Tests of Equivalence and Confidence Intervals for Effect Sizes

---

Point or sharp null hypotheses specify that a parameter has a particular value -- for example, $(\mu_1 - \mu_2) = 0$, or $\rho = 0$. Such null hypotheses are highly unlikely ever to be absolutely true. They may, however, be close to true, and it may be more useful to test range or loose null hypotheses that state that the value of the parameter of interest is close to a hypothetical value. For example, one might test the null hypothesis that the difference between the effect of drug G and that of drug A is so small that the drugs are essentially equivalent. Biostatisticians do exactly this, and they call it bioequivalence testing.

Steiger (2004) presents a simple example of bioequivalence testing. Suppose that we wish to determine whether or not generic drug G is bioequivalent to brand name drug B. Suppose that the FDA defines bioequivalence as bioavailability within 20% of that of the brand name drug. Let θ_1 represent the lower limit (bioavailability 20% less than that of the brand name drug), θ_2 the upper limit (bioavailability 20% greater than that of the brand name drug), and θ_G the bioavailability of the generic drug. A test of bioequivalence amounts to pitting the following two hypotheses against one another:

H_{NE}: $\theta_G < \theta_1$ or $\theta_G > \theta_2$ -- the drugs are not equivalent
H_E: $\theta_1 \leq \theta_G \leq \theta_2$ -- the drugs are equivalent -- note that this a range hypothesis

In practice, this amounts to testing two pairs of directional hypotheses:
H_0: $\theta_G \leq \theta_1$ versus H_1: $\theta_G > \theta_1$ and H_0: $\theta_G \geq \theta_2$ versus H_1: $\theta_G < \theta_2$.

If both of these null hypotheses are rejected, then we conclude that the drugs are equivalent. Alternatively, we can simply construct a confidence interval for θ_G -- if the confidence interval falls entirely within θ_1 to θ_2 , then bioequivalence is established.

Steiger (2004) opines that tests of equivalence (also described as tests of close fit) have a place in psychology too, especially when we are interested in demonstrating that an effect is trivial in magnitude. Steiger recommends the use of confidence intervals, dispensing with the traditional NHST procedures (computation of test statistic, *p* value, decision).

Suppose, for example, that we are interested in determining whether or not two different therapies for anorexia are equivalent. Our criterion variable will be the average amount of weight gained during a two month period of therapy. By how much would the groups need differ before we would say they differ by a nontrivial amount? Suppose we decide that a difference of less than three pounds is trivial. The hypothesis that the

difference (D) is trivial in magnitude can be evaluated with two simultaneous one-sided tests:

$H_{0:}$ $D \leq -3$ versus H_1: $D > 3$, and H_0: $D \geq 3$ versus H_1: $D < 3$

After obtaining our data, we simply construct a confidence interval for the difference in the two means. If that confidence interval is entirely enclosed within the "range of triviality," -3 to +3, then we retain the loose null hypothesis that the two therapies are equivalent. What if the entire confidence interval is outside the range of triviality? We would then conclude that there is a nontrivial difference between the therapies. If part of the confidence interval is within the range of triviality and part outside the range, then we suspend judgment and wish we had obtained more data and/or less error variance. Of course, if the confidence interval extended into the range of triviality but not all the way to the point of no difference then we would probably want to conclude that there is a difference but confess that it might be trivial. See “Study 6” in Blume et al (2018).

Psychologists often use instruments which produce measurements in units that are not as meaningful as pounds and inches. For example, suppose that we are interested in studying the relationship between political affiliation and misanthropy. We treat political affiliation as dichotomous (Democrat or Republican) and obtain a measure of misanthropy on a 100 point scale. The point null is that mean misanthropy in Democrats is exactly the same as that in Republicans. While this hypothesis is highly unlikely to be true, it could be very close to true. Can we construct a loose null hypothesis, like we did for the anorexia therapies? What is the smallest difference between means on the misanthropy scale that we would consider to be nontrivial? Is a 5 point difference small, medium, or large? Faced with questions like this, we often resort to using standardized measures of effect sizes. In this case, we could use Cohen's *d*, the standardized difference between means. Suppose that we decide that the smallest difference that would be nontrivial is $d = .1$. All we need to do is get our data and then construct a confidence interval for *d*. If that interval is totally enclosed within the range -.1 to .1, then we conclude that affiliates of the two parties are equivalent in misanthropy, and if the entire confidence interval is outside the range, then we conclude that there is a nontrivial difference between the parties.

So, how do we get a confidence interval for *d*? Regretfully, it is not as simple as finding the confidence interval in the raw unit of measure and then dividing the upper and lower limits by the pooled standard deviation. Because we are estimating both means and standard deviations, we will be dealing with noncentral distributions (see Cumming & Finch, 2001; Fidler & Thompson, 2001; Smithson, 2001). Iterative computations that cannot reasonably be done by hand will be required. There are, out there on the Internet, statistical programs designed to construct confidence intervals for standardized effect size estimates, but I think it unlikely that such confidence intervals will be commonly used unless and until they are incorporated in major statistical packages such as SAS, SPSS, BMDP, Minitab, and so on. I have, on my SAS Program

Page and my SPSS Program Page, programs for constructing confidence intervals for Cohen's *d*.

Steiger (2004) argues that when testing for close fit, the appropriate confidence interval for testing range hypotheses is a 100(1 - 2α) confidence interval. For example, with the traditional .05 criterