Equivalence Testing

Statnews #85

Cornell Statistical Consulting Unit

Created January 2013. Last updated August 2020

Introduction

Equivalence testing seeks to claim that observations from two groups are similar enough for practical purposes. For example, a pharmaceutical company may wish to determine whether a new drug is as effective as an existing, "gold standard" drug in producing weight loss.

Equivalence testing should not be confused with the more familiar method of significance testing for comparing two population means. A common statistical analysis comparing two population means tests the null hypothesis that the population means are equal. In contrast, in an equivalence test the null hypothesis is that the difference between two population means is greater than a particular amount, denoted Δ, which may be referred to as an interval or margin of tolerable difference. If μ_C and μ_T are the population mean responses in the control and treatment groups, respectively, then the null hypothesis in equivalence testing is $|\mu_C - \mu_T| > \Delta$.

The threshold Δ represents a difference that is not large enough to be clinically meaningful. The choice of Δ may be based on expert opinion or previous estimates of the effect of the treatment compared to a placebo.

Example

To illustrate the difference between equivalence testing and significance testing for population mean differences, suppose we are interested in comparing a new diet pill to an existing gold standard pill based on the total weight loss, in pounds, during six months of use. Suppose that there are $n = 100$ observations in both the control (gold standard) group and treatment (new pill) group, and the sample means are $\overline{x}_C = 24.9$ pounds in the control group and $\overline{x}_T = 24.2$ pounds in the treatment group. The standard deviations are 2.4 in the control group and 1.8 in the treatment group.

In a significance test, the null hypothesis is $H_0: \mu_C = \mu_T$ and the alternative hypothesis is $H_A: \mu_C \neq \mu_T$. To perform this test, we compute the t statistic

$$t = \frac{24.9 - 24.2}{s\sqrt{1/100 + 1/100}} = 2.333,$$

where $s = \sqrt{\frac{(100-1)2.4^2+(100-1)1.8^2}{100+100-2}} = 2.121$ is the pooled standard deviation estimate. With 198 degrees of freedom, we reject the null hypothesis at the 5 percent level, concluding that the average weight loss among the subjects receiving the new diet pill is not equal to that of the subjects receiving the standard pill.

Now consider an equivalence test. Suppose that based on previous clinical studies, a difference of 2 pounds in average weight loss is not clinically meaningful. That is $\Delta = 2$ is the margin of difference between the population means for which the two pills will be considered equivalent. The null hypothesis is $H_0: |\mu_C - \mu_T| > 2$, which is equivalent to the hypothesis

$$H_0: \mu_C - \mu_T > 2 \text{ or } \mu_C - \mu_T < -2.$$

To test this null hypothesis of non-equivalence, we perform two one-sided tests of the hypotheses $H_{01}: \mu_C - \mu_T > 2$ and $H_{02}: \mu_C - \mu_T < -2$. The null hypothesis of non-equivalence is rejected if and only if both H_{01} and H_{02} are rejected. This is the "two one-sided test" (TOST) approach to equivalence testing. For H_{01}, the test statistic is

$$t_1 = \frac{24.9 - 24.2 - 2}{s\sqrt{1/100 + 1/100}} = -4.333$$

and for H_{02} the test statistic is

$$t_2 = \frac{24.9 - 24.2 - (-2)}{s\sqrt{1/100 + 1/100}} = 9.$$

The $\alpha = 0.05$ critical value for a t distribution with 198 degrees of freedom is 1.653, so both H_{01} and H_{02} are rejected. Thus, H_0 is rejected and the two treatments are considered equivalent.

Equivalence testing with confidence intervals

The above procedure can be further understood by constructing a confidence interval. If the endpoints of a $1 - 2\alpha$ confidence interval for $\mu_T - \mu_C$ are contained within the interval $[-\Delta, \Delta]$, then the null hypothesis of non-equivalence will be rejected using the TOST procedure and the two groups are said to be equivalent. With $\alpha = 0.05$, the $1 - 2\alpha$ critical value for a t distribution with 198 degrees of freedom is -1.286, so the 90-percent confidence interval is $(1.086, 0.314)$, which is contained in the interval $[-2, 2]$.

