Designation: E2489 – 16

An American National Standard

Standard Practice for Statistical Analysis of One-Sample and Two-Sample Interlaboratory Proficiency Testing Programs[1]

This standard is issued under the fixed designation E2489; 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.

1. Scope

1.1 This practice describes methods for the statistical analysis of laboratory results obtained from interlaboratory proficiency testing programs. As in accordance with Practice E1301, proficiency testing is the use of interlaboratory comparisons for the determination of laboratory testing or measurement performance. Conversely, collaborative study (or collaborative trial) is the use of interlaboratory comparisons for the determination of the precision of a test method, as covered by Practice E691.

1.1.1 Method A covers testing programs using single test results obtained by testing a single sample (each laboratory submits a single test result).

1.1.2 Method B covers testing programs using paired test results obtained by testing two samples (each laboratory submits one test result for each of the two samples). The two samples should be of the same material or two materials similar enough to have approximately the same degree of variation in test results.

1.2 Methods A and B are applicable to proficiency testing programs containing a minimum of 10 participating laboratories.

1.3 The methods provide direction for assessing and categorizing the performance of individual laboratories based on the relative likelihood of occurrence of their test results, and for determining estimates of testing variation associated with repeatability and reproducibility. Assumptions are that a majority of the participating laboratories execute the test method properly and that samples are of sufficient homogeneity that the testing results represent results obtained from each laboratory testing essentially the same material. Each laboratory receives the same instructions or protocol.

1.4 *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 and health practices and determine the applicability of regulatory limitations prior to use.*

2. Referenced Documents

2.1 *ASTM Standards:*[2]

E177 Practice for Use of the Terms Precision and Bias in ASTM Test Methods

E178 Practice for Dealing With Outlying Observations

E456 Terminology Relating to Quality and Statistics

E691 Practice for Conducting an Interlaboratory Study to Determine the Precision of a Test Method

E1301 Guide for Proficiency Testing by Interlaboratory Comparisons (Withdrawn 2012)[3]

E2586 Practice for Calculating and Using Basic Statistics

3. Terminology

3.1 *Definitions*—The terminology defined in Terminology E456 applies to this practice unless modified herein.

3.1.1 *collaborative study, n*—interlaboratory study in which each laboratory uses the defined method of analysis to analyze identical portions of homogeneous materials to assess the performance characteristics obtained for that method of analysis. **Horwitz**[4]

3.1.2 *collaborative trial, n*—see *collaborative study*.

3.1.3 *interlaboratory comparison, n*—organization, performance, and evaluation of tests on the same or similar test items by two or more laboratories in accordance with predetermined conditions.

3.1.4 *median, $\tilde{X}$, n*—the 50th percentile in a population or sample. **E2586**

3.1.4.1 *Discussion*—The sample median is the *[(n + 1)/2]* order statistic if the sample size *n* is odd and is the average of the *[n/2]* and *[n/2 + 1]* order statistics if *n* is even.

3.1.5 *outlier, n*—see *outlying observation*. **E178**

[1] This practice is under the jurisdiction of ASTM Committee E11 on Quality and Statistics and is the direct responsibility of Subcommittee E11.20 on Test Method Evaluation and Quality Control.

Current edition approved Nov. 15, 2016. Published November 2016. Originally approved in 2006. Last previous edition approved in 2011 as E2489 – 11. DOI: 10.1520/E2489-16.

[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] The last approved version of this historical standard is referenced on www.astm.org.

[4] Horwitz, W., "Protocol for the Design, Conduct and Interpretation of Collaborative Studies," *Pure and Applied Chemistry*, Vol 60, No. 6, 1988, pp. 855–864.

3.1.6 *outlying observation, n*—observation that appears to deviate markedly in value from other members of the sample in which it appears. **E178**

3.1.7 *proficiency testing, n*—determination of laboratory testing performance by means of interlaboratory comparisons.

3.1.8 *repeatability standard deviation (S_r), n*—standard deviation of test results obtained under repeatability conditions. **E177**

3.1.9 *reproducibility standard deviation (S_R), n*—standard deviation of test results obtained under reproducibility conditions. **E177**

3.2 *Definitions of Terms Specific to This Standard:*

3.2.1 *hinge (upper or lower), n*—median of the upper or lower half of a set of data when the data is arranged in order of size.

3.2.1.1 *Discussion*—When there is an odd number of items in the data set, the middle value is included in both the upper and lower halves. The upper hinge is an estimate of the 75th percentile; the lower hinge is an estimate of the 25th percentile.

3.2.2 *inner fence (upper or lower), n*—value equal to the upper or lower hinge of a data set plus (upper) or minus (lower) 1.5 times the difference between upper and lower hinges.

3.2.3 *interquartile range, n*—distance between the upper and lower hinges of a data set.

3.2.4 *outer fence (upper or lower), n*—value equal to the upper or lower hinge of a data set plus (upper) or minus (lower) three times the difference between upper and lower hinges.

4. Summary of Practice

4.1 This practice describes methods of displaying interlaboratory data that visually show individual laboratory results.

4.2 The methods described in this practice can be applied to large and small sample populations from any distribution expected to have a general mound shape. It is recommended that in cases in which it is suspected that the data may be highly unsymmetrical or very unusual in some other manner a statistician should be consulted regarding the applicability of the analysis method.

4.2.1 The median is used as the "consensus" value of the measured test property.

4.2.2 The interquartile range (IQR) is used as the basis for estimating the spread in the data. Because the median and the interquartile range are not affected by the magnitude of extreme values of a data set, the analysis approach presented in this practice effectively eliminates the need to identify outlying observations (outliers).

4.3 Laboratory results are categorized according to how far the results lie outside of the interquartile range.

4.4 The upper and lower ends of the interquartile range are referred to as the hinges. The limits for categorizing laboratory results lying outside of the interquartile range are determined by multiplying the extent of the interquartile range by the fixed factors of 1.5 and 3.0. The upper and lower limits lying a distance of 1.5 times the range of the IQR beyond the hinges are referred to as the inner fences. The upper and lower limits for results lying at 3.0 times the range of the IQR beyond the hinges are referred to as the outer fences.

4.5 Guidance is provided for proficiency testing programs wishing to establish additional limits (or fences). The user is directed to Guide E1301 for additional guidance.

4.6 When using the methods in this practice, the number of participating laboratories should be at least ten. Since the degree of confidence is lower for analyses performed on small sample populations, caution should be used in applying statistics obtained from small sample populations.

4.7 When possible, it is generally desirable to have 30 or more participants when estimating the precision of test methods.

4.8 Estimates of the repeatability standard deviation and the reproducibility standard deviation are determined by dividing the interquartile ranges of appropriate data sets by a factor of 1.35.

4.8.1 The number 1.35 used in determining the repeatability and reproducibility standard deviations is based on an assumption of similarity to a normal distribution. Therefore, the estimate of the standard deviation using the methods described in this practice may not supply the desired accuracy if the distribution differs too much from the general shape of a normal curve. It is beyond the scope of this practice to describe procedures for determining when the analysis methods described in this practice are not applicable.

5. Significance and Use

5.1 This practice is specifically designed to describe simple robust statistical methods for use in proficiency testing programs.

5.2 Proficiency testing programs can use the methods in this practice for the purpose of comparing testing results obtained from a group of participating laboratories. The laboratory comparisons can then be used for evaluation of individual laboratory performance.

