the ranges as x_1 to x_7, beginning with 4549, we find

$$r_{10} = \frac{x_2 - x_1}{x_7 - x_1} = \frac{4730 - 4549}{4838 - 4549} = \frac{181}{289} = .626$$

which is only a little less than the 1% critical value, .637, for $n = 7$. So, if the test is being conducted at any significance level greater than the 1% level, we would conclude that 4549 is an outlier. Since the lowest of the original set of ranges, 4420, is even more outlying than the one we have just tested, it can be classified as an outlier without further testing. We note here, however, that this test did not use all of the sample observations.

4.10 *Rejection of Several Outliers.* So far we have discussed procedures for detecting one or two outliers in the same sample, but these techniques are not generally recommended for repeated rejection, since if several outliers are present in the sample the detection of one or two spurious values may be "masked" by the presence of other anomalous observations. Outlying observations occur due to a shift in level (or mean), or a change in scale (i.e., change in variance of the observations), or both. Ferguson [6, 7] has studied the power of the various rejection rules relative to changes in level or scale. For several outliers and repeated rejection of observations, Ferguson points out that the sample coefficient of skewness

$$\sqrt{b_1} = \sqrt{n} \sum_{i=1}^{n} (x_i - \bar{x})^3/(n - 1)^{\frac{3}{2}}s^3 = \sqrt{n} \sum_{i=1}^{n} (x_i - \bar{x})^3/[\sum (x_i - \bar{x})^2]^{\frac{3}{2}}$$

should be used for "one-sided" tests (change in level of several observations in the same direction), and the sample coefficient of kurtosis

$$b_2 = n \sum_{i=1}^{n} (x_i - \bar{x})^4/(n - 1)^2 s^4 = n \sum_{i=1}^{n} (x_i - \bar{x})^4/[\sum (x_i - \bar{x})^2]^2$$

is recommended for "two-sided" tests (change in level to higher and lower values) and also for changes in scale (variance)*. In applying the above tests, the $\sqrt{b_1}$ or the b_2, or both, are computed and if their observed values exceed those for significance levels given in the following tables, then the observation farthest from the mean is rejected and the same procedure repeated until no further sample values are judged as outliers. [As is well-known $\sqrt{b_1}$ and b_2 are also used as tests of Normality].

4.10.1 The significance levels in the following tables for sample sizes of 5, 10, 15 and 20 (and 25 for b_2) were obtained by Ferguson on an IBM 704 Computer using a sampling experiment or "Monte Carlo" procedure. The

*In the above equations for $\sqrt{b_1}$ and b_2, s is defined as used in this paper, i.e.

$$s = \sum_{i=1}^{n} (x_i - \bar{x})^2/(n - 1)$$

significance levels for the other sample sizes are from E. S. Pearson, "Table of Percentage Points of $\sqrt{b_1}$ and b_2 in Normal Samples; a Rounding Off," Biometrika (1965), Vol. 52, pp. 282–285.

Significance Levels for $\sqrt{b_1}$

Sig Level	n									
	5*	10*	15*	20*	25	30	35	40	50	60
1%	1.34	1.31	1.20	1.11	1.06	.98	.92	.87	.79	.72
5%	1.05	.92	.84	.79	.71	.66	.62	.59	.53	.49

Significance Levels for b_1

Sig Level	n							
	5*	10*	15*	20*	25*	50	75	100
1%	3.11	4.83	5.08	5.23	5.00	4.88	4.59	4.39
5%	2.89	3.85	4.07	4.15	4.00	3.99	3.87	3.77

* These values were obtained by Ferguson, using a Monte Carlo procedure. For $n = 25$, Ferguson's Monte Carlo values of b_2 agree with Pearson's computed values.

4.10.2 The $\sqrt{b_1}$ and b_2 statistics have the optimum property of being "locally" best against one-sided and two-sided alternatives, respectively. The $\sqrt{b_1}$ test is good for up to 50% spurious observations in the sample for the one-sided case and the b_2 test is optimum in the two-sided alternatives case for up to 21% "contamination" of sample values. For only one or two outliers the sample statistics of the previous paragraphs are recommended, and Ferguson [7] discusses in detail their optimum properties of *pointing out* one or two outliers.

