Benchmarks Online

Skip Navigation Links

Page One

Campus Computing News

New and Improved Microsoft Elearning Titles Available to Faculty and Staff

TurnItIn.com: Don't Be Cheated On!

Pedagogy 101 for Distance Learning

Spring Break Hours

Today's Cartoon

RSS Matters

The Network Connection

Link of the Month

WWW@UNT.EDU

Short Courses

IRC News

Staff Activities

Subscribe to Benchmarks *Online*

Research and Statistical Support
University of North Texas

# RSS Matters

*Link to the last RSS article here: A.M.: Faculty Evaluations After the Mainframe - Ed.*

## Equivalence Tests

**By Dr. Mike Clark, Research and Statistical Support Services Consultant**

A common use of statistical analysis entails a comparison of groups to one another. The familiar Student’s t-test is used to distinguish whether two groups are significantly different from one another. Starting with a null hypothesis that states there is no difference between the two samples (e.g. $\mu_1 - \mu_2 = 0$), one proceeds to determine the probability with finding the difference seen in the data if that hypothesis were true. If the probability of the result is low (typically p < .05), then one rejects the null hypothesis that claims no difference, and concludes the groups are statistically different from one another. Alternatively, one may set up a design to minimize Type I error rate (a), and conclude rejection of the null hypothesis if the observed t-statistic is at or beyond the t-critical value ($t_{cv}$) associated with a chosen error rate and sample size.

Take for example, an examination of the effectiveness of a particular teaching strategy for high school math students. Some students are randomly assigned either to classes in which the method will be implemented in their courses, or those which will receive no special treatment and can thus serve as a control group. After the semester is over they are given a math proficiency exam, which resulted in the following data:

$$\bar{X}_{teach} = 75 \quad s = 3 \quad n = 20$$
$$\bar{X}_{cont} = 71 \quad s = 5 \quad n = 20$$

As mentioned, our null hypothesis would be that there is no difference between the groups outside of sampling error. As the basics of the t-test are well-known, I will not rehash the details of the procedure here, but one can review them in introductory statistical texts. The data above would give us a t-statistic of 3.07, which would allow us to reject the null hypothesis (a = .05, two-tailed test, df = 38) and claim the teaching strategy results in higher math scores.

*Equivalence*

What would we have done had we not reached our specified significance criterion? Common practice is to assume the two groups are equivalent, and such a conclusion seems hardly far-fetched. However there are several problems with doing so. First, a small sample size can make it very hard to find statistical significance, so if it was our goal to establish equivalence we could just have a small sample. Secondly, the procedure establishes evidence against the null hypothesis only, not for it[i]. Furthermore, problems with the data itself (e.g. outliers), may hamper our ability to find a result at the specified significance level, and so again could come to a conclusion of equivalence just because we had messy data.

One may be left wondering what to do to determine equivalence between two groups. A statistical analysis that is just now coming into wider acceptance among the social sciences is *equivalence testing*. Often used in biomedical studies to examine different treatments' relative effectiveness, it provides a method for establishing whether two samples of data are functionally equivalent with regard to some statistic of interest.

The first step involves establishing a range of acceptable values such that any observed difference less than a certain amount may be shown to imply equivalence among the groups. As sampling error could result in our difference falling into that range, one performs inferential analysis to determine statistical equivalence.

*Two one-sided tests*

Return to the previous example. Say that based on those results we began teaching our math classes using the new method. Someone comes along later and says they have an alternative method of teaching the courses that could be even more effective. The principal is hesitant to alter the existing setup because changing the system before was very costly in terms of training and materials, and doesn't want to go through another overhaul unless there is substantial improvement to be had. Based on her knowledge of the proficiency test and other practical considerations, she states that if there is no more than a 5 point improvement she will maintain the current situation. The resulting data at the end of the semester is as follows.

$$\bar{X}_{new} = 78.5 \quad s = 4 \quad n = 20$$
$$\bar{X}_{old} = 76.0 \quad s = 5 \quad n = 20$$

Although we do not reach the established criterion of interest in raw form, the presence of sampling error still does not allow us to determine statistical equivalence right off. One method requires testing the following joint null hypothesis:

$$H_{01}: \mu_1 - \mu_2 \geq \Delta$$
$$H_{02}: \mu_1 - \mu_2 \leq -\Delta$$

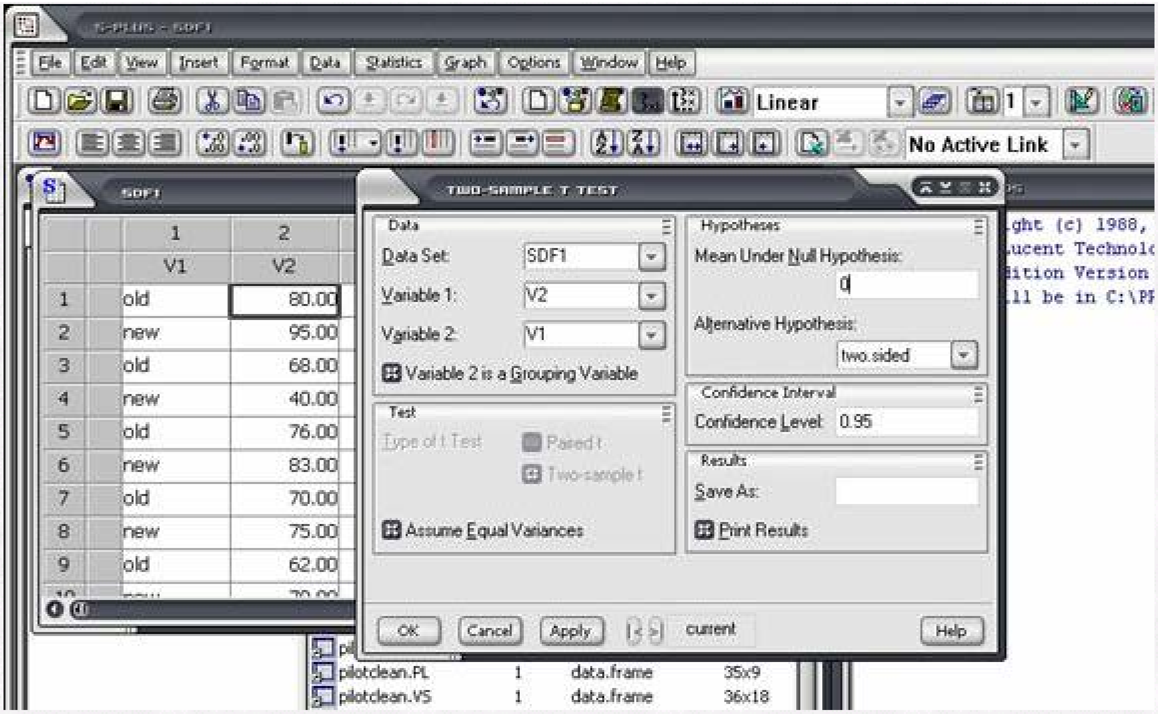
where Δ refers to our specified maximum difference allowed to still profess equivalence, in this case, 5 as determined by the principal of the school. Rejection of $H_{01}$ implies $\mu_1 - \mu_2 \leq 5$. Rejection of $H_{02}$ implies that $\mu_1 - \mu_2 \geq -5$. Rejection of both suggests our difference falls in the range of -5 to 5 and so we could include they are equivalent. In order to do so we perform two one-sided t-tests.

$$t_1 = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta}{s_{\bar{X}_1 - \bar{X}_2}}$$
$$t_2 = \frac{(\bar{X}_1 - \bar{X}_2) - (-\Delta)}{s_{\bar{X}_1 - \bar{X}_2}}$$

If $t_1$ is less than the negative $t_{cv}$ we reject $H_{01}$, and if $t_2$ is greater than the positive $t_{cv}$ we reject $H_{02}$.

We do this because starting off we don’t know how the means could have turned out relative to one another (new greater than old or vice versa) and so could have a positive or negative difference. However, another perhaps more simple way to look at it would be that we are testing whether our absolute difference is significantly less than that specified for non-equivalence, or that our null hypothesis would be $H_0: |\mu_1 - \mu_2| \geq \Delta$. By rejecting the null hypothesis, we conclude that the difference observed qualifies for practical equivalence.

Such a test can be performed in a statistical package such as S-Plus. From the Statistics menu, go to Compare samples/Two samples/t-test



S-PLUS - SDF1

SDF1

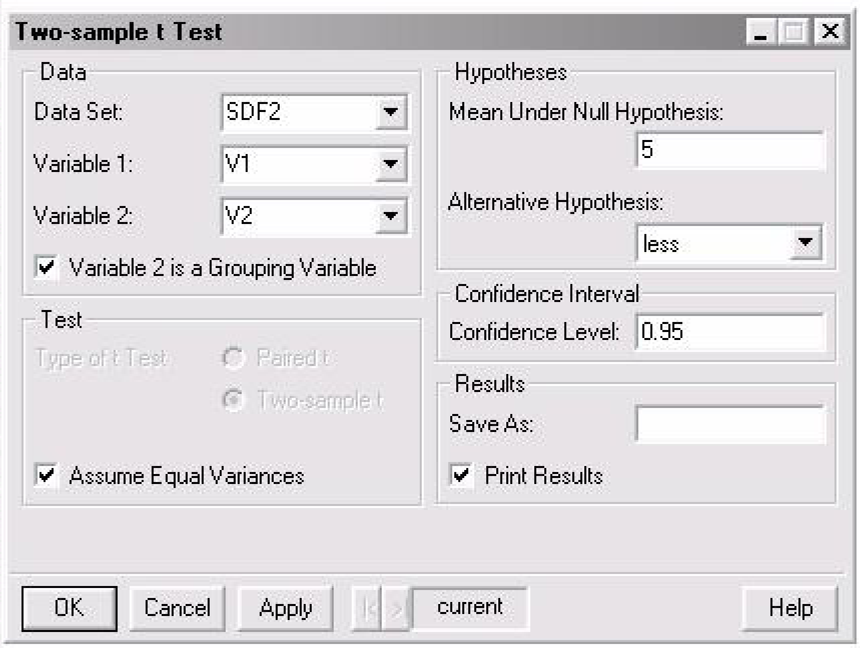
| | 1 | 2 |
|---|---|---|
| | V1 | V2 |
| 1 | old | 80.00 |
| 2 | new | 95.00 |
| 3 | old | 68.00 |
| 4 | new | 40.00 |
| 5 | old | 76.00 |
| 6 | new | 83.00 |
| 7 | old | 70.00 |
| 8 | new | 75.00 |
| 9 | old | 62.00 |