5.3 In addition, the data obtained in proficiency testing programs may contain information regarding repeatability (within-lab) and reproducibility (between-lab) testing variation. Repeatability information is possible only if the program uses more than one sample. See Method B. Proficiency testing programs often have a greater number of participants than might be available for conducting an interlaboratory study to determine the precision of a test method (such as described in Practice E691). Precision estimates obtained for the larger number of participants in a proficiency testing program, along with the corresponding wider variation of test conditions, can provide useful information to standards developers regarding the precision of test results that can be expected for a test method when in actual use in the general testing community.

5.4 To estimate the precision of a test method, the participants must use the same test method to obtain their test results, and testing must be performed under the conditions required for repeatability and reproducibility. The precision estimates are applicable to the property levels and material types included in the testing program. The precision of a test method

may vary considerably for different material types and at different property levels.

5.5 This practice may be useful to proficiency testing program administrators and provides examples of statistical methods along with explanations of some of the advantages of the suggested methods of analysis. The analyses resulting from the application of methods described in this practice may be used by laboratories as part of their quality control procedures, accrediting bodies to assist in the evaluation of laboratory performance, and ASTM International technical committees (and other organizations charged with the task of writing, maintaining, or improving test methods) to obtain information regarding reproducibility and repeatability.

5.6 There are many types of proficiency testing programs in existence and many methods exist for analyzing the data resulting from the interlaboratory testing. It is not the intention of this practice to call into question the integrity of programs using other methods of analysis. Testing programs using replicate testing of one or more samples (each laboratory submits two or more results for each sample) are directed to Practice E691 or other practices for the description of a method of analysis that may be more suitable to that type of program.

6. Analysis of a One-Sample Program (Method A)

6.1 *Display of Data:*

6.1.1 When possible, display the data in a table to show the actual results submitted by each laboratory. This may not be practical if the number of participants is too large.

6.1.1.1 To assist in maintaining confidentiality, give each laboratory an identification number if one does not already exist.

6.1.1.2 List the laboratory results in increasing order by laboratory identification number to make it easy to locate the results for a particular laboratory. See Table 1.

6.1.2 Sort the laboratory results in decreasing order by test result to show the range and distribution of the test results. See Table 2. Besides the laboratory identification number and corresponding test results, Table 2 contains columns of additional information that will be explained in the following sections of this practice.

6.1.3 Display the data in a dot diagram to show the location of each laboratory's test result in the distribution of all test results. For each test result, plot occurrence number of that test result value versus the value of the test result. As points are plotted from the top of Table 2 to the bottom, the first time a test value occurs assign it an occurrence of "one." The next time that test result value occurs, assign it an occurrence of "two." If the test result value appears a third time, assign it an occurrence of "three" and so forth. If a test result value appears three times in the data, plot the test result value three times, once with an occurrence of "one," once with an occurrence of "two." and once with an occurrence of "three." The consequence is that each laboratory's test result will be plotted as an individual dot and no dots will be concealed by being plotted on top of one another.

6.1.3.1 Fig. 1 shows the dot diagram for the data in Table 2. There are no repeat values in the test results, so Column 3 of Table 2 shows that the number of occurrences is "one" for each test result and the dots in Fig. 1 appear in a single horizontal row. The dot diagram in Fig. 1 also shows that the test result for Laboratory 5, at (2.75, 1), is slightly removed from the rest of the data. The test result for Laboratory 27, at (4.89, 1), is farther removed.

TABLE 1 Original Data for a One-Sample Program

Lab	Test Result
1	1.22
2	1.62
3	1.82
4	0.60
5	2.75
6	1.55
7	1.17
8	1.76
9	1.35
10	1.18
11	1.19
12	1.71
13	2.03
14	1.10
15	1.84
16	1.39
17	1.13
18	1.66
19	1.28
20	1.24
21	0.69
22	1.54
23	1.43
24	0.84
25	0.98
26	1.97
27	4.89
28	1.85
29	1.09
30	1.07

6.1.3.2 A dot diagram with a different appearance can be obtained by classifying the results into multiple contiguous size classes such that each class contains a portion of the data, but together, the classes cover the entire data range. Table 3 shows the number of occurrences in each size class when the range of each class is 0.10. When the numbers of occurrences in each size class are plotted versus the corresponding values of the lower ends of each size class (see Fig. 2), the display has the advantage of being more compact, and it is more apparent how test results are clustered. The dot diagram in Fig. 2 still shows that the test result for Laboratory 5 is slightly removed from the rest of the data and that the test result for Laboratory 27 is farther removed.

6.1.3.3 Other ranges for the size classes are permitted to be used to classify the test results. For example, each size class could have a range of 0.20 or 0.05. The corresponding dot diagrams will each have a different appearance.

6.1.3.4 The range of the size classes used for grouping the laboratory test results should be chosen carefully to show as much information (regarding individual laboratory test results and the overall distribution of the test results) as possible in the dot diagram. One consideration should be the number of test results that must be plotted. Generally, it is desirable to limit the number of classes to be plotted along the x-axis of the dot diagram. For larger data sets, the range of each of the classes must be wider to contain a larger number of test results. Another consideration should be the overall range of the test

TABLE 2 Data in Descending Order for One-Sample Program

Count of Labs	Lab	Test Result	Number of Occurrences	Category
	27	4.89	1	Extremely Unusual
	5	2.75	1	Unusual
	13	2.03	1	Typical
	26	1.97	1	Typical
	28	1.85	1	Typical
	15	1.84	1	Typical
	3	1.82	1	Typical
8th from Top	8	1.76	1	Typical
	12	1.71	1	Typical
	18	1.66	1	Typical
	2	1.62	1	Typical
	6	1.55	1	Typical
	22	1.54	1	Typical
	23	1.43	1	Typical
15th from Top	16	1.39	1	Typical
16th from Top	9	1.35	1	Typical
	19	1.28	1	Typical
	20	1.24	1	Typical
	1	1.22	1	Typical
	11	1.19	1	Typical
	10	1.18	1	Typical
	7	1.17	1	Typical
8th from Bottom	17	1.13	1	Typical
	14	1.10	1	Typical
	29	1.09	1	Typical
	30	1.07	1	Typical
	25	0.98	1	Typical
	24	0.84	1	Typical
	21	0.69	1	Typical
	4	0.60	1	Typical

Shown Below Is Determination of "Fences" for Data Above

Median of All Test Results = 1.37
Upper hinge (Median of Top Half) = 1.76
Lower Hinge (Median of Bottom Half) = 1.13
Interquartile Range (IQR) = (1.76 – 1.13) = 0.63

(3 × IQR) = 1.89
Outer Fence (Upper) = (1.76 + 1.89) = 3.65
Outer Fence (Lower) = (1.13 – 1.89) = –0.76

(1.5 × IQR) = 0.945
Inner Fence (Upper) = (1.76 + 0.945) = 2.705
Inner Fence (Lower) = (1.13 – 0.945) = 0.185

Reproducibility Standard Deviation = (IQR / 1.35) = 0.467

results in the data set. All size classes should have the same width and each size class must be sufficiently wide to limit the number of classes to be plotted along the *x*-axis of the dot diagram.