5. RECOMMENDED CRITERION USING INDEPENDENT STANDARD DEVIATION

5.1 Suppose that an independent estimate of the standard deviation is available from previous data. This estimate may be from a single sample of previous similar data or may be the result of combining estimates from several such previous sets of data. In any event, each estimate is said to have degrees of freedom equal to one less than the sample size that it is based on. The proper combined estimate is a weighted average of the several values of s^2, the weights being proportional to the respective degrees of freedom. The total degrees of freedom in the combined estimate is then the sum of the individual degrees of freedom. When one uses an independent estimate of the standard deviation, s_ν, the test criterion recommended here for an outlier is as follows:

$$T'_1 = \frac{\bar{x} - x_1}{s_\nu} \quad (\nu = \text{total number of degrees of freedom})$$

or

$$T'_n = \frac{x_n - \bar{x}}{s_\nu}.$$

5.2 The critical values for T_1' and T_n' for the 5% and 1% significance levels are due to David [3] and are given in Table 5. In Table 5 the subscript ν = df indicates the total number of degrees of freedom associated with the independent estimate of standard deviation σ and n indicates the number of observations

TABLE 5

Critical Values for T When Standard Deviation s_ν is Independent of Present Sample

$$T' = \frac{x_n - \bar{x}}{s_\nu} \text{ or } \frac{\bar{x} - x_1}{s_\nu}$$

n	3	4	5	6	7	8	9	10	12
ν = df					1% points				
10	2.78	3.10	3.32	3.48	3.62	3.73	3.82	3.90	4.04
11	2.72	3.02	3.24	3.39	3.52	3.63	3.72	3.79	3.93
12	2.67	2.96	3.17	3.32	3.45	3.55	3.64	3.71	3.84
13	2.63	2.92	3.12	3.27	3.38	3.48	3.57	3.64	3.76
14	2.60	2.88	3.07	3.22	3.33	3.43	3.51	3.58	3.70
15	2.57	2.84	3.03	3.17	3.29	3.38	3.46	3.53	3.65
16	2.54	2.81	3.00	3.14	3.25	3.34	3.42	3.49	3.60
17	2.52	2.79	2.97	3.11	3.22	3.31	3.38	3.45	3.56
18	2.50	2.77	2.95	3.08	3.19	3.28	3.35	3.42	3.53
19	2.49	2.75	2.93	3.06	3.16	3.25	3.33	3.39	3.50
20	2.47	2.73	2.91	3.04	3.14	3.23	3.30	3.37	3.47
24	2.42	2.68	2.84	2.97	3.07	3.16	3.23	3.29	3.38
30	2.38	2.62	2.79	2.91	3.01	3.08	3.15	3.21	3.30
40	2.34	2.57	2.73	2.85	2.94	3.02	3.08	3.13	3.22
60	2.29	2.52	2.68	2.79	2.88	2.95	3.01	3.06	3.15
120	2.25	2.48	2.62	2.73	2.82	2.89	2.95	3.00	3.08
∞	2.22	2.43	2.57	2.68	2.76	2.83	2.88	2.93	3.01
					5% points				
10	2.01	2.27	2.46	2.60	2.72	2.81	2.89	2.96	3.08
11	1.98	2.24	2.42	2.56	2.67	2.76	2.84	2.91	3.03
12	1.96	2.21	2.39	2.52	2.63	2.72	2.80	2.87	2.98
13	1.94	2.19	2.36	2.50	2.60	2.69	2.76	2.83	2.94
14	1.93	2.17	2.34	2.47	2.57	2.66	2.74	2.80	2.91
15	1.91	2.15	2.32	2.45	2.55	2.64	2.71	2.77	2.88
16	1.90	2.14	2.31	2.43	2.53	2.62	2.69	2.75	2.86
17	1.89	2.13	2.29	2.42	2.52	2.60	2.67	2.73	2.84
18	1.88	2.11	2.28	2.40	2.50	2.58	2.65	2.71	2.82
19	1.87	2.11	2.27	2.39	2.49	2.57	2.64	2.70	2.80
20	1.87	2.10	2.26	2.38	2.47	2.56	2.63	2.68	2.78
24	1.84	2.07	2.23	2.34	2.44	2.52	2.58	2.64	2.74
30	1.82	2.04	2.20	2.31	2.40	2.48	2.54	2.60	2.69
40	1.80	2.02	2.17	2.28	2.37	2.44	2.50	2.56	2.65
60	1.78	1.99	2.14	2.25	2.33	2.41	2.47	2.52	2.61
120	1.76	1.96	2.11	2.22	2.30	2.37	2.43	2.48	2.57
∞	1.74	1.94	2.08	2.18	2.27	2.33	2.39	2.44	2.52

The above percentage points are reproduced from H. A. David, "Revised upper percentage points of the extreme studentized deviate from the sample mean," Biometrika, Vol. 43 (1956), pp. 449–451. (Reference [3]).

