datapoints
a statistics q&a

What Are Repeatability and Reproducibility?

Part 2: The E11 Viewpoint[1]

BY NEIL ULLMAN

Q: Many ASTM standard test methods contain repeatability and reproducibility statements and values. What variations can be expected? What is the standard deviation for the repeatability and reproducibility of the method?

A: The interlaboratory study is the first step to obtaining contrasting values for the repeatability and reproducibility of a test method. It is a way to learn how well or poorly a test method behaves when performed in different situations.

ASTM standard E691, Practice for Conducting an Interlaboratory Study to Determine the Precision of a Test Method, is the basic standard describing how to perform an ILS and obtain values for the variations that one might expect for tests done at typical laboratories. As described in the preceding article of this series (see page 16 of the March/April *SN*), taking some number of repeats within each laboratory in a very short time by the same operator and equipment yields a best case situation, which should be the smallest variation among readings. This becomes a measure of the repeatability of measurements and is represented by calculating the *repeatability standard deviation*.

Usually only a small number of repeats for each test protocol in every laboratory is needed. By averaging results over many laboratories we learn how a typical lab might perform. Of course, all of this depends on how many laboratories participated and how well they represent the real world of laboratories. Thus, if you only have 10 laboratories involved, especially if they have been the ones developing the standard, it may be questionable to assume that any other random laboratory would perform as well. This could certainly be the case for new methods.

When we perform the test method in many different laboratories on the same material, we hope to discover all of the potential variations that can occur when the test method is used. Because we now have different operators, different equipment and different environmental conditions, all of the intermediate conditions and more will have been introduced. Thus, we should expect to have greater variability among the results from different laboratories. The measure of this larger variation due to readings taken among laboratories is found as the *reproducibility standard deviation*.

Again, the problem of interpreting this variation is that it is dependent on how many laboratories have participated. When only a very small sample of all the labs that might run the method is used, you must be cautious in believing that these results are typical of what all laboratories might have done with the same test. In addition, it is also important to look at the results to see if some laboratories consistently perform differently. Often the major reason for variation among laboratories is a consequence of some type of bias, or systematic difference, that occurs for one or more of the labs. This is especially a problem to recognize when only a very small number of labs actually participated (say, fewer than 10).

If we use the terms repeatability and reproducibility as describing the nature of the variation, then that variation is best computed as a standard deviation. Let's take a closer look at how these terms are defined in an ILS.

When the terms *r* and *R* were first presented, the idea was to provide a simple approximate comparison for a very special case of using the results of the ILS. The value of *r*, called the *repeatability interval*, is found by simply multiplying the repeatability standard deviation by 2.8; it is similar to the statistical estimate of a 95 percent confidence interval for the difference

between two readings. So, by using *r* we reduce the statistical jargon. The same goes for *R*, which is the reproducibility standard deviation times 2.8. With these calculations we arrive at a *reproducibility interval*, which we then use to compare the difference among a pair of actual test results that we might observe from two labs.

These interval types assume the following:

- The data are normally distributed; that is, the laboratories chosen are randomly picked from a distribution of labs that have approximately the same variation as the ILS results.
- We are only comparing two independent readings; that is, either two repeats by a single operator or single tests conducted in only two laboratories.
- That when we see a difference greater than the interval, there is a 5 percent chance that it is actually due to random chance, not an actual difference in the testing process.
- That the actual tests we look at fall in the range of those performed in the ILS.

A few comments should also be noted:

- As mentioned earlier, serious biases may be involved in tests performed in the ILS, especially when only a small number of labs were used, which calls into question how heavily one should rely on the estimate for a particular sample.
- E691 currently describes the case where two samples are taken. The International Organization for Standardization's ILS standard, ISO 5725, Accuracy (Trueness and Precision) of Measurement Methods and Results, suggests ways to compare different size samples either for repeatability or reproducibility based on modifying the multiplier of the standard deviation.
- In addition, if you want to reduce the chance of making the error of declaring a difference in results from the 5 percent value to, say only 1 percent, then the multiplier would need to change from 2.8 to about 3.6.

Even more important is the repeated use of the comparison of samples. For example, if you make many paired comparisons, the chance that one would randomly be different rapidly increases.

EXAMPLE — E691 SERUM IN GLUCOSE STUDY

For a serum in glucose study, eight laboratories testing five different reference samples of blood, ranging from low sugar levels to very high, measured each material three times. In the table we have the key summary statistics (Table 11 in E691). The first column contains the sample material averages, followed by the repeatability standard deviation, s_r, and then the reproducibility standard deviation, s_R. In all but the first case, s_R is larger than s_r. Recall that *r* is 2.8 times s_r, and *R* is 2.8 times s_R.

To interpret these values of standard deviation, if we had a reading of about 135 (material C), and we had a single operator in one laboratory run many tests on that material, then 95 percent of the readings would fall within a range of approximately +3.0 units (or about 1.96 times the repeatability standard deviations or a total

Material	$\bar{X}$	S_r	S_R	*r*	*R*
A	41.5183	1.0632	1.0632	2.98	2.98
B	79.6796	1.4949	1.5796	4.19	4.42
C	134.7264	1.5434	2.1482	4.33	6.02
D	194.7170	2.6251	3.3657	7.35	9.42
E	294.4920	3.9350	4.1923	11.02	11.74

range of about 6.0 units). But if only two readings were run at random, then 95 percent of the time the difference between those two readings should not be more than 4.33 units (the value of *r* for material C). Similarly, if many laboratories ran a single test then 95 percent of the single readings would fall in a range of about 8.6 units, but pairs of readings would rarely have a difference of greater than 6.02 units.

Next Issue: Part 3 – Repeatability and reproducibility in measurement systems analysis, or "gage R&R" methodology.

REFERENCE

1. ASTM Committee E11 on Quality and Statistics

NEIL ULLMAN *is a retired professor of mathematics and mechanical technology and a past chair of Committee E11. Currently chair of Subcommittee E11.20 on Test Method Evaluation and Quality Control, he is a fellow of ASTM and the American Society for Quality. He recently received the E11 Harold F. Dodge Award.*

DEAN NEUBAUER *is the DataPoints column coordinator and E11.90.03 publications chair.*

BONUS Q&A

Q: Regarding the use of the intermediate precision condition as shown in the March/April DataPoints column: Is there a manipulation of one of these calculations to be used when one is looking for limits relating to the intermediate precision condition? – Jeff Monson

A. The short answer is yes. Intermediate precision (*R'*) is defined in ASTM standard D6299, Practice for Applying Statistical Quality Assurance and Control Charting Techniques to Evaluate Analytical Measurement System Performance, as 2.77 times the standard deviation of a lab's quality control test results obtained under site precision conditions. A typical application of *R'* would be to determine if the retained material from a particular batch of released product has degraded over time. For that application, a lab will test the retained material and compare the test result with the original test result. If the difference exceeds *R'*, then there is strong evidence that the retained material has degraded, subject to the assumption that the test method is in statistical control when the original and retain test results are obtained, and the QC material that *R'* is based on is compositionally similar to the retain.

– Alex Lau

Statistics play an important role in the ASTM International standards you write, such as the development of precision and bias statements for test methods, running interlaboratory studies, knowing how to round numbers properly and determining sample size. A panel of experts is ready to answer your questions about how to use statistical principles in ASTM standards. Please send your questions to SN *Editor in Chief Maryann Gorman at mgorman@astm.org or ASTM International, 100 Barr Harbor Drive, P.O. Box C700, W. Conshohocken, PA 19428-2959.*