6.1.3.5 Various computer software programs can be used to generate similar types of diagrams. When other types of diagrams are used, it is generally preferable to choose one in which each individual laboratory's result is displayed as a single point on the diagram. For example, Fig. 2 is similar in appearance to a histogram, but a typical histogram does not show individual data points. Another example is a stem-and-leaf plot.

6.2 *Steps for Evaluating Laboratory Performance:*

6.2.1 Visually examine the dot plot (or graphic of the data) to confirm that the distribution is approximately mound shaped and unimodal. If either condition is not met, the analysis prescribed may not be appropriate. See 4.2.

6.2.2 The steps for evaluating a laboratory's performance are to determine the median and interquartile range (IQR), locate the inner and outer fences, and then categorize the laboratories according to where their results lie relative to the fences.

6.2.3 The method for determining the median depends on whether there is an odd or even number of results in the data set.

6.2.3.1 Sort the data set into ascending or descending order. If there is an odd number of results in the data set, after the results are placed in ascending or decreasing order, the median is the middle number of the data set. For example, consider the five results in the data set 9, 1, 5, 4, 5. When placed in ascending order, the result is 1, 4, 5, 5, 9. The middle number, or median, is the underlined 5. It does not matter that one of the numbers is repeated.

6.2.3.2 If there is an even number of results in the data set, after the results are placed in ascending or descending order, the median is the average of the middle two numbers in the data set. For example, consider the eight results in the data set 2, 8, 5, 11, 4, 6, 9, 4. When placed in ascending order, the result is 2, 4, 4, 5, 6, 8, 9, 11. The middle two numbers are 5 and 6. The average is (5 + 6)/2 or 5.5, so the median is 5.5.

6.2.4 The method for determining the interquartile range is to determine the middle number (or median) of the top and bottom halves of the data set.

6.2.4.1 If there are an odd number of results in the data set, the median of the entire data set is included in both halves. For example, consider again the data set 1, 4, 5, 5, 9. The underlined 5 is included in both halves. So, the middle number (or median) of the top half of the data set, 5, 5, 9, is 5. The median of the top half of the data set is referred to as the upper hinge. The middle number (or median) of the bottom half of the data set, 1, 4, 5, is 4. The median of the bottom half of the data set is referred to as the lower hinge.

6.2.4.2 The IQR is the range from the upper hinge (the median of the top half of the data set) to the lower hinge (the median of the bottom half of the data set).

6.2.4.3 Since the IQR of the data set 1, 4, 5, 5, 9, is the range from the upper hinge, 5, to the lower hinge, 4, the IQR is (5 – 4), or 1.

6.2.4.4 If there is an even number of results in the data set, the data set is simply divided into a top half and a bottom half, each containing an equal number of test results. For example, consider the data set 2, 4, 4, 5, 6, 8, 9, 11. The top half contains 6, 8, 9, 11 and the median (or upper hinge) is the average of 8 and 9, or 8.5. The bottom half contains 2, 4, 4, 5 and the median (or lower hinge) is the average of 4 and 4, or 4.

6.2.4.5 Since the IQR of the data set 2, 4, 4, 5, 6, 8, 9, 11, is the range from the upper hinge, 8.5, to the lower hinge, 4, the IQR is (8.5 – 4), or 4.5.

6.2.5 Once the IQR is determined, the outer fence is located three times the range of the IQR, (3 × IQR), beyond the upper and lower hinges. See Fig. 3 and guidance provided in 4.8.1.

$$\text{Outer Fence (Upper)} = (\text{Upper Hinge}) + (3 \times \text{IQR}) \quad (1)$$

$$\text{Outer Fence (Lower)} = (\text{Lower Hinge}) - (3 \times \text{IQR}) \quad (2)$$

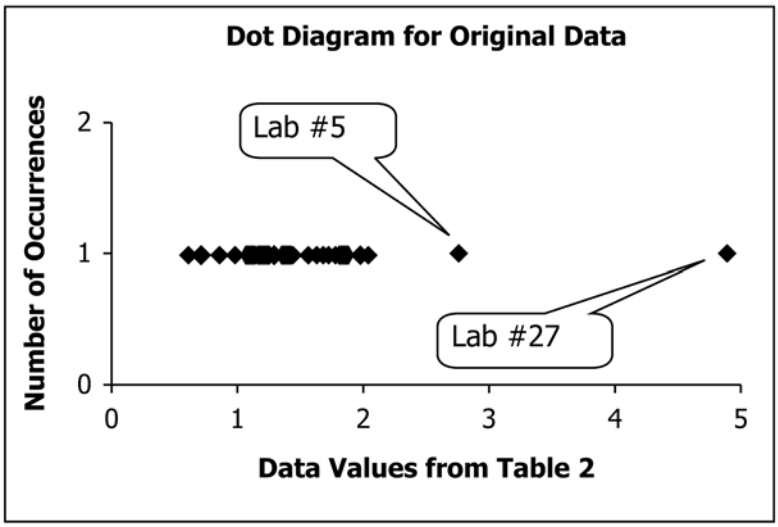

FIG. 1 Dot Diagram for Original Data

TABLE 3 Data Classified by Tenths

Lab	Test Result	Size Class Range			Number of Occurrences
		Lower End		Upper End	
27	4.89	4.80	≤ X <	4.90	1
5	2.75	2.70	≤ X <	2.80	1
13	2.03	2.00	≤ X <	2.10	1
26	1.97	1.90	≤ X <	2.00	1
28	1.85	1.80	≤ X <	1.90	1
15	1.84	1.80	≤ X <	1.90	2
3	1.82	1.80	≤ X <	1.90	3
8	1.76	1.70	≤ X <	1.80	1
12	1.71	1.70	≤ X <	1.80	2
18	1.66	1.60	≤ X <	1.70	1
2	1.62	1.60	≤ X <	1.70	2
6	1.55	1.50	≤ X <	1.60	1
22	1.54	1.50	≤ X <	1.60	2
23	1.43	1.40	≤ X <	1.50	1
16	1.39	1.30	≤ X <	1.40	1
9	1.35	1.30	≤ X <	1.40	2
19	1.28	1.20	≤ X <	1.30	1
20	1.24	1.20	≤ X <	1.30	2
1	1.22	1.20	≤ X <	1.30	3
11	1.19	1.10	≤ X <	1.20	1
10	1.18	1.10	≤ X <	1.20	2
7	1.17	1.10	≤ X <	1.20	3
17	1.13	1.10	≤ X <	1.20	4
14	1.10	1.10	≤ X <	1.20	5
29	1.09	1.00	≤ X <	1.10	1
30	1.07	1.00	≤ X <	1.10	2
25	0.98	0.90	≤ X <	1.00	1
24	0.84	0.80	≤ X <	0.90	1
21	0.69	0.60	≤ X <	0.70	1
4	0.60	0.60	≤ X <	0.70	2

6.2.5.1 For testing performed in strict accordance with a testing protocol, laboratory test results beyond the outer fence have an extremely low likelihood of occurrence. Laboratory results occurring beyond the outer fence are categorized as "extremely unusual."