Standardization of Sodium Hydroxide Solutions as Determined by Plant Laboratories
Standard Used: Potassium Acid Phthalate (P.A.P)

Laboratory	(P.A.P.-.096000) × 10³	Sums	Averages	Deviation of Average from Grand Average
1	1.893 1.972 1.876	5.741	1.914	+ .043
2	2.046 1.851 1.949	5.846	1.949	+ .078
3	1.874 1.792 1.829	5.495	1.832	− .039
4	1.861 1.998 1.983	5.842	1.947	+ .076
5	1.922 1.881 1.850	5.653	1.884	+ .013
6	2.082 1.958 2.029	6.069	2.023	+ .152
7	1.992 1.980 2.066	6.038	2.013	+ .142
8	2.050 2.181 1.903	6.134	2.045	+ .174
9	1.831 1.883 1.855	5.569	1.856	− .015
10	.735 .722 .777	2.234	.745	−1.126
11	2.064 1.794 1.891	5.749	1.916	+ .045
12	2.475 2.403 2.102	6.980	2.327	+ .456

Grand Sum   67.350

Grand Average   1.871

in the sample under study. We illustrate with an example on interlaboratory testing.

5.3 *Example 6—Interlaboratory Testing.* In an analysis of interlaboratory test procedures, data representing normalities of sodium hydroxide solutions were determined by twelve different laboratories. In all the standardizations, a tenth normal sodium hydroxide solution was prepared by the Standard Methods Committee using carbon-dioxide-free distilled water, Potassium acid phthalate (P. A. P.), obtained from the National Bureau of Standards, was used as the test standard.

Test data by the twelve laboratories are given in the table below. The P. A. P. readings have been coded to simplify the calculations. The variances between the three readings within all laboratories were found to be homogeneous. A one-way classification in the analysis of variance was first analyzed to determine if the variation in laboratory results (averages) was statistically significant. This variation was significant, so tests for outliers were then applied to isolate the particular laboratories whose results gave rise to the significant variation. We are indebted to Dr. Grant Wernimont of the Eastman Kodak Co. for the data on Standardization of Sodium Hydroxide Solutions.

Analysis of Variance

Source of Variation	Degrees of Freedom d.f.	Sum of Squares SS	Mean Square MS	F-ratio
Between Labs	11	4.70180	.4274	F = 48.61 (Highly
Within Labs	24	.21103	.008793	Significant)
TOTAL	35	4.91283		

The above analysis of variance shows that the variation between laboratories is highly significant. To test if this (very significant) variation is due to one (or perhaps two) laboratories that obtained "outlying" results (i.e. perhaps showing non-standard technique), we can test the laboratory averages for outliers. From the analysis of variance, we have an estimate of the variance of an individual reading as .008793, based on 24 degrees of freedom. The estimated standard deviation of an individual measurement is $\sqrt{.008793} = .094$ and the estimated standard deviation of the average of three readings is therefore $.094/\sqrt{3} = .054$.

Since the estimate of within-laboratory variation is independent of any difference between laboratories, we can use the statistic T'_1 of section 5.1 to test for outliers. An examination of the deviations of the laboratory averages from the grand average indicates that Laboratory 10 obtained an average reading much lower than the grand average, and that Laboratory 12 obtained a high average compared to the overall average. To first test if Laboratory 10 is an outlier, we compute

$$T' = \frac{1.871 - .745}{.054} = 20.9$$

This value of T' is obviously significant at a very low level of probability

($P \ll .01$. Refer to Table 5 with $n = 12$ and $\nu = 24$ d.f.). We conclude therefore that the test methods of Laboratory 10 should be investigated.

Excluding Laboratory 10, we compute a new grand average of 1.973 and test if the results of Laboratory 12 are outlying. We have

$$T' = \frac{2.327 - 1.973}{.054} = 6.56$$

and this value of T' is significant at $P \ll .01$ (Refer to Table 5 with $n = 11$ and $\nu = 24$ d.f.). We conclude that the procedures of Laboratory 12 should also be investigated.

To verify that the remaining laboratories did indeed obtain homogeneous results, we might repeat the analysis of variance omitting Laboratories 10 and 12. This calculation gives

Analysis of Variance
(omitting labs 10 and 12)

Source of Variation	d.f.	SS	MS	F-ratio
Between Labs	9	.13889	.01543	F = 2.36
Within Labs	20	.13107	.00655	$F_{.05}(9, 20) = 2.40$
				$F_{.01}(9, 20) = 3.45$
TOTAL	29	.26996		

For this analysis, the variation between labs is not significant at the 5% level and we conclude that all the laboratories except No. 10 and No. 12 exhibit the same capability in testing procedure.

