Designation: ~~E178 – 16a~~ E178 – 21

An American National Standard

Standard Practice for Dealing With Outlying Observations[1]

This standard is issued under the fixed designation E178; the number immediately following the designation indicates the year of original adoption or, in the case of revision, the year of last revision. A number in parentheses indicates the year of last reapproval. A superscript epsilon (ε) indicates an editorial change since the last revision or reapproval.

~~Note—Corrections were made to Table 2 and the year date was changed on Sept. 7, 2016.~~

1. Scope

1.1 This practice covers outlying observations in samples and how to test the statistical significance of outliers.

1.2 The system of units for this standard is not specified. Dimensional quantities in the standard are presented only as illustrations of calculation methods. The examples are not binding on products or test methods treated.

1.3 *This standard does not purport to address all of the safety concerns, if any, associated with its use. It is the responsibility of the user of this standard to establish appropriate ~~safety~~ safety, health, and ~~health~~environmental practices and determine the applicability of regulatory ~~requirements~~limitations prior to use.*

1.4 *This international standard was developed in accordance with internationally recognized principles on standardization established in the Decision on Principles for the Development of International Standards, Guides and Recommendations issued by the World Trade Organization Technical Barriers to Trade (TBT) Committee.*

2. Referenced Documents

2.1 *ASTM Standards:*[2]

E456 Terminology Relating to Quality and Statistics

E2586 Practice for Calculating and Using Basic Statistics

3. Terminology

3.1 *Definitions*—~~The terminology defined in Terminology~~—Unless otherwise noted in this ~~E456 applies to this standard unless modified herein.~~standard, all terms relating to quality and statistics are defined in Terminology E456.

3.1.1 *null hypothesis, H_0, n*—a statement about a parameter of a probability distribution or about the type of probability distribution, tentatively regarded as true until rejected using a statistical hypothesis test. **E2586**

3.1.2 *order statistic $x_{(k)}$, n*—value of the kth observed value in a sample after sorting by order of magnitude. **E2586**

3.1.2.1 *Discussion*—

In this practice, x_k is used to denote order statistics in place of $x_{(k)}$, to simplify the notation.

3.1.3 *outlier*—see **outlying observation**.

[1] This practice is under the jurisdiction of ASTM Committee E11 on Quality and Statistics and is the direct responsibility of Subcommittee E11.10 on Sampling / Statistics.

Current edition approved ~~Sept. 7, 2016~~June 1, 2021. Published ~~September 2016~~June 2021. Originally approved in 1961. Last previous edition approved in 2016 as ~~E178 – 16.~~E178 – 16a. DOI: ~~10.1520/E0178-16A.~~10.1520/E0178-21.

[2] For referenced ASTM standards, visit the ASTM website, www.astm.org, or contact ASTM Customer Service at service@astm.org. For *Annual Book of ASTM Standards* volume information, refer to the standard's Document Summary page on the ASTM website.

3.1.4 *outlying observation, n*—an extreme observation in either direction that appears to deviate markedly in value from other members of the sample in which it appears.

3.1.4.1 *Discussion*—
The identification of a value as outlying, and therefore a doubtful observation, is a judgement of the analyst and can be made before any statistical test.

4. Significance and Use

4.1 An outlying observation, or "outlier," is an extreme one in either direction that appears to deviate markedly from other members of the sample in which it occurs.

4.2 Statistical rules test the null hypothesis of no outliers against the alternative of one or more actual outliers. The procedures covered were developed primarily to apply to the simplest kind of experimental data, that is, replicate measurements of some property of a given material or observations in a supposedly random sample.

4.3 A statistical test may be used to support 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.

5. Procedure

5.1 In dealing with an outlier, the following alternatives should be considered:

5.1.1 An outlying observation might be the result of gross deviation from prescribed experimental procedure or an error in calculating or recording the numerical value. When the experimenter is clearly aware that a deviation from prescribed experimental procedure has taken place, the resultant observation 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 is available, the observation may sometimes be corrected and retained.