6.2.6 The inner fence is located 1.5 times the range of the IQR (1.5 × IQR) beyond the upper and lower hinges.

$$\text{Inner Fence (Upper)} = (\text{Upper Hinge}) + (1.5 \times \text{IQR}) \quad (3)$$

$$\text{Inner Fence (Lower)} = (\text{Lower Hinge}) - (1.5 \times \text{IQR}) \quad (4)$$

6.2.6.1 Laboratory test results lying beyond the inner fence, but within the outer fence, have a low probability of occurrence when testing is properly performed in accordance with the prescribed testing protocol. Laboratory results occurring beyond the inner fence but within the outer fence are categorized as "unusual."

6.2.7 Most of the test results will fall within the inner fence. Laboratory test results falling at or within the inner fence are categorized as "typical."

6.2.8 If desired, other limits or fences can be used. Table 4 suggests several intervals that could be used to establish other fences and gives the probabilities for results lying outside of each of the intervals listed in the table.

6.3 *Example for Evaluating Laboratory Performance Using the Data in Table 2:*

6.3.1 Table 2 shows test results for 30 laboratories, an even number of results, in descending order by test result. The median of the data set is the average of the results for the 15th and 16th laboratories from the top of the table. The 15th and 16th laboratories are #16 and #9. The median is the average of the results, (1.39 + 1.35)/2 or 1.37. See the analysis at the bottom of Table 2.

6.3.2 There are 15 results in the top half of the data in Table 2 and 15 in the bottom half. The middle (or median) value of the top half is the eighth test result from the top, 1.76. This value, 1.76, is referred to as the upper hinge. The middle (or median) of the bottom half is the eighth result from the bottom, 1.13. This value, 1.13, is referred to as the lower hinge.

6.3.3 The IQR is the range from the upper hinge (median of the top half) to the lower hinge (median of the bottom half), (1.76 – 1.13), or 0.63.

6.3.4 Since the outer fence is located three times the IQR beyond the hinges, the outer fence for the upper end of the data set is located at [1.76 + (3 × IQR)], or [1.76 + (3 × 0.63)], or 3.65. The outer fence for the lower end of the data set is located at [1.13 – (3 × 0.63)], or –0.76.

6.3.5 Test results greater than 3.65 or less than –0.76 are categorized as "extremely unusual." Only one test result, 4.89,

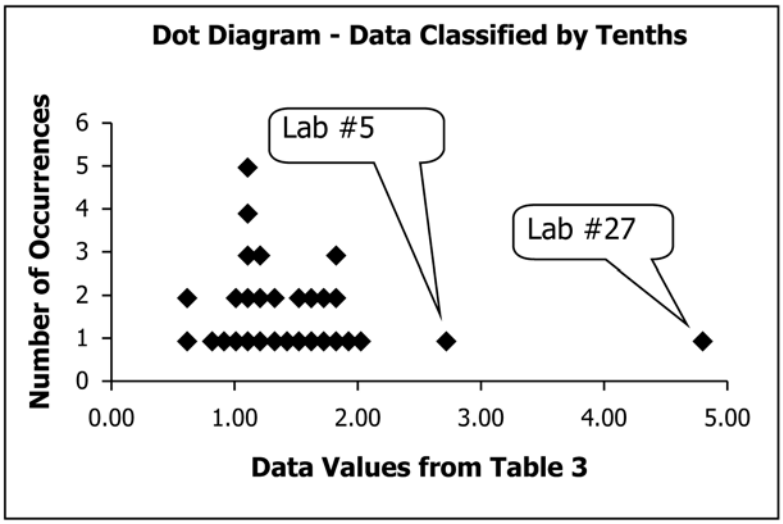

FIG. 2 Dot Diagram—Data Classified by Tenths

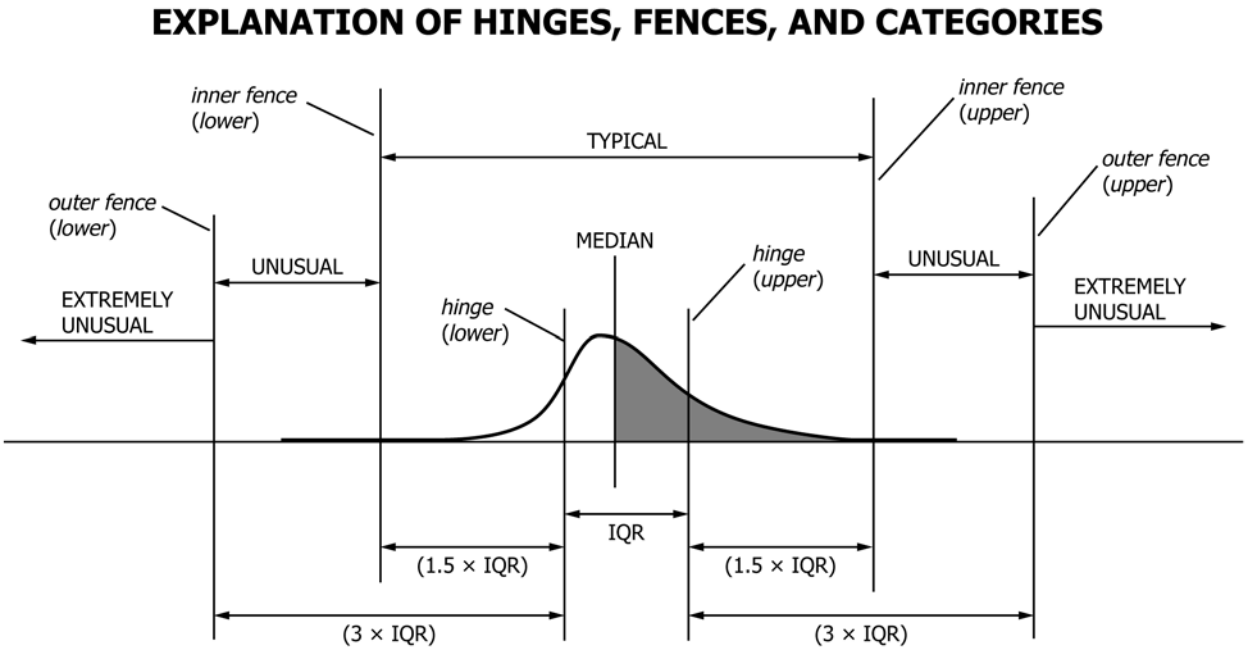

FIG. 3 Explanation of Hinges, Fences, and Categories

is beyond the outer fence. That test result, for laboratory #27, is greater than 3.65 and is categorized as "extremely unusual." See Column 5 of Table 2. There are no test results below –0.76.

6.3.6 The inner fence is located 1.5 times the IQR beyond the hinges. The inner fence for the upper end of the data set is located at [1.76 + (1.5 × 0.63)], or 2.705. The inner fence for the lower end of the data set is located at [1.13 – (1.5 × 0.63)], or 0.185.

6.3.7 On the upper end of the data, test results lying beyond the inner fence, but not beyond the outer fence (greater than 2.705, but less than or equal to 3.65) are categorized as "unusual." Test result 2.75, for laboratory #5, falls into that range and is categorized as "unusual." Correspondingly, at the lower end of the data set, test results less than 0.185 and greater than or equal to –0.76 are also categorized as "unusual." Table 2 contains no test results in that range.