In conclusion, there should be a systematic investigation of test methods for Laboratories No. 10 and No. 12 to determine why their test precedures are apparently different from the other ten laboratories.

(For the above example, procedures for ranking means after the initial analysis of variance test could, of course, have been used. For example, Duncan's Multiple Range Test, Scheffe's Test, Tukey's procedure, etc., could have been used. Also, the test of Halperin, Greenhouse and Cornfield [9] could have been used. We have used David's tables [3] as an example here since they seem tailor-made for one or two specific laboratories.)

6. Recommended Criteria for Known Standard Deviation

6.1 Frequently the population standard deviation σ may be known accurately. In such cases, Table 6 may be used for single outliers and we illustrate with the following example.

6.2 *Example 7 (σ known).* Passage of the Echo I (Balloon) Satellite was recorded on star-plates when it was visible. Photographs were made by means of a camera with shutter automatically timed to obtain a series of points for the Echo path. Since the stars were also photographed at the same times as the Satellite, all the pictures show star-trails and so are called "star-plates."

TABLE 6

Critical Values of $T'_{1\infty}$ and $T'_{n\infty}$ When the Population Standard Deviation σ is Known

Number of Observations n	5% Significance Level	1% Significance Level	0.5% Significance Level
2	1.39	1.82	1.99
3	1.74	2.22	2.40
4	1.94	2.43	2.62
5	2.08	2.57	2.76
6	2.18	2.68	2.87
7	2.27	2.76	2.95
8	2.33	2.83	3.02
9	2.39	2.88	3.07
10	2.44	2.93	3.12
11	2.48	2.97	3.16
12	2.52	3.01	3.20
13	2.56	3.04	3.23
14	2.59	3.07	3.26
15	2.62	3.10	3.29
16	2.64	3.12	3.31
17	2.67	3.15	3.33
18	2.69	3.17	3.36
19	2.71	3.19	3.38
20	2.73	3.21	3.39
21	2.75	3.22	3.41
22	2.77	3.24	3.42
23	2.78	3.26	3.44
24	2.80	3.27	3.45
25	2.81	3.28	3.46

$$x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_n \qquad T'_{1\infty} = (\bar{x} - x_1)/\sigma \qquad T'_{n\infty} = (x_n - \bar{x})/\sigma$$

This table is taken from the paper of Grubbs, Reference [8].

The x- and y-coordinate of each point on the Echo path are read from a photograph, using a stereo-comparator. To eliminate bias of the reader, the photograph is placed in one position and the coordinates are read; then the photograph is rotated 180° and the coordinates reread. The average of the two readings is taken as the final reading. Before any further calculations are made, the readings must be "screened" for gross reading or tabulation errors. This is done by examining the difference in the readings taken at the two positions of the photograph.

Recorded below are a sample of six readings made at the two positions and the differences in these readings. On the third reading, the differences are rather large. Has the operator made an error in positioning the cross-hair on the point?

For this example, an independent estimate of σ is available since extensive tests on the stero-comparator have shown that the standard deviation in reader's error is about 4 microns. The determination of this standard error was based on such a large sample that we can assume $\sigma = 4$ microns. The standard deviation of the difference in two readings is therefore $\sqrt{4^2 + 4^2} = \sqrt{32}$ or 5.7 microns.

Measurement (Microns)

x-Coordinate			y-Coordinate		
Pos. 1	Pos. 1 + 180°	Δx	Pos 1	Pos. 1 + 180°	Δy
−53011	−53004	− 7	70263	70258	+ 5
−38112	−38103	− 9	−39729	−39723	− 6
− 2804	− 2828	+24	81162	81140	+22
18473	18467	+ 6	41477	41485	− 8
25507	25497	+10	1082	1076	+ 6
87736	87739	− 3	− 7442	− 7434	− 8

For the six readings above, the mean difference in the x-coordinates is $\bar{\Delta}_x = 3.5$ and the mean difference in the y-coordinates is $\bar{\Delta}_y = 1.8$. For the questionable third reading, we have

$$T'_x = \frac{24 - 3.5}{5.7} = 3.60$$

$$T'_y = \frac{22 - 1.8}{5.7} = 3.54$$

From Table 6 we see that for $n = 6$, values of $T'_{n\infty}$ as large as the calculated values would occur by chance less than 1% of the time so that a significant reading error seems to have been made on the third point.

