Data Significance

Understanding Statistical and Practical Significance

By Thomas J. Bzik

Q What is the difference between statistical and practical significance?

A Many data-based ASTM standards use relevant data along with a test or multiple tests of statistical significance as their primary measure of interpretation. In this regard they are analogous to most of the examples found in statistical textbooks, primarily focused on statistical significance. This can be too narrow a perspective. This article discusses another type of significance, practical significance, and its interrelationship to statistical significance. Understanding statistical problems simultaneously in the context of both types of significance is highly useful.

Two common applications of statistical significance in ASTM standards involve statistical significance testing for a statistically significant difference between the averages or variances of two samples. Statistical significance testing leads to a binary decision, either a statistically significant difference between averages or variances is identified or not. Statistical significance indicates that statistically strong evidence of a real measurable difference between the tested groups has been identified. Failure to obtain statistical significance is a weaker conclusion, that no statistically significant difference has been identified. There is no strong conclusion that they are the same, only that the observed difference was not large enough to be judged statistically different. In practice, not finding statistical significance often is taken to mean there is no difference or need for further action.

As an example, suppose the ASTM test is intended to establish the equivalency of two sets of data. Statistical methodology sets up two possible conclusions. The first is a null hypothesis that there is no difference between the two sets of data and an alternative hypothesis that the two sets are different. For example, if the chosen statistical test is to compare the variances of each of two sets of data, a Folded F-test would be used. Our focus is the binary nature of the results from hypothesis testing.

In the hypothesis test, if the variances differ sufficiently in terms of a statistical distance, the testing would indicate statistically significant evidence of a difference and the null hypothesis would be rejected. The sets of data would be judged not equivalent. Alternatively, if the difference fails to be judged statistically significant, the null hypothesis of no difference is accepted in common practice. Hence, if an ASTM procedure does not incorporate practical significance, the two data sets would be judged as equivalent.

Practical significance is another binary significance concept that is independent of statistical significance. Practical significance involves looking at the size of the observed difference in the problem context. If this size difference is consequential, then the difference is said to have practical significance; otherwise not. These two types of binary significance judgements lead to a 2x2 table of possible results (Tables 1 and 2). In Table 1, significance results are either in agreement or disagreement. Table 2 expresses Table 1 in terms of whether further action is required.

Send your statistics questions to Maryann Gorman at mgorman@astm.org.

Consider when statistical and practical significance are in agreement. If there is both statistical and practical significance, then the distributions represented by the samples should be treated as different in the context of the application. When there is neither statistical nor practical significance, then treat the distributions in question as being essentially equivalent in the application context.

When significance measures disagree, things get more interesting.

Consider where there is statistical significance but not practical significance. Here there is strong data-based evidence of a measurable difference, but the difference is judged to have a small enough impact not to be actionable. As an example, testing the equivalency of a new analytical instrument's standard deviation to the process of record standard deviation found a statistically significant difference. The new instrument has an observed standard deviation of 2.6 ppb, and the process of record has an observed standard deviation of 2.4 ppb. Suppose that the instrument manufacturer has stated instrument standard deviation can vary from instrument to instrument by up to 0.5 ppb. Here the observed difference is not practically significant relative to the instrument manufacturing process. Trying to fix the "instrument problem" would be ill-advised. It's not that a measurable difference has not been identified, it's that taking further action because of this result is questionable unless the instrument manufacturer were to improve the instrument in question. Here practical significance serves as a value judgement that addressing the statistically significant difference is of relatively limited value or practicality. This scenario becomes more likely with larger sample size. The use of more data permits relatively smaller differences to be identified as statistically significant.

Now consider the case where statistical significance is not found, but the result is of practical significance (the red case in Tables 1 and 2). Continue with the instrument example but now the new instrument has an observed standard deviation of 5.3 ppb, and the difference was found to be not statistically significantly different from 2.4 ppb. Suppose that the process engineer knows that if the instrument standard deviation is really 5.3 ppb this implies that 8 percent of future product is expected to test out of specification solely due to this increase in analytical uncertainty. If the standard deviation is really 5.3 ppb, then this translates to $30 million dollars less product to sell annually (e.g., the difference is of practical significance). In this case, the observed difference, if real, is consequential, but the lack of statistical significance implies the observed difference may have solely resulted from data noise. This situation is more likely to occur with smaller sample sizes than with larger sample sizes. The appropriate action is to collect more data. With enough additional data, the issue will be resolved. With additional data either (1) the observed difference will decrease to below practical significance, or (2) the additional data will make the result statistically significant. Once one or both significance measures flip, the issue is resolved.

An ASTM standard that incorporates aspects of practical significance is E2935, Practice for Conducting Equivalence Testing in Laboratory Applications. Many standard practices for equivalence testing in ASTM standards do not have the level of testing sophistication contained in E2935. The DataPoints article, "Testing for Equivalence: Why the TOST Procedure Works," contains additional useful reading material on both practical and statistical significance.[1]

Statistical methods are very valuable tools if one's approach is not artificially narrowed by focusing solely on statistical significance. Use both significance measures judiciously.

REFERENCE

1. Murphy, T.D., "Testing for Equivalence: Why the TOST Procedure Works," ASTM *Standardization News*, Sept./Oct. 2014, Vol. 42, No. 5, pp. 16-17 (www.astm.org/standardization-news/data-points/testing-for-equivalence-so14.html).

Table 1 — Statistical vs. Practical Significance

		Statistical Significance	
		YES	NO
Practical Significance	YES	Agree	Disagree
	NO	Disagree	Agree

Table 2 — Action Version of Table 1

		Statistical Significance	
		YES	NO
Practical Significance	YES	Take action as if the observed difference is real	Need more data before deciding how to proceed
	NO	Note difference, but take no further action	Take no further action

Thomas J. Bzik is a statistical consultant, Macungie, Pennsylvania; he serves as vice chairman of E11 on Quality and Statistics and as chairman of Subcommittee E11.10 on Sampling / Statistics.

John Carson, Ph.D., of CB&I Federal Services LLC and P&J Carson Consulting LLC, is the Data Points column coordinator. Vice chairman of E11.30 on Statistical Quality Control, part of E11 on Quality and Statistics, he is also a member of Committees E50 on Environmental Assessment, Risk Management, and Corrective Action, and of D02 on Petroleum Products, Liquid Fuels, and Lubricants.