TABLE 4 Alternative Intervals for Fences

Interval Beyond the Upper and Lower Hinges	Approx. Number of Standard Deviations from the Consensus Value (Median)[A]	Approx. Two-Tailed Probability for Results Outside of the Interval[A]	Suggested Descriptive Label for Results Occurring Outside of the Interval
1.0 IQR	2	0.04	typical
1.5 IQR	2.7	0.007	unusual
2.0 IQR	3.37	0.0008	very unusual
3.0 IQR	4.725	0.000 002	extremely unusual

[A] The number of standard deviations from the consensus value and the probabilities for being outside of the intervals are based on the assumption of a normal distribution. The probabilities may vary for distributions that cannot be approximated by a normal distribution.

6.3.8 Test results lying within the inner fence (from 0.185 to 2.705 inclusive) are categorized as "typical," as shown in Table 2, Column 5.

6.4 *Estimating the Reproducibility Standard Deviation:*

6.4.1 In a one-sample program, each test result in the data set contains the random error (within-laboratory error) and systematic error (bias) components of testing variation associated with the laboratory providing the test result. The sample standard deviation of the data set of test results describes the spread of the distribution of test results and is used to estimate the total between-laboratory variation or reproducibility standard deviation.

6.4.2 The IQR also describes the spread of the distribution of test results in the data set of test results. Since the IQR and the sample standard deviation both describe the spread of the test results, the two are related. Determine an estimate of the sample standard deviation, or reproducibility standard deviation, by dividing the IQR by a fixed factor, 1.35 as follows:

$$\text{Reproducibility Standard Deviation, } S_R = \text{IQR}/1.35 \qquad (5)$$

6.4.3 For example, the reproducibility standard deviation estimate for the data in Table 2 is (IQR/1.35), or (0.63/1.35), or 0.467. See the analysis at the bottom of Table 2.

7. Analysis of a Two-Sample Program (Method B)

7.1 In two-sample programs, samples are issued to participant laboratories in pairs, with each laboratory providing a single test result for each sample.

7.2 *Display of Data for a Two-Sample Program:*

7.2.1 When possible, display the data in a table to show the actual results submitted by each laboratory. This may not be practical if the number of participants is too large.

7.2.1.1 To assist in maintaining confidentiality, give each laboratory an identification number if one does not already exist. Call the first sample "X" and the second sample "Y."

7.2.1.2 List the laboratory results in increasing order by laboratory number to make it easy to locate the results for a particular laboratory. See Table 5. (The data for Sample X in Table 5 is the same as was used in Table 2 for the one-sample program.)

7.2.2 Display the results of each participating laboratory on a scatter diagram by plotting the results of the second sample (Sample Y) versus the results of the first sample (Sample X). Plot each laboratory's results as a coordinate pair, with the result for the first sample (Sample X) plotted on the *x*-axis and the results for the second sample (Sample Y) plotted on the *y*-axis. This display provides an obvious view of the distribution of the test results.

7.2.3 For the usual circumstances in which the within-laboratory test variation is less than the test variation between laboratories, the plotted points typically appear in the approximate shape of an elliptical cloud with the major axis of the ellipse along a diagonal line having a slope of approximately 45° from the horizontal axis. The two median values (one median for each of the two samples, X and Y) represent the "consensus" values of the test property for each sample, respectively. A coordinate pair consisting of the median values for each sample would appear near the center of the ellipse.

TABLE 5 Original Data for a Two-Sample Program

Lab	Sample X Test Result	Sample Y Test Result
1	1.22	1.26
2	1.62	1.91
3	1.82	1.20
4	0.60	1.00
5	2.75	2.41
6	1.55	1.26
7	1.17	1.57
8	1.76	1.28
9	1.35	0.93
10	1.18	1.21
11	1.19	1.26
12	1.71	0.42
13	2.03	2.21
14	1.10	1.33
15	1.84	1.81
16	1.39	1.12
17	1.13	1.08
18	1.66	1.45
19	1.28	1.20
20	1.24	0.71
21	0.69	0.77
22	1.54	1.39
23	1.43	1.52
24	0.84	1.27
25	0.98	0.99
26	1.97	2.24
27	4.89	5.28
28	1.85	1.78
29	1.09	1.20
30	1.07	1.24

7.2.4 The distance from the center of the ellipse to a point on the scatter diagram is related to how much a laboratory's results differ from the "consensus," or median, values.

7.2.5 If a laboratory obtains similar results on both samples, their results will be plotted as a point along the 45° diagonal (near the major axis of the ellipse). The greater the difference between a laboratory's two results, the further the point representing that laboratory's test results will appear from the diagonal. Therefore, the distance of a laboratory's results from the diagonal is related to a laboratory's within-laboratory variation.

7.2.6 Similarly, the distance from the center of the ellipse to the orthogonal projection of the point representing a laboratory's test results onto the 45° diagonal is related to the laboratory's between-laboratory variation.

7.2.7 Information can be derived from the general shape of the cloud of points on the scatter diagram. If the between-laboratory variation is large relative to the within-laboratory variation, the elliptical cloud will appear long and slender. For a well-defined test method in which the equipment is very similar from laboratory to laboratory and there is very little variation between technicians, the between-laboratory variation may be the same as the within-laboratory variation, and the general shape of the cloud of points on the scatter diagram will be circular.

7.2.8 Information regarding an individual laboratory's results can be inferred from the location on the scatter diagram of the point representing that laboratory's results. Points lying in the upper right quadrant indicate that the laboratory's results on

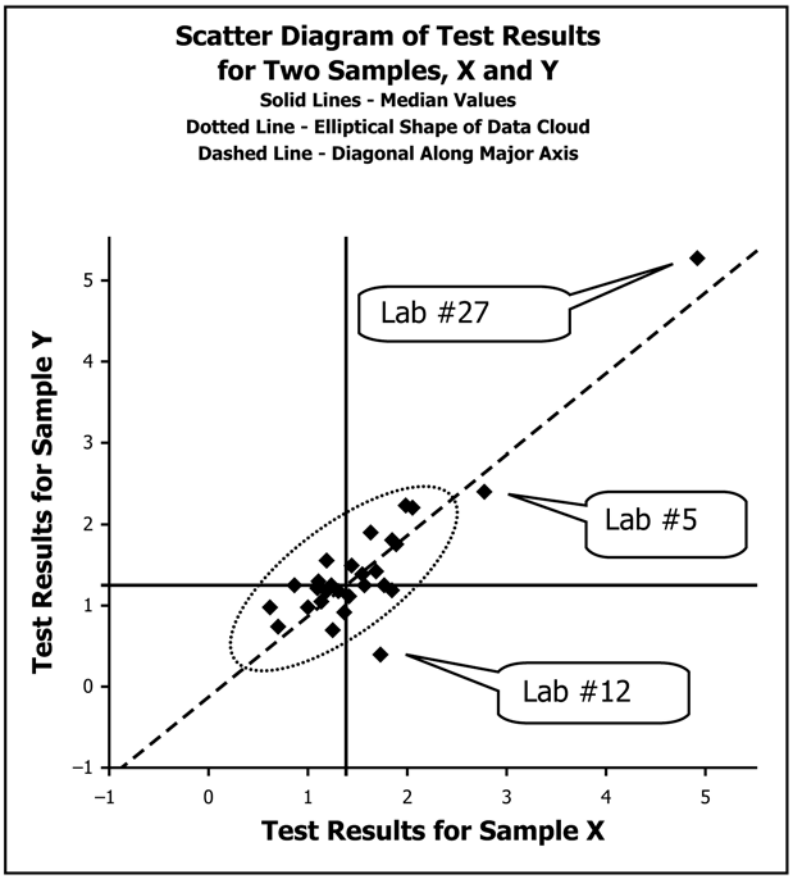

FIG. 4 Scatter Diagram of Test Results for Two Samples, X and Y

both samples were above the medians (or consensus values). Points lying in the lower left quadrant indicate the results were lower than median (or consensus value) for each of the two samples. Points lying in the lower right quadrant indicate that the result for the sample plotted on the *x*-axis was higher than the median and that the result for the sample plotted on the *y*-axis was lower than the median. Points lying in the upper left quadrant indicate that the result for the sample plotted on the *y*-axis was higher than the median and the result for the sample plotted on the *x*-axis was lower than the median.

