

The Role of Statistical Significance Tests

Lynn D. Torbeck

Significance tests prevent accepting an apparent result to be real when it could be due to chance.

Statistical analysis provides scientists with an additional tool for making scientific interpretations and conclusions. Often, the analysis focuses on identifying significant differences, that is, practical and statistical differences.

Practical significance comes from comparing a difference (i.e., a signal) to an absolute reference. Statistical significance comes from comparing a difference to a relative reference that contains noise or random variability. Practical significance always takes precedence over statistical significance. In fact, statistical significance should not be checked until practical significance is found.

Statistical-significance testing compares signal with noise and is often expressed as a ratio of signal to noise. The result is not meant to be a statement of truth or reality.

If the signal can be shown to be larger than the noise (i.e., more of a difference than expected by chance variation alone), then the scientist may conclude it to be "statistically significant." Otherwise, we say we can't show it to be significant. If more data are obtained, the noise could be made smaller and then perhaps demonstrate that the signal is significant. In fact, if the noise is small enough or if the sample is large enough, even wildly impractical differences can be shown to be statistically significant.

The primary purpose of statistical-significance testing is to prevent the declaration of an apparent practical significance when in fact it could be due to random variation. The real difficulty is explaining statistical significance that has no practical meaning. Wang illustrates this point well:

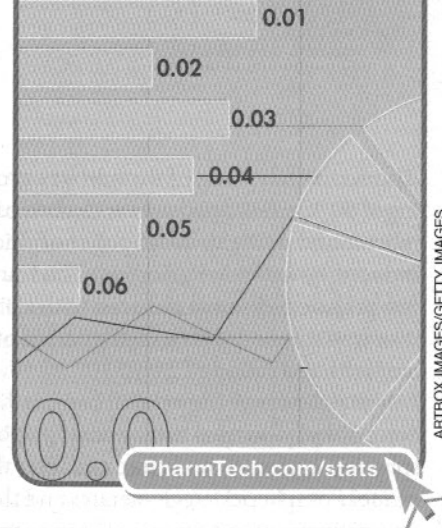
After years of reflection, I believe that statistical users will be better off if they take note of a two-stage test-of-significance as follows: Step 1: Is the difference practically significant? If the answer is NO, don't bother with the next step. Step 2: Is the difference statistically significant? How do we know if a difference is 'practically significant'? The answer is that you have to use subject matter knowledge. In general, there is no statistical formula readily applicable (1).

Given the expectation to write validation protocols with predetermined criteria, scientists sometimes find themselves in the unfortunate position of explaining why a highly significant statistical result is meaningless in practical terms. For example, the triplicate determinations below of two samples were measured using a high-performance liquid chromatography method that typically has a standard deviation of 3% for intermediate precision (2). Assume the specification is 90–110%.

Sample 1: 96.9, 101.2, 101.9

Sample 2: 99.7, 99.9, 100.4

Are these two samples different? Comparing the individual values with the specification, there is no practical difference. The averages are both 100.0%;



ARTBOX IMAGES/GETTY IMAGES

thus, there is no practical or statistical difference in the means.

The standard deviations are: $S_1=2.707\%$ and $S_2=0.361\%$. But are they different? When compared with the usual laboratory variability of the method (3%), there is no practical difference in the variability.

Are they statistically different? This difference can be tested with the F test of variances. The variances are the standard deviations squared. Thus, $V_1=7.330\%$, $V_2=0.130\%$, and $F=7.330/0.130$, which = 56.385. For samples of size 3 and an α of 0.05, the critical F is 19.00. With the calculated F larger than the critical F, the variability of the two samples are said to be statistically significantly different. From a scientific and practical viewpoint, this is meaningless.

We have found the right answer to the wrong question. The real issue is as defined in the International Conference on Harmonization guideline on analytical-procedure validation, "The objective of validation of an analytical procedure is to demonstrate that it is suitable for its intended purpose" (2).

To conclude, statistical-significance tests provide one more valuable tool for the scientist. But they are not a magic panacea, an end in themselves, or truth. They cannot provide a substitute for education, experience, subject-matter knowledge, or assertive confidence in our own abilities. In the end, we must use and accept good science and common sense.

References

1. C. Wang, *Sense and Nonsense of Statistical Inference* (Marcel Dekker, New York, 1993), pp. 1–4.
2. FDA, ICH Q2(R1) *Validation of Analytical Procedures: Definitions and Terminology* (FDA, Rockville, MD, 1995). **PT**



Lynn D. Torbeck is a statistician at Torbeck and Assoc., 2000 Dempster Plaza, Evanston, IL 60202, tel. 847.424.1314, Lynn@Torbeck.org, www.torbeck.org.