Two-Sample t Test

Data Set: SDF1
Variable 1: V2
Variable 2: V1
☒ Variable 2 is a Grouping Variable
☒ Assume Equal Variances
Mean Under Null Hypothesis: 0
Alternative Hypothesis: two.sided
Confidence Level: 0.95
Save As:
☒ Print Results

pilotclean.PL 1 data.frame 35x9
pilotclean.VS 1 data.frame 36x18

Choose your variables under consideration and select whether one is a grouping variable. The only thing that remains is to change the "Mean Under Null Hypothesis" (MUNH) and select our alternative hypothesis[ii]. Let us begin with the first part of the joint null hypothesis. Enter in a value of 5 for MUNH and select as our Alternative Hypothesis "less", which is to say that the alternative hypothesis in competition with the null hypothesis suggests that our result is significantly less than 5. See below.



Two-sample t Test

Data Set: SDF2
Variable 1: V1
Variable 2: V2
☑ Variable 2 is a Grouping Variable
☑ Assume Equal Variances
Mean Under Null Hypothesis: 5
Alternative Hypothesis: less
Confidence Level: 0.95
Save As:
☑ Print Results

To test the other in this situation we simply change (MUNH) to -5 and our competing alternative hypothesis to "greater". If the t-statistics obtained exceed our critical values in the above mentioned fashion, as they do here ($t_{cv}$ = ±1.68), we can reject the joint null and claim statistical equivalence.

# The Confidence Interval Approach

Another way in which to test for equivalence calls for the construction of a confidence interval (CI) about the difference between means observed. Having initially specified our Δ boundaries, we then see if e.g. the 95% CI for the difference between means falls *entirely* within that range. To obtain the interval for the above example: $(\bar{X}_1-\bar{X}_2)\pm t_{cv}(s_{\bar{X}_1-\bar{X}_2})$ . If we want this approach to provide the same results as the two one-sided tests performed above, we construct a CI at the 100(1-2a)% level used in the previous approach. Going the other way, if our decision was to provide a CI at the 95% confidence level for the difference, we would perform the two one-sided tests above at a = .025.

*Inferential Confidence Intervals*

An alternative confidence interval method comes from Tryon (2001). His approach involves what he refers to as *inferential* confidence intervals (ICI). These are different from ordinary confidence intervals in that they use they use a reduced critical value such that non-overlap of the two group's ICIs for their respective means suggests statistical difference at the specified alpha level for a significant difference. To test equivalence, one takes the lower bound of the lesser mean's ICI and the upper bound of the greater mean's ICI to establish a range of the difference between them ($R_g$ = Upper bound – Lower bound). If this $R_g$ is less than that established for non-equivalence (the above Δ), we conclude equivalency among the groups. Tryon's method may be preferable in that in contains within it a means of testing both equivalence and difference. Furthermore, it provides a third possible outcome, indeterminancy, where neither equivalence nor difference can be established statistically.

*Summary*

The social sciences are overdue for a much wider implementation of tests of equivalence. It is often the case that researchers claim evidence of a particular hypothesis through non-rejection of the null hypothesis of no difference. However, it is incorrect to conclude equivalent groups just because we do not have enough evidence against the null, as Fisher pointed out decades ago. Methods are readily available to determine whether two groups are statistically equivalent, and should be used accordingly.

## Some Resources

Altman, D.G., & Bland, J.M. (1995). Absence of Evidence is not Evidence of Absence. *British Medical Journal,* 311, 485.

Rogers, J.L., Howard, K.I., & Vessey, J.T. (1993). Using Significance Tests to Evaluate Equivalence Between Two Experimental Groups. *Psychological Bulletin, Vol 113(3),* 553-565.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6,* 371-386.

[i] In fact, our result doesn't speak to any hypothesis in a particular, only the probability of some outcome, which we then use to make a decision regarding the null hypothesis.

[ii] This procedure already differs from most t-tests done in social sciences as despite repeated pleas from quantitative types typically we test a 'nill' hypothesis of no difference.

Return to top