6.3 A great number of points are read and automatically tabulated on star-plates. Here we have chosen a very small sample of these points. In actual practice, the tabulations would probably be scanned quickly for very large errors such as tabulator errors; then some rule-of-thumb such as ±3 standard deviations of reader's error might be used to scan for outliers due to operator error. (Note 5). In other words, the data are probably too extensive to allow repeated use of precise tests like those described above, (especially for varying sample size) but this example does illustrate the case where σ is assumed known. If gross disagreement is found in the two readings of a coordinate, then the reading could be omitted or reread before further computations are made.

7. ADDITIONAL COMMENTS

7.1 In the above, we have covered only that part of screening samples to detect outliers statistically. However, a large area remains after the decision has been reached that outliers are present in data. Once some of the sample observations are branded as "outliers", then a thorough investigation should be initiated to determine the cause. In particular, one should look for gross errors, personal errors, errors of measurement, errors in calibration, etc. If reasons are found for aberrant observations, then one should act accordingly and perhaps scrutinize also the other observations. Finally, if one reaches the point that some observations are to be discarded or treated in a special manner

Note 5: Note that the values of Table 6 vary between about 1.4σ and 3.5σ.

based solely on statistical judgment, then it must be decided what action should be taken in the further analysis of the data. We do not propose to cover this problem here, since in many cases it will depend greatly on the particular case in hand. However, we do remark that there could be the outright rejection of aberrant observations once and for all on physical grounds (and preferably not on statistical grounds generally), and only the remaining observations would be used in further analyses or in estimation problems. On the other hand, some may want to replace aberrant values with newly taken observations and others may want to "Winsorize" the outliers i.e., replace them with the next closest values in the sample. Also with outliers in a sample, some may wish to use the median instead of the mean, and so on. Finally, we remark that perhaps a fair or appropriate practice might be that of using truncated sample theory (Note 6) for cases of samples where we have "censored" or rejected some of the observations. We cannot go further into these problems here. For additional reading on outliers, however, see References [1], [2], [3], [10], [12], [13], and [14].

REFERENCES

[1] ANSCOMBE, F. J., 1960. Rejection of outliers. Technometrics, Vol. 2, No. 2, pp. 123–147.

[2] CHEW, VICTOR, 1964. Tests for the rejection of outlying observations. RCA Systems Analysis Technical Memorandum No. 64–7, Patrick Air Force Base, Florida.

[3] DAVID, H. A., 1956. Revised upper percentage points of the extreme studentized deviate from the sample mean. Biometrika, Vol. 43, pp. 449–451.

[4] DAVID, H. A., HARTLEY, H. O. and PEARSON, E. S., 1954. The distribution of the ratio in a single sample of range to standard deviation. Biometrika, Vol. 41, pp. 482–493.

[5] DIXON, W. J., 1953. Processing data for outliers. Biometrics, Vol. 9, No. 1, pp. 74–89.

[6] FERGUSON, THOMAS S., 1961. On the rejection of outliers. Fourth Berkeley Symposium on Mathematical Statistics and Probability, edited by Jerzy Neyman. University of California Press, Berkeley and Los Angeles.

[7] FERGUSON, THOMAS S., 1961. Rules for rejection of outliers. Revue Inst. Int. de Stat., Vol. 3, pp. 29–43.

[8] GRUBBS, FRANK E., 1950. Sample criteria for testing outlying observations. Annals of Mathematical Statistics, Vol. 21, pp. 27–58.

[9] HALPERIN, M., GREENHOUSE, S. W. and CORNFIELD, J., 1955. Tables of percentage points for the studentized maximum absolute deviation in normal samples. Journal of the American Statistical Association, Vol. 50, No. 269, pp. 185–195.

[10] KRUSKAL, W. H., 1960. Some remarks on wild observations. Technometrics, Vol. 2, No. 1, pp. 1–3.

[11] KUDO, A., 1956. On the testing of outlying observations. Sankhyā, The Indian Journal of Statistics, Vol. 17, Part 1, pp. 67–76.

[12] PROSCHAN, F., 1957. Testing suspected observations. Industrial Quality Control, Vol. XIII, No. 7, pp. 14–19.

[13] SARHAN, A. E. and GREENBERG, B. G., Editors, 1962. Contributions to Order Statistics. John Wiley and Sons, Inc.

[14] THOMPSON, W. R., 1935. On a criterion for the rejection of observations and the distribution of the ratio of the deviation to the sample standard deviation. The Annals of Mathematical Statistics, Vol. 6, pp. 214–219.

Note 6: See Reference [13], for example.