Figure 1 illustrates the equivalence test, in which the alternative hypothesis is true when the population mean difference is between -2 and 2, and the null hypothesis is rejected in this case because the endpoints of the 90-percent confidence interval are both between -2 and 2. Figure 1 illustrates how a confidence interval would be used to perform a significance test for a population mean difference at the 0.05 level: the **95**-percent confidence interval does not contain zero, so $H_0: \mu_T - \mu_C = 0$ is rejected.

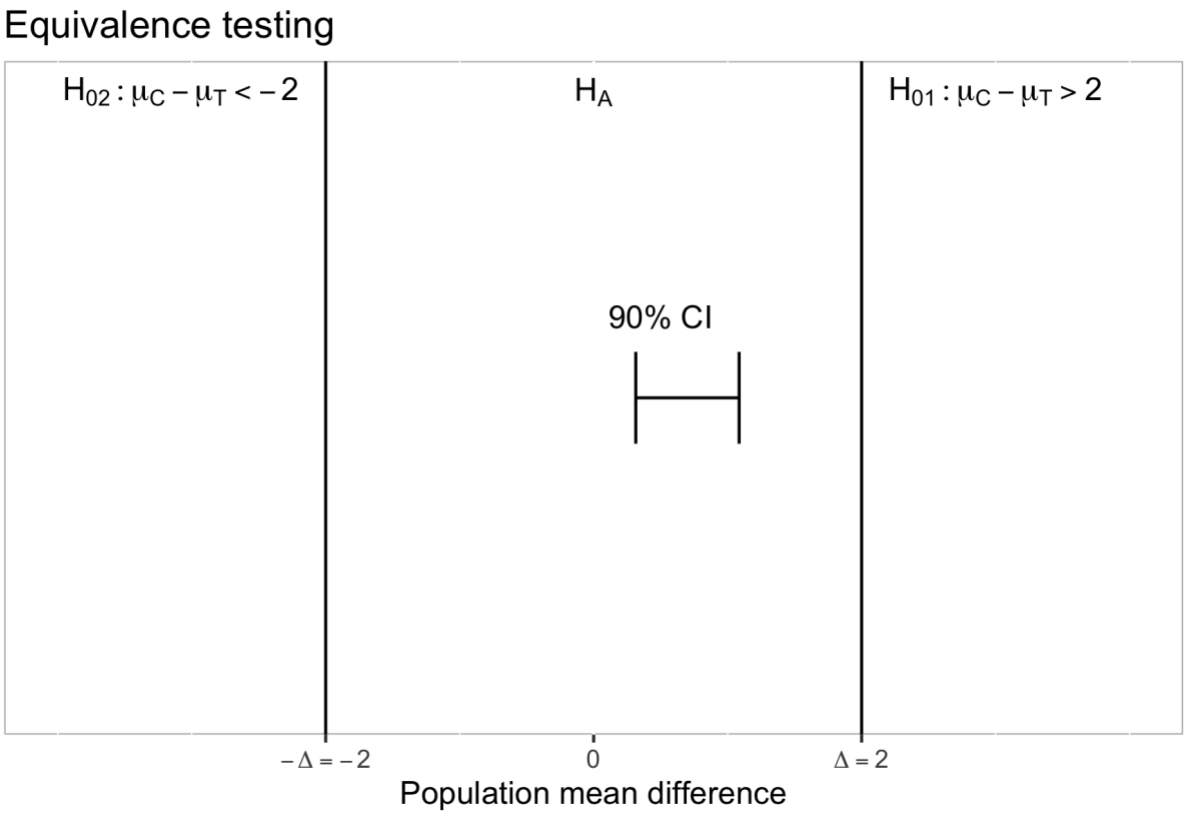

Figure 1: Equivalence testing with a confidence interval: the $100 \times (1 - 2\alpha)$*-percent confidence interval is contained in the interval* $[-\Delta, \Delta]$*, so the null hypothesis is rejected.*