7.2.9 Fig. 4 shows one point (4.89, 5.28) that appears particularly far from the center of the ellipse along the diagonal when compared to other laboratories. The point located at (2.75, 2.41) also appears to be farther along the diagonal than most of the other points. In addition, there is one point (1.71, 0.42) that appears to be suspiciously farther from the diagonal than the remainder of the laboratories. These three points correspond to the test results for laboratories #27, #5, and #12.

7.2.10 Since the slope of the major axis of the elliptical cloud of test results is related to the spread of the test results along the *y*-axis relative to the spread along the *x*-axis, a slope appearing to differ from 45° is an indication that the variation in test results is different for each of the two samples. The spread of the test results along the *y*-axis can be estimated by the reproducibility standard deviation of "Sample Y," S_{RY}, and the spread of the test results along the *x*-axis can be estimated by the reproducibility standard deviation of "Sample X," S_{RX}. The slope of the major axis of the elliptical cloud is then related to the ratio of S_{RY}/S_{RX}. If the ratio of S_{RY}/S_{RX} is greater than 1.1 or less than 0.9 (this corresponds to a slope of 45 ± 3°), the two samples may not be similar enough to use the statistical methods described in Method B of this practice. Specifically, if the slope falls outside of that range, it may not be appropriate to consider repeatability (within-laboratory variation) when using Method B and it may not be appropriate to apply a "pooled" estimate of reproducibility standard deviation (between-laboratory variation) to both samples. It may be more appropriate to apply Method A separately to each of the two samples. See 7.8.4.

7.3 *Steps for Evaluating Intra(Within)-Laboratory Performance:*

7.3.1 A single laboratory's test result for Sample X, X_i, contains variation from the "consensus" value because of random error (or within-laboratory error) and systematic error (or bias). The same is true of the laboratory's test result for Sample Y, Y_i. Assuming that the systematic error is the same for both samples, the difference $(X_i - Y_i)$ will contain only random error because the systematic error will be subtracted out of the resulting quantity. Subtracting the quantity $(X_{med} - Y_{med})$ from the quantity $(X_i - Y_i)$ will remove any difference that may appear between the medians for Samples X and Y, where X_{med} is the median for Sample X and Y_{med} is the median for Sample Y. (If the materials used for Samples X and Y are

TABLE 6 Random Error Quantities in Descending Order for Two-Sample Program

Count of Labs	Lab	Sample X Test Result	Sample Y Test Result	Random Error Quantities $(X - Y) - (X_{med} - Y_{med})$	Within-Lab Category
	12	1.71	0.42	1.18	Unusual
	3	1.82	1.20	0.51	Typical
	20	1.24	0.71	0.42	Typical
	8	1.76	1.28	0.37	Typical
	9	1.35	0.93	0.31	Typical
	5	2.75	2.41	0.23	Typical
	6	1.55	1.26	0.18	Typical
8th from Top	16	1.39	1.12	0.16	Typical
	18	1.66	1.45	0.10	Typical
	22	1.54	1.39	0.04	Typical
	19	1.28	1.20	–0.03	Typical
	28	1.85	1.78	–0.04	Typical
	17	1.13	1.08	–0.06	Typical
	15	1.84	1.81	–0.08	Typical
15th from Top	25	0.98	0.99	–0.12	Typical
16th from Top	10	1.18	1.21	–0.14	Typical
	1	1.22	1.26	–0.15	Typical
	11	1.19	1.26	–0.18	Typical
	21	0.69	0.77	–0.19	Typical
	23	1.43	1.52	–0.20	Typical
	29	1.09	1.20	–0.22	Typical
	30	1.07	1.24	–0.28	Typical
8th from Bottom	13	2.03	2.21	–0.29	Typical
	14	1.10	1.33	–0.34	Typical
	26	1.97	2.24	–0.38	Typical
	2	1.62	1.91	–0.40	Typical
	27	4.89	5.28	–0.50	Typical
	4	0.60	1.00	–0.51	Typical
	7	1.17	1.57	–0.51	Typical
	24	0.84	1.27	–0.54	Typical

Shown Below Is Determination of Inner and Outer Fences for Random Error Quantities Data Above

Median for Sample X, (X_{med}) = 1.37
Median for Sample Y, (Y_{med}) = 1.26

Median of Random Error Quantities = –0.13
Median of Top Half of Random Error Quantities = 0.16 (Upper Hinge)
Median of Bottom Half of Random Error Quantities = –0.29 (Lower Hinge)
Interquartile Range (IQRr) of Random Error Quantities = 0.45

(3 × IQRr) = 1.35
Outer Fence (Upper) = (0.16 + 1.35) = 1.51
Outer Fence (Lower) = (–0.29 – 1 .35) = –1.64

(1.5 × IQRr) = 0.675
Inner Fence (Upper) = (0.16 + 0.675) = 0.835
Inner Fence (Lower) = (–0.29 – 0.675) = –0.965

Repeatability Standard Deviation = (IQRr / 1.35)/sqrt(2) = 0.236

identical, then X_{med} and Y_{med} should be close in value and $(X_{med} - Y_{med})$ should be small or zero.) Therefore, the quantity $[(X_i - Y_i) - (X_{med} - Y_{med})]$ contains only the laboratory's random error, and that quantity can be used to compare the within-laboratory performance of one laboratory to the within-laboratory performance of other participants.

7.3.2 The steps for evaluating a laboratory's within-laboratory performance are to determine the random error quantity $[(X_i - Y_i) - (X_{med} - Y_{med})]$ for each laboratory, determine the interquartile range of the random error quantities, find the outer and inner fences, and then rate categorize the laboratories by how far their results lie outside of the interquartile range relative to the inner and outer fences established for the data set containing the random error quantities.

7.3.3 Determine the IQR for the random error quantities (IQR_r) in the same manner as explained for finding the IQR for one-sample programs in 6.2.3.

7.3.4 Once the IQR_r is determined, the outer and inner fences and laboratory performance evaluation categories are determined in the same manner as explained for the one-sample program in 6.2.5 – 6.2.8.

7.4 *Example for Evaluating Intra(Within)-Laboratory Performance for Two-Sample Programs Using Data in Table 5:*

7.4.1 Column 5 of Table 6 shows the random error quantities calculated for each of the 30 laboratories listed in Table 5. To make it possible to determine the IQR_r by observation, the quantities in Column 5 are listed in descending order.

