Cornell University
Cornell Statistical Consulting Unit

StatNews #85

Equivalence Testing

January 2013

Equivalence testing is an important statistical tool, utilized when an investigator seeks to claim that observations from two groups are similar enough for practical purposes. Equivalence testing is growing in popularity and is frequently used by pharmaceutical companies to determine, for example, whether a new drug is as effective as the gold standard in producing weight loss.

Equivalence testing should not be confused with the more familiar method of significance testing when comparing two means. The two approaches share an overall strategy, where a researcher assumes a "*null hypothesis*" and tests whether the data provides sufficient evidence to *reject* it (thus, concluding that the "alternative hypothesis" is true). Equivalence testing and significance testing for differences between two means differ in how the null hypothesis is stated. In the traditional significance testing framework the null hypothesis is that the two mean responses are similar, while in equivalence testing the null hypothesis is that the difference between them is greater than a prescribed amount, denoted by Δ, which is referred to as "interval of tolerable difference". In mathematical notation, let μ_C and μ_T be the mean responses in the control and treatment groups, respectively. Then the null hypothesis of significance is $\mu_C - \mu_T = 0$, whereas in equivalence testing, the null hypothesis is $|\mu_C - \mu_T| > \Delta$.

The threshold Δ represents a difference that is not large enough to have any clinical implications. Oftentimes, researchers will calculate Δ using a percentage of the parameter of interest. Alternatively, the investigator can use prior knowledge or expertise to set Δ.

To illustrate the difference between equivalence and significance testing consider a simple example. Suppose we are interested in comparing the weight loss efficacy of a new diet pill vs. the gold standard as measured by the total weight loss (in pounds, over six months)

Group	Observed mean	Standard deviation
Control (gold standard) n=100	$\bar{x}_C = 24.9$	2.4
Treatment (new pill) n=100	$\bar{x}_T = 24.2$	1.8

With significance testing, we compute the t-statistic
$t_S = (24.9 - 24.2)/\sqrt{2.4^2/100 + 1.8^2/100} = 2.333$, which is significant at the 5% level (two-sided test, df=198). Hence, we reject the null hypothesis and conclude that the

average weight loss among the subjects that received the new diet pill is significantly different from the weight loss in the control group.

In contrast, suppose that with the equivalence testing approach we specify in advance that, based on previous clinical studies, a difference of 2 pounds in weight loss is not meaningful. In the above notation, $\Delta=2$. Formally, the null hypothesis is $|\mu_C - \mu_T| > \Delta$, which can be written as two inequalities: $\mu_C - \mu_T > \Delta$ and $\mu_C - \mu_T < -\Delta$. Hence, we have to compute two one-sided test statistics and apply the Bonferroni correction for multiple testing, so that the following two null hypotheses are tested at the $\alpha/2$ level:

$$H_{01}: \mu_C - \mu_T > 2 \quad t_1 = \left(\left(24.9 - 24.2\right) - 2\right) / \sqrt{2.4^2/100 + 1.8^2/100} = -4.333$$

$$H_{02}: \mu_C - \mu_T < -2 \quad t_2 = \left(\left(24.9 - 24.2\right) - (-2)\right) / \sqrt{2.4^2/100 + 1.8^2/100} = 9$$

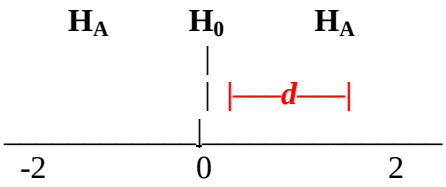
We reject the null hypothesis $|\mu_C - \mu_T| > \Delta$ by comparing these t-statistics with the critical values at the 2.5% level and conclude that the two treatments are equivalent.

These differing approaches can be summarized visually using confidence intervals as follows. The red segments represent a 95% confidence interval for the difference between the means (denoted by *d*). The confidence interval does not include 0, so according to the significance testing approach, the difference between the groups is statistically significant, but it may not be a large enough difference in practical terms, because *d* is in the interval of tolerable difference.

Significance Testing:

$H_0: \mu_T - \mu_C = 0$

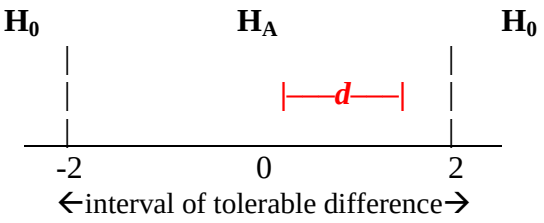
$H_A: \mu_T - \mu_C \neq 0$



Equivalence Testing:

$H_0: |\mu_T - \mu_C| > \Delta$

$H_A: |\mu_T - \mu_C| \leq \Delta$



It is important to note that in the significance testing approach for testing equality of two means, failure to reject the null hypothesis does not imply that the two means are equal. In other words, a small p-value provides a measure of the evidence against the null, but a large p-value does *not* provide evidence *for* the null. As Altman and Bland put it, "absence of evidence is not evidence of absence".

Also, note that it can be that neither the null hypothesis of significance ($\mu_C - \mu_T = 0$), nor the null hypothesis of equivalence ($|\mu_C - \mu_T| > \Delta$) are rejected. In such cases, it is not possible to determine if the two means are different or equivalent. For example, if in our hypothetical example we had n=16 (rather than 100), we would have obtained $t_S=0.933$ (thus, cannot reject the null hypothesis in significance testing), and $t_1= -1.733$ (thus, cannot reject the null hypothesis in equivalence testing). This is a case of insufficient

power, and hence, it is important to perform sample size calculations in advance for both hypotheses (significance/equivalence test).

In order to determine which method of testing to conduct, one must develop a clear analytic objective, and state the null hypothesis accordingly. For those seeking to prove that a new treatment is as effective as another, equivalence testing is an appropriate approach, provided that the width of the interval of tolerable difference can be determined.

If you have any questions regarding this topic, please contact the CSCU Office.

Authors: Jonathan Shtaynberger and Haim Bar

References

Ebbut, E., Jones, B., Jarvis, P., Lewis, J. (1996) Trials to assess equivalence: the importance of rigorous methods. British Medical Journal

Barker, L., Luman, E., McCauley, M., Chu, S. (2002). Assessing Equivalence: An Alternative to the Use of Difference Tests for Measuring Disparities in Vaccination Coverage. *American Journal of Epidemiology*

Altman DG, Bland JM. (1995). Absence of evidence is not evidence of absence. *BMJ*