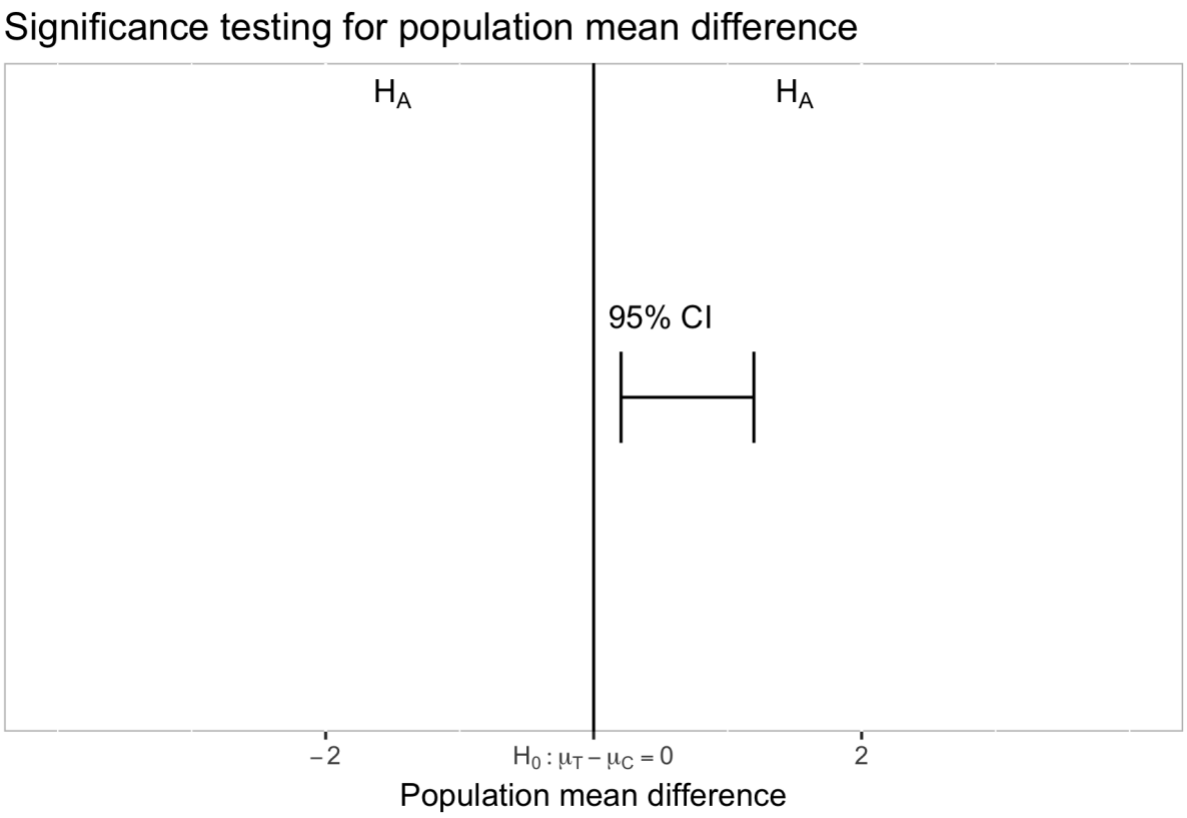

Figure 2: Significance testing for a population mean difference using a confidence interval: the $100 \times (1 - \alpha)$*-percent confidence interval excludes zero and hence the null hypothesis of equal population means is rejected.*

Final thoughts

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.

Also, note that it is possible to fail to reject both the null hypothesis of significance ($\mu_C - \mu_T = 0$) and the null hypothesis of equivalence ($|\mu_C - \mu_T| > \Delta$). In such cases, it is not possible to determine if the two means are different or equivalent. Also note that power calculations for equivalence tests are not the same as power calculations for significance tests of a population mean difference.

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

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. Journal of pharmacokinetics and biopharmaceutics, 15(6), 657-680.

Lakens, D. (2017). Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. Social psychological and personality science, 8(4), 355-362.