Statistical Testing in Context

How Analysis Provides Perspective on Results and Data

By Thomas J. Bzik

Q When are two things statistically different?

A First, the *difference pyramid:*

When are two things different?

When are two things meaningfully different?

When are two distributions statistically different?

When do two distributions have a particular different parameter?

Before engaging in statistical analysis, it is useful to have a well-oriented philosophical perspective on statistical testing.

Consider the t-test, which compares the sample averages from two distributions to determine if a statistically significant difference in population average can be found between them. Statistical testing is built upon a pyramid of proxy logic and assumptions, which will be denoted as the "difference pyramid." It is built upon the uncertain concept that the statistical decision at the base of the difference pyramid is somehow highly supportive of answering the fundamental question of "When are two things different?"

The question of when two things are different is a *trick question.* Statistical methods are entirely unnecessary to answer this question. Everything that is comparable is different regardless of whatever degree of similarity they may possess. Time, entropy, copying (manufacturing) imperfections, and quantum effects engineer this.

Does this render the rest of the pyramid, i.e., statistical testing, useless? A philosophical purist unequivocally would answer yes, there is no need to apply a statistical test to answer this question, nor is there a need to continue reading further.

If you are more of a pragmatist than a purist, however, continue reading. Statistics has *something* to offer in how it observes data in a tightly focused manner.

Statistics can answer a modified and narrower version of the question of when two things are different. This is a fundamentally different, nuanced question that must be well understood before statistical method application. Restart this examination at the base of the difference pyramid in the statistical context of the t-test.[1]

Good statistical practice is the business of determining when there is currently a "meaningful" difference context between the two things that are to be compared. Statistical testing is insufficient for determining what is a "meaningful" difference. Good statistical practice requires knowing something about the application context.[2]

The t-test compares the averages of two sample distributions to infer whether the two population distributions from which the samples have been drawn

are likely to have different underlying population averages (statistically significant difference found) or that no statistically meaningful difference between population averages has been found.

The question of "When are two things meaningfully different?" has been substantially narrowed to a simplified proxy argument. Statistical methods typically compare a distributional parameter as the proxy to assess whether underlying data distributions are meaningfully different. The data distributions themselves serve as proxies for assessing things as being measurably and/or meaningfully different from one another. Statistical methodologies typically identify statistically significant distributional parameter differences, which differs from finding meaningful differences. Using both practical significance (meaningful context significance) and statistical significance (statistically meaningful difference) is required for good statistical practice.

There are many observations (measurements) that can be made of "things," but the t-test narrows our focus to a single measurement parameter property of the things being compared. The measurements themselves serve as proxies for estimating the true value of the property.

Is this property — and the corresponding statistical test hypothesis — relevant or key to why the data is being analyzed? Why are the data being statistically analyzed? You do know why the data is being analyzed, don't you? Are other observable test properties more relevant, less relevant, or irrelevant to the purpose for which the study is being performed? Should more than just the average be statistically compared? For instance, are measures of variability or probabilities that the property exceeds a certain value also important? Has relevant data in sufficient quantity been collected for the project's purposes? Are suitably relevant observations even present in the current data?

Statistical practice effectiveness is much more than just mechanical application of statistical methods.

The most useful results from statistical testing occur when the statistical testing is closely aligned with the purpose for which the statistical tests are being implemented. Without this focus, if you do not know where you are going, any statistical test (road) will take you there. (The Cheshire Cat, in a moment of intense statistical reflection, provided this slightly modified quote.)

Statistical practitioners still need to understand their fundamental role as storyteller. Suppose the t-test has indicated a statistically significant difference in average. What story does one tell, and how is this statistical story made meaningful or more fully descriptive? Could the observed difference be due wholly or in part to differential observation (measurement system artifacts) rather than a more fundamental difference? Did time instability of samples wholly or in part trigger significance? Was the sampling representative? Did analyst judgment play a role in data collection or measurement? Was all the data provided or was some systematically withheld ("data weeding") or perhaps over-summarized ("destructive summarization")? Can systematic failures in observation (bias) be an explanation or not? Could failure to detect a difference be due to imprecision of the measurement system? Could significance or lack thereof be an artifact of the data violating the statistical test's fundamental assumptions? What are the implications of reducing statistical significance to only a yes or no result?

The purpose here is to illuminate analysis perspectives that are typically missing or underemphasized in basic statistical training. Current basic statistical training provides the statistical mechanics and resulting statistical interpretation but typically fails to describe the limitations and conceptual shortcuts used at the core of statistical testing. Data analysts have a strong tendency to jump prematurely to conclusions that are not necessarily complete or correct when statistical testing is performed.

Effective statistical analysis involves thinking deeply about what might be going on as well as likely data relevance and limitations before crafting the story of what statistical significance means or might mean. Statistical testing is a useful tool for developing data understanding and interpretation, but good statistical practice requires contextual knowledge to write a good data story (statistical analysis). The most important details are ignored when statistical methodologies are used in isolation.

Perhaps you have heard the story about the three blind scholars (blind = statisticians without data context) and the elephant. Looking at an isolated piece of a problem without wider reflection leads one to imagine the wrong story. Effective statistical analysis involves thinking deeply about what might be going on as well as likely data relevance and limitations before crafting the story of what statistical significance means or might mean. ■

REFERENCES

1. The practice for calculating and using basic statistics (E2586) provides related information.
2. Bzik, T.J., "Data Significance: Understanding Statistical and Practical Significance," Data Points, ASTM *Standardization News,* Vol. 44, No. 4, July/Aug. 2016.

Thomas J. Bzik is current chair of the committee on quality and statistics (E11). He works as a statistical consultant.

John Carson, Ph.D., senior statistician for Neptune and Co., is the Data Points column coordinator. He is a member of the committee on quality and statistics (E11), and a member of the committees on petroleum products, liquid fuels, and lubricants (D02), air quality (D22), and environmental assessment, risk management, and corrective action (E50).