5.1.2 An outlying observation might be merely an extreme manifestation of the random variability inherent in the data. If this is true, the value should be retained and processed in the same manner as the other observations in the sample. Transformation of data or using methods of data analysis designed for a non-normal distribution might be appropriate.

5.1.3 Test units that give outlying observations might be of special interest. If this is true, once identified they should be segregated for more detailed study. Outliers may contain important information for a possible root cause analysis and action on the process or procedure.

5.2 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 clearcut decision to be made.

5.2.1 When the experimenter cannot identify abnormal conditions, ~~he~~they should report the discordant values and indicate to what extent they have been used in the analysis of the data.

5.3 Thus, as part of the over-all process of experimentation, the process of screening samples for outlying observations and acting on them is the following:

5.3.1 *Physical Reason Known or Discovered for Outlier(s):*

5.3.1.1 Reject observation(s) and possibly take additional observation(s).

5.3.1.2 Correct observation(s) on physical grounds.

5.3.2 *Physical Reason Unknown—Use Statistical Test:*

5.3.2.1 Reject observation(s) and possibly take additional observation(s).

5.3.2.2 Transform observation(s) to improve fit to a normal distribution.

5.3.2.3 Use estimation appropriate for non-normal distributions.

5.3.2.4 Segregate samples for further study.

6. Basis of Statistical Criteria for Outliers

6.1 In testing outliers, 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 level" or "percentage point" and can be thought of as the risk of erroneously rejecting a good observation. If a real shift or change in the value of an observation arises from nonrandom causes (human error, loss of calibration of instrument, change of measuring instrument, or even change of time of measurements, and so forth), then the observed value of the sample criterion used will exceed the "critical value" based on random-sampling theory. Tables of critical values are usually given for several different significance levels. In particular for this practice, significance levels 10, 5, and 1 % are used.

NOTE 1—In this practice, 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, that is, if the null-hypothesis stating "all observations in the sample come from the same normal population" may be assumed correct.

6.2 Almost all criteria for outliers are based on an assumed underlying normal (Gaussian) population or distribution. The null hypothesis that we are testing in every case is that all observations in the sample come from the same normal population. In choosing an appropriate alternative hypothesis (one or more outliers, separated or bunched, on same side or different sides, and so forth) it is useful to plot the data as shown in the dot diagrams of the figures. When the data are not normally or approximately normally distributed, the probabilities associated with these tests will be different. The experimenter is cautioned against interpreting the probabilities too literally.

6.3 Although our primary interest here is that of detecting outlying observations, some of the statistical criteria presented may also be used to test the hypothesis of normality or that the random sample taken come from a normal or Gaussian population. The end result is for all practical purposes the same, that is, we really wish to know whether we ought to proceed as if we have in hand a sample of homogeneous normal observations.

6.4 One should distinguish between data to be used to estimate a central value from data to be used to assess variability. When the purpose is to estimate a standard deviation, it might be seriously underestimated by dropping too many "outlying" observations.

7. Recommended Criteria for Single Samples

7.1 *Criterion for a Single Outlier*—~~Let~~Sort the ~~sample of~~ n observations ~~be denoted~~ in order of increasing magnitude by $x_1 \le x_2 \le x_3 \le \ldots \le x_n$. called *order statistics.* Let the largest value, x_n, be the doubtful value, that is the largest value. The test criterion, T_n, for a single outlier is as follows:

$$T_n = (x_n - \bar{x})/s \quad (1)$$

where:

$\bar{x}$ = arithmetic average of all n values, and
s = estimate of the population standard deviation based on the sample data, calculated as follows:
~~s~~ =

$$\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}x_i^2 - n\cdot\bar{x}^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2/n}{n-1}}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i \quad (2)$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\cdot\bar{x}^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 /n}{n-1}} \quad (3)$$

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

$$T_1 = (\bar{x} - x_1)/s \quad (4)$$

The critical values for either case, for the 1, 5, and 10 % levels of significance, are given in Table 1.

