Why We Don't "Accept" the Null Hypothesis

by Keith M. Bower, M.S. and James A. Colton, M.S.

Reprinted with permission from the American Society for Quality

When performing statistical hypothesis tests such as a one-sample t-test or the Anderson-Darling test for normality, an investigator will either *reject* or *fail to reject* the null hypothesis, based upon sampled data. Frequently, results in Six Sigma projects contain the verbiage "accept the null hypothesis," which implies that the null hypothesis has been proven true. This article discusses why such a practice is incorrect, and why this issue is more than a matter of semantics.

Overview of Hypothesis Testing

In a statistical hypothesis test, two hypotheses are evaluated: the null (H_0) and the alternative (H_1). The null hypothesis is assumed true until proven otherwise. If the weight of evidence leads us to believe that the null hypothesis is highly unlikely (based upon probability theory), then we have a statistical basis upon which we may reject the null hypothesis.

A common misconception is that statistical hypothesis tests are designed to select the more likely of two hypotheses. Rather, a test will stay with the null hypothesis until enough evidence (data) appears to support the alternative.

The amount of evidence required to "prove" the alternative may be stated in terms of a confidence level (denoted X%). The confidence level is often specified before a test is conducted as part of a sample size calculation. We view the confidence level as equaling one minus the Type I error rate (α). A Type I error is committed when the null hypothesis is incorrectly rejected. An α value of 0.05 is typically used, corresponding to 95% confidence levels.

The p-value is used to determine if enough evidence exists to reject the null hypothesis in favor of the alternative. The p-value is the probability of incorrectly rejecting the null hypothesis.

The two possible conclusions, after assessing the data, are to:

1. Reject the null hypothesis (p-value <= α) and conclude that the alternative hypothesis is true at the pre-determined confidence level of X%, or at the observed and more specific confidence level of 100*(1 – p-value)%.

2. Fail to reject the null hypothesis (p-value > α) and conclude that there is not enough evidence to state that the alternative is true at the pre-determined confidence level of X%. Note that it is possible to state the alternative to be true at the lower confidence level of 100*(1 – p-value)%.

Ronald A. Fisher succinctly discusses the key point of our paper:

> *In relation to any experiment we may speak of... the "null hypothesis," and it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.*[1]

A Helpful Analogy: The U.S. Legal System

Consider the example of the legal system in the United States of America. A person is considered innocent until proven guilty in a court of law. We may state this particular decision-making process in the form of a hypothesis test, as follows:

H_0: Person is innocent
vs.
H_1: Person is not innocent (i.e., guilty)

The responsibility then falls upon the prosecution to build a case to prove *guilt* beyond a reasonable doubt. It should be borne in mind that a jury will never find a person to be "innocent." The defendant would be found "not guilty" in such a situation; i.e., the jury has *failed to reject* the null hypothesis.

Decisions Based on Data

We must keep in mind, of course, that it is **always** possible to draw an incorrect conclusion based upon sampled data. There are two kinds of error we can make:

- **Type I error.** When the null hypothesis is rejected, practitioners refer to the Type I error when they present results, using language such as: *"We reject the null hypothesis at the 5% significance level,"* or *"We reject the null hypothesis at the 95% confidence level."*

- **Type II error.** A second possible mistake involves incorrectly failing to reject the null hypothesis. The power of a test is defined as one minus the Type II error rate, and is therefore the probability of *correctly* rejecting H_0. The sample size plays an important role in determining the statistical power of a test.

When statisticians address small sample sizes, they often refer to the power to justify their concerns. One may argue that the sample size would be too low to correctly detect a difference from the hypothesized value, if that difference truly existed.

Example of a Test with Low Power

Consider a test that compares the mean of a process to a target value. The null and alternative hypotheses are, respectively:

H_0: Process mean on target
vs.
H_1: Process mean different from target

Suppose two observations are collected daily to monitor for a change in the process mean (i.e., n = 2). Assume a one-sample t-test is carried out at the $\alpha = 0.05$ significance level (95% confidence level) and the resulting p-value is above 0.05.

Fig. 1 One-Sample t-Test

Power and Sample Size

```
1-Sample t Test

Testing mean = null (versus not = null)
Calculating power for mean = null + difference
Alpha = 0.05  Sigma = 1

            Sample
Difference    Size   Power
         6       2  0.4944
```

As is shown in Figure 1, there is less than a 50% chance (power = 0.4944) such a test will correctly reject the null hypothesis even when the difference between the process mean and the target is six standard deviations. This is obviously an enormous statistical difference, yet the test (owing to the small sample size) would not be sensitive to it. The danger in concluding the process is on target with a sample size of two, for this example, is evident.

Implications

Assessing and relaying findings in a cogent manner is critical for Six Sigma practitioners. In statistical hypothesis testing procedures, this means that investigators should avoid misleading language such as that which implies "acceptance" of the null hypothesis.

Reference

1. Ronald A. Fisher, *The Design of Experiments*, 8th ed. (New York: Hafner Publishing Company Inc., 1966), 17.

Bibliography

1. Lenth, Russell V. "Some Practical Guidelines for Effective Sample Size Determination." *The American Statistician* 55, no. 3 (2001): 187-193.
2. Tukey, John W. "Conclusions vs. Decisions." *Technometrics* 2, no. 4 (1960): 423-433.