7.4.2 The medians for Samples X and Y are needed to calculate the random error quantities in Column 5 of Table 6. Since the data in Column 3 of Table 6 is identical to the data in Column 3 of Table 2 (for the one-sample program), the median for Sample X, X_{med}, is 1.37 as determined in 6.3.1. In a similar manner, the median of Sample Y, Y_{med}, is determined to be 1.26. Note that the test results for Samples X and Y in Columns 3 and 4, respectively, are not in descending order, so the values for X_{med} and Y_{med} are not obvious by looking at the table.

7.4.3 Column 5 of Table 6 shows the values for the quantity $[(X_i - Y_i) - (X_{med} - Y_{med})]$ for 30 laboratories, an even number of results, in descending order. The median of the values in Column 5 is the average of the values for the 15th and 16th laboratories from the top of the table. The 15th and 16th laboratories are #25 and #10. The median is the average of the values (–0.12 + –0.14)/2 or –0.13. See the analysis at the bottom of Table 6.

7.4.4 There are 15 results in the top half of the data in Column 5 and 15 in the bottom half. The middle (or median) value of the top half (the upper hinge) is the eighth test result from the top, 0.16. The middle (or median) of the bottom half (the lower hinge) is the eighth result from the bottom, –0.29.

7.4.5 The IQR_r is the range from the upper hinge to the lower hinge [0.16 – (–0.29)] or 0.45.

7.4.6 Since the outer fences are located three times the IQR_r beyond the hinges (see Fig. 3 and Eq 1 and 2), the outer fence for the upper end of the data set is located at [0.16 + (3 × IQR_r)] or [0.16 + (3 × 0.45)] or 1.51. The outer fence for the lower end of the data set is located at [-0.29 – (3 × 0.45)] or –1.64.

7.4.7 Quantities greater than 1.51 or less than –1.64 are categorized as "extremely unusual." None of the quantities in Column 5 lie in either range, so none of the quantities are categorized as "extremely unusual."

7.4.8 The inner fences are located 1.5 times the IQR_r beyond the hinges (see Eq 3 and 4). The inner fence for the upper end of the data set is located at [0.16 + (1.5 × 0.45)] or 0.835. The inner fence for the lower end of the data set is located at [–0.29 – (1.5 × 0.45)] or –0.965.

7.4.9 Test results greater than the upper inner fence, 0.835, and less than or equal to the upper outer fence, 1.51, are categorized as "unusual." The quantity 1.18, for Laboratory #12, falls into that range and is categorized as "unusual." Correspondingly, at the lower end of the data set, test results less than –0.965 and greater than or equal to –1.64 are also categorized as "unusual." Column 5 contains no quantities in that range.

7.4.10 Test results falling at or within the inner fences (from –0.965 to 0.835, inclusive) are categorized as "typical," as shown in Table 6.

7.5 *Steps for Evaluating Inter(Between)-Laboratory Performance*—Evaluate each of the two samples, X and Y, for between-laboratory performance separately using the same procedure as for the single-sample program described in 6.2.5 – 6.2.8.

7.6 *Example for Evaluating Inter(Between)-Laboratory Performance for Two-Sample Programs Using the Data in Table 5:*

7.6.1 Since the sample data set used for Sample X in Table 5 for the two-sample program is the same as the data in Table 1 for the one-sample program, the between-laboratory evaluation categories for the test results for Sample X are identical to those shown in Table 2 for the one-sample program. Also, see Table 7, Column 3. Note that the Sample X test results in Table 7 are not in descending order. The results in Table 7 are in descending order for Column 5 to show how to determine evaluation categories for Sample Y.

7.6.2 The inter(between)-laboratory evaluation categories for the test results for Sample Y are based on the inner and outer fences determined at the bottom of Table 7. Again, as was done for Sample X, the categories are assigned in the same manner as used for the one-sample program.

7.6.3 Test results beyond the outer fences, 2.92 for the upper outer fence and –0.23 for the lower outer fence, are categorized as "extremely unusual." See Column 6 of Table 7.

7.6.4 Test results lying beyond the inner fences, 2.245 for the upper inner fence and 0.445 for the lower inner fence, but within the outer fences, are categorized as "unusual."

7.6.5 Test results falling at or within the inner fences are categorized as "typical."

7.7 Table 8 shows a summary of the ratings for Samples X and Y.

7.7.1 The results of Laboratories #5, #12, and #27 are of particular interest. These are the laboratories whose results were noted in Fig. 1, Fig. 2, and Fig. 4 as being different from the results of the other laboratories.

7.7.2 Laboratory #27 reported test results of 4.89 for Sample X and 5.28 for Sample Y. Even though the laboratory's random error (within-laboratory variation) is categorized as "typical" (see Column 7 of Table 8), the results reported for Samples X and Y differ greatly from the results of other laboratories and the results are categorized as "extremely unusual" for both samples.

7.7.3 Laboratory #12 reported 1.71 for Sample X and the result is categorized as "typical." However, the laboratory's test result for Sample Y was 0.42 and is categorized as "unusual." The large difference between the laboratory's two test results indicates a larger than usual random error, so the within-laboratory variation is categorized as "unusual."

7.7.4 The within-laboratory variation for Laboratory #5 is "typical," but the test results for Samples X and Y, 2.75 and 2.41, are categorized as "unusual."

7.8 *Estimating the Repeatability Standard Deviation for a Two-Sample Program:*

7.8.1 As explained in 7.3.1, in a two-sample program, the random error quantity, $[(X_i - Y_i) - (X_{med} - Y_{med})]$, calculated for each laboratory provides a value that represents the laboratory's random error.

7.8.2 As in the case for one-sample programs, the IQR is used to describe the spread of the distribution of the data set of random error quantities. An estimate of the sample standard deviation of the random error quantities is determined by dividing the IQR of the random error quantities (IQR_r) by the

TABLE 7 Test Results for Sample Y in Descending Order for Two-Sample Program

Count of Labs	Lab	Sample X Test Result	Between-Laboratory Category for Sample X	Sample Y Test Result	Between-Laboratory Category for Sample Y
	27	4.89	Extremely Unusual	5.28	Extremely Unusual
	5	2.75	Unusual	2.41	Unusual
	26	1.97	Typical	2.24	Typical
	13	2.03	Typical	2.21	Typical
	2	1.62	Typical	1.91	Typical
	15	1.84	Typical	1.81	Typical
	28	1.85	Typical	1.78	Typical
8th from Top	7	1.17	Typical	1.57	Typical
	23	1.43	Typical	1.52	Typical
	18	1.66	Typical	1.45	Typical
	22	1.54	Typical	1.39	Typical
	14	1.10	Typical	1.33	Typical
	8	1.76	Typical	1.28	Typical
	24	0.84	Typical	1.27	Typical
15th from Top	6	1.55	Typical	1.26	Typical
16th from Top	1	1.22	Typical	1.26	Typical
	11	1.19	Typical	1.26	Typical
	30	1.07	Typical	1.24	Typical
	10	1.18	Typical	1.21	Typical
	3	1.82	Typical	1.20	Typical
	19	1.28	Typical	1.20	Typical
	29	1.09	Typical	1.20	Typical
8th from Bottom	16	1.39	Typical	1.12	Typical
	17	1.13	Typical	1.08	Typical
	4	0.60	Typical	1.00	Typical
	25	0.98	Typical	0.99	Typical
	9	1.35	Typical	0.93	Typical
	21	0.69	Typical	0.77	Typical
	20	1.24	Typical	0.71	Typical
	12	1.71	Typical	0.42	Unusual

Shown Below Is Determination of Inner and Outer Fences for Sample Y Data Above

Median of Sample Y Test Results = 1.26
Median of Top Half for Sample Y (Upper Hinge) = 1.57
Median of Bottom Half for Sample Y (Lower Hinge) = 1.12
Interquartile Range for Sample Y (IQR_Y) = (1.57 – 1.12) = 0.45

$(3 \times IQR_Y)$ = 1.35
Outer Fence (Upper) for Sample Y = (1.57 + 1.35) = 2.92
Outer Fence (Lower) for Sample Y= (1.12 – 1.35) = –0.23

$(1.5 \times IQR_Y)$ = 0.675
Inner Fence (Upper) for Sample Y = (1.57 + 0.675) = 2.245
Inner Fence (Lower) for Sample Y = (1.12 – 0.675) = 0.445

Reproducibility Standard Deviation for Sample Y = (IQR_Y /1.35) = (0.45 /1.35) = 0.333

fixed factor, 1.35. However, since the random error quantities contain the differences between two measurements, the result (IQR_r/1.35) must also be divided by the square root of two to obtain an estimate of the repeatability standard deviation as follows:

$$\text{Repeatability Standard Deviation, } s_r = (IQR_r/1.35)/(\text{square root of } 2) \quad (6)$$

where:

IQR_r = interquartile range of the random error quantities.

7.8.3 For example, the repeatability standard deviation estimated for the data in Table 6 is [(0.45/1.35)/1.414], or 0.236, as shown at the bottom of the table.

7.8.4 This method of estimating the repeatability (within-laboratory) standard deviation determines a "pooled" estimate based on the test results for both samples that tends to overestimate the repeatability standard deviation. The estimate may not be appropriate if the test results for the two materials used for the two samples, X and Y, show significantly different degrees of test variation. See 7.9.6.

7.9 *Estimating the Reproducibility Standard Deviation for a Two-Sample Program:*

7.9.1 Determine the estimate for the reproducibility standard deviation for each sample, X and Y, by dividing the interquartile ranges of the test results for each sample by the fixed factor 1.35 as follows:

$$\text{Reproducibility Standard Deviation for Sample X, } S_{RX} = IQR_X/1.35 \quad (7)$$

$$\text{Reproducibility Standard Deviation for Sample Y, } S_{RY} = IQR_Y/1.35 \quad (8)$$

where:

IQR_X = interquartile range of the test results for Sample X, and

IQR_Y = interquartile range of the test results for Sample Y.

TABLE 8 Summary of Results for Samples X and Y for Two-Sample Program

Lab	Sample X Test Result	Between-Laboratory Category for Sample X	Sample Y Test Result	Between-Laboratory Category for Sample Y	Random Error Quantities $(X - Y) - (X_{med} - Y_{med})$	Within-Lab Category
1	1.22	Typical	1.26	Typical	–0.15	Typical
2	1.62	Typical	1.91	Typical	–0.40	Typical
3	1.82	Typical	1.20	Typical	–0.51	Typical
4	0.60	Typical	1.00	Typical	–0.51	Typical
5	2.75	Unusual	2.41	Unusual	–0.23	Typical
6	1.55	Typical	1.26	Typical	–0.18	Typical
7	1.17	Typical	1.57	Typical	–0.51	Typical
8	1.76	Typical	1.28	Typical	0.37	Typical
9	1.35	Typical	0.93	Typical	0.31	Typical
10	1.18	Typical	1.21	Typical	–0.14	Typical
11	1.19	Typical	1.26	Typical	–0.18	Typical
12	1.71	Typical	0.42	Unusual	1.18	Unusual
13	2.03	Typical	2.21	Typical	–0.29	Typical
14	1.10	Typical	1.33	Typical	–0.34	Typical
15	1.84	Typical	1.81	Typical	–0.08	Typical
16	1.39	Typical	1.12	Typical	0.16	Typical
17	1.13	Typical	1.08	Typical	–0.06	Typical
18	1.66	Typical	1.45	Typical	0.10	Typical
19	1.28	Typical	1.20	Typical	–0.03	Typical
20	1.24	Typical	0.71	Typical	0.42	Typical
21	0.69	Typical	0.77	Typical	–0.19	Typical
22	1.54	Typical	1.39	Typical	–0.04	Typical
23	1.43	Typical	1.52	Typical	–0.20	Typical
24	0.84	Typical	1.27	Typical	–0.54	Typical
25	0.98	Typical	0.99	Typical	–0.12	Typical
26	1.97	Typical	2.24	Typical	–0.38	Typical
27	4.89	Extremely Unusual	5.28	Extremely Unusual	–0.50	Typical
28	1.85	Typical	1.78	Typical	–0.04	Typical
29	1.09	Typical	1.20	Typical	–0.22	Typical
30	1.07	Typical	1.24	Typical	–0.28	Typical

Precision Estimates for Two-Sample Program Data Shown Above

Reproducibility Standard Deviation for Sample X = 0.467
Reproducibility Standard Deviation for Sample Y = 0.333
Pooled Reproducibility Standard Deviation, s_R = 0.406

Repeatability Standard Deviation, s_r = 0.236

7.9.2 To obtain one overall precision estimate (applicable to the test method used to obtain the test results and applicable for both samples, X and Y) rather than to have separate estimates of the reproducibility standard deviation for each sample, X and Y, the "pooled" reproducibility standard deviation must be determined.

7.9.3 Determine the "pooled" reproducibility standard deviation using the following equation:

$$\text{Pooled Reproducibility Standard Deviation, } S_R \quad (9)$$

$$= \text{square root of } \{[(n_X - 1)S_{RX}{}^2 + (n_Y - 1)S_{RY}{}^2]/(n_X + n_Y - 2)\}$$

where:

n_X = number of test results for Sample X,
n_Y = number of test results for Sample Y,
S_{RX} = reproducibility standard deviation for Sample X, and
S_{RY} = reproducibility standard deviation for Sample Y.

7.9.4 For the data in Table 8 for Samples X and Y, calculate the pooled reproducibility standard deviation as follows:

$$\text{Pooled Reproducibility Standard Deviation, } S_R \quad (10)$$

$$= \text{square root of } \{[(30 - 1)(0.467)^2 + (30 - 1)(0.333)^2]/(30 + 30 - 2)\}$$

$$= 0.406$$

7.9.5 When all laboratories have submitted results for both samples, n_X is equal to n_Y and Eq 9 can be simplified to the following:

$$\text{Pooled Reproducibility Standard Deviation, } S_R \quad (11)$$

$$= \text{square root of } [(S_{RX}{}^2 + S_{RY}{}^2)/2]$$

7.9.6 A large difference between the reproducibility standard deviations of the two samples is an indication that the materials used for the two samples were significantly different and that a pooled estimate of precision may not be appropriate for estimating the reproducibility and repeatability standard deviations. As a rule of thumb, if the ratio of S_{RY} to S_{RX} is greater than 1.1 or less than 0.9, a pooled estimate of precision should not be used without careful consideration of reasons for the difference between the test variations of the two samples. It may be more appropriate to consider the two samples as different materials and use Method A to determine a separate estimate of the reproducibility standard deviation for each sample.

8. Keywords

8.1 collaborative study; collaborative trial; interlaboratory comparison; proficiency testing