7.1.1 The test criterion T_n can be equated to the Student's t test statistic for equality of means between a population with one observation x_n and another with the remaining observations $x_1, \ldots, x_{n-1}$, and the critical value of T_n for significance level α can be approximated using the α/n percentage point of Student's t with n – 2 degrees of freedom. The approximation is exact for small enough values of α, depending on n, and otherwise a slight overestimate unless both α and n are large:

$$T_n(\alpha) \le \frac{t_{\alpha/n,n-2}}{\sqrt{1+\frac{nt^2_{\alpha/n,n-2}-1}{(n-1)^2}}} \quad (5)$$

7.1.2 To test outliers on the *high side*, use the statistic $T_n = (x_n - \bar{x})/s$ and take as critical value the 0.05 point of Table 1. To test outliers on the *low side*, use the statistic $T_1 = (\bar{x} - x_1)/s$ and again take as a critical value the 0.05 point of Table 1. If we are interested in outliers occurring on *either side*, use the statistic $T_n = (x_n - \bar{x})/s$ or the statistic $T_1 = (\bar{x} - x_1)/s$ whichever 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. Similar considerations apply to the other tests given below.

7.1.3 *Example 1*—As an illustration of the use of T_n and Table 1, consider the following ten observations on breaking strength (in

TABLE 1 Critical Values for T (One-Sided Test) When Standard Deviation is Calculated from the Same Sample[A]

Number of Observations, n	Upper 10 % Significance Level	Upper 5 % Significance Level	Upper 1 % Significance Level
3	1.1484	1.1531	1.1546
4	1.4250	1.4625	1.4925
5	1.602	1.672	1.749
6	1.729	1.822	1.944
7	1.828	1.938	2.097
8	1.909	2.032	2.221
9	1.977	2.110	2.323
10	2.036	2.176	2.410
11	2.088	2.234	2.485
12	2.134	2.285	2.550
13	2.175	2.331	2.607
14	2.213	2.371	2.659
15	2.247	2.409	2.705
16	2.279	2.443	2.747
17	2.309	2.475	2.785
18	2.335	2.504	2.821
19	2.361	2.532	2.854
20	2.385	2.557	2.884
21	2.408	2.580	2.912
22	2.429	2.603	2.939
23	2.448	2.624	2.963
24	2.467	2.644	2.987
25	2.486	2.663	3.009
26	2.502	2.681	3.029
27	2.519	2.698	3.049
28	2.534	2.714	3.068
29	2.549	2.730	3.085
30	2.563	2.745	3.103
35	2.628	2.811	3.178
40	2.682	2.866	3.240
45	2.727	2.914	3.292
50	2.768	2.956	3.336

[A] Values of T are taken from Grubbs (1),[3] Table 1. All values have been adjusted for division by n – 1 instead of n in calculating s. Use Ref. (1) for higher sample sizes up to n = 147.

pounds) of 0.104-in. hard-drawn copper wire: 568, 570, 570, 570, 572, 572, 572, 578, 584, 596. See Fig. 1. 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 more often 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.

7.2 *Dixon Criteria for a Single Outlier*—An alternative system, the Dixon criteria (**2**),[3] based entirely on ratios of differences between the observations 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. In most situations, the Dixon criteria is less powerful at detecting an outlier than the criterion given in 7.1.

7.2.1 *Example 2*—As an illustration of the use of Dixon's test, consider again the observations on breaking strength given in Example 1. Table 2 indicates use of:

$$r_{11} = (x_n - x_{n-1})/(x_n - x_2) \tag{6}$$

Thus, for n = 10:

$$r_{11} = (x_{10} - x_9)/(x_{10} - x_2) \tag{7}$$

For the measurements of breaking strength above:

$$r_{11} = (596 - 584)/(596 - 570) = 0.462$$

Which is a little less than 0.478, 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. These results illustrate how borderline cases may be accepted under one test but rejected under another.

7.3 *Recursive Testing for Multiple Outliers in Univariate Samples*—For testing multiple outliers in a sample, recursive application of a test for a single outlier may be used. In recursive testing, a test for an outlier, x_1 or x_n, is first conducted. If this is found to be significant, then the test is repeated, omitting the outlier found, to test the point on the opposite side of the sample, or an additional point on the same side. The performance of most tests for single outliers is affected by masking, where the probability of detecting an outlier using a test for a single outlier is reduced when there are two or more outliers. Therefore, the recommended procedure is to use a criterion designed to test for multiple outliers, using recursive testing to investigate after the initial criterion is significant.

7.4 *Criterion for Two Outliers on Opposite Sides of a Sample*—In testing the least and the greatest observations simultaneously as probable outliers in a sample, use the ratio of sample range to sample standard deviation test of David, Hartley, and Pearson (**5**):

$$w/s = (x_n - x_1)/s \tag{8}$$

The significance levels for this sample criterion are given in Table 3. Alternatively, the largest residuals test of Tietjen and Moore (7.5) could be used.

7.4.1 *Example 3*—This classic set consists of a sample of 15 observations of the vertical semidiameters of Venus made by Lieutenant Herndon in 1846 (**6**). In the reduction of the observations, Prof. Pierce found the following residuals (in seconds of arc) which have been arranged in ascending order of magnitude. See Fig. 2, above.

FIG. 1 Ten Observations of Breaking Strength from Example 1

[3] The boldface numbers in parentheses refer to a list of references at the end of this standard.

TABLE 2 Dixon Criteria for Testing of Extreme Observation (Single Sample)[A]

n	Criterion	Significance Level (One-Sided Test)		
		10 %	5 %	1 %
3	$r_{10} = (x_2 - x_1)/(x_n - x_1)$ if smallest value is suspected;	0.886	0.941	0.988
4	$= (x_n - x_{n-1})/(x_n - x_1)$ if largest value is suspected	0.679	0.766	0.889
5		0.558	0.642	0.781
6		0.484	0.562	0.698
7		0.434	0.507	0.637
8	$r_{11} = (x_2 - x_1)/(x_{n-1} - x_1)$ if smallest value is suspected;	0.480	0.554	0.681
9	$= (x_n - x_{n-1})/(x_n - x_2)$ if largest value is suspected.	0.440	0.511	0.634
10		0.410	0.478	0.597
11	$r_{21} = (x_3 - x_1)/(x_{n-1} - x_1)$ if smallest value is suspected;	0.517	0.575	0.674
12	$= (x_n - x_{n-2})/(x_n - x_2)$ if largest value is suspected.	0.490	0.546	0.643
13		0.467	0.521	0.617
14	$r_{22} = (x_3 - x_1)/(x_{n-2} - x_1)$ if smallest value is suspected;	0.491	0.546	0.641
15	$= (x_n - x_{n-2})/(x_n - x_3)$ if largest value is suspected.	0.470	0.524	0.618
16		0.453	0.505	0.598
17		0.437	0.489	0.580
18		0.424	0.475	0.564
19		0.412	0.462	0.550
20		0.401	0.450	0.538
21		0.391	0.440	0.526
22		0.382	0.430	0.516
23		0.374	0.421	0.506
24		0.366	0.413	0.497
25		0.359	0.406	0.489
26		0.353	0.399	0.482
27		0.347	0.393	0.474
28		0.342	0.387	0.468
29		0.336	0.381	0.462
30		0.332	0.376	0.456
35		0.311	0.354	0.431
40		0.295	0.337	0.412
45		0.283	0.323	0.397
50		0.272	0.312	0.384

[A] $x_1 \leq x_2 \leq \ldots \leq x_n$. Original Table in Dixon **(2)**, Appendix. Critical values updated by calculations by Bohrer **(3)** and Verma-Ruiz **(4)**.

−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

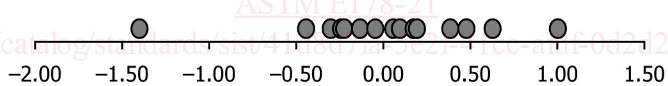

FIG. 2 Fifteen Residuals from the Semidiameters of Venus from Example 3

7.4.2 The deviations –1.40 and 1.01 appear to be outliers. Here the suspected observations lie at each end of the sample. The mean of the deviations is $x^{-} = 0.018$, the standard deviation is s = 0.551, and:

$$w/s = [1.01 - (-1.40)]/0.551 = 2.41/0.551 = 4.374$$

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.

7.4.3 The lowest measurement, –1.40, is 1.418 below the sample mean, and the highest measurement, 1.01, is 0.992 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:

$$T_1 = (\bar{x} - x_1)/s = [0.018 - (-1.40)]/0.551 = 2.574$$

This value is greater than the critical value for the 5 % level, 2.409 from Table 1, 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:

$$T_n = (x_n - \bar{x})/s \quad (9)$$

or with Dixon's r_{22}. Omitting –1.40 and renumbering the observations, we compute:

$$\bar{x} = 1.67/14 = 0.119, \; s = 0.401$$

and:

TABLE 3 Critical Values[A] (One-Sided Test) for *w/s* (Ratio of Range to Sample Standard Deviation)

Number of Observations, *n*	10 % Significance Level	5 % Significance Level	1 % Significance Level
3	1.9973	1.9993	2.0000
4	2.409	2.429	2.445
5	2.712	2.755	2.803
6	2.949	3.012	3.095
7	3.143	3.222	3.338
8	3.308	3.399	3.543
9	3.449	3.552	3.720
10	3.574	3.685	3.875
11	3.684	3.803	4.011
12	3.782	3.909	4.133
13	3.871	4.005	4.244
14	3.952	4.092	4.344
15	4.025	4.171	4.435
16	4.093	4.244	4.519
17	4.156	4.311	4.597
18	4.214	4.374	4.669
19	4.269	4.433	4.736
20	4.320	4.487	4.799
21	4.368	4.539	4.858
22	4.413	4.587	4.913
23	4.456	4.633	4.965
24	4.497	4.676	5.015
25	4.535	4.717	5.061
26	4.572	4.756	5.106
27	4.607	4.793	5.148
28	4.641	4.829	5.188
29	4.673	4.863	5.226
30	4.704	4.895	5.263
35	4.841	5.040	5.426
40	4.957	5.162	5.561
45	5.057	5.265	5.674
50	5.144	5.356	5.773

[A] Each entry calculated by 50 000 000 simulations.

$$T_{14} = (1.01 - 0.119)/0.401 = 2.22$$

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. The Dixon test criterion is:

$$\begin{aligned} r_{22} &= (x_{14} - x_{12})/(x_{14} - x_3) \\ &= (1.01 - 0.48)/(1.01 + 0.24) \\ &= 0.53/1.25 \\ &= 0.424 \end{aligned}$$

From Table 2 for $n = 14$, we see that the 5 % critical value for r_{22} is 0.546. Since our calculated value (0.424) 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.

7.5 *Criteria for Two or More Outliers on Opposite Sides of the Sample*—For suspected observations on both the high and low sides in the sample, and to deal with the situation in which some of $k \geq 2$ suspected outliers are larger and some smaller than the remaining values in the sample, Tietjen and Moore (7) suggest the following statistic. Let the sample values be x_1, x_2, x_3, ..., x_n. Compute the sample mean, $\bar{x}$, and the n absolute residuals:

$$r_1 = |x_1 - \bar{x}|, r_2 = |x_2 - \bar{x}|, \ldots, r_n = |x_n - \bar{x}| \tag{10}$$

Now relabel the original observations x_1, x_2, ..., x_n as z's in such a manner that z_i is that x whose r_i is the i^{th} smallest absolute residual above. This now means that z_1 is that observation x which is closest to the mean and that z_n is the observation x which is farthest from the mean. The Tietjen-Moore statistic for testing the significance of the k largest residuals is then:

$$E_k = \left[\sum_{i=1}^{n-k} (z_i - \bar{z}_k)^2 / \sum_{i=1}^{n} (z_i - \bar{z})^2\right] \tag{11}$$

where:

$$\bar{z}_k = \sum_{i=1}^{n-k} z_i/(n-k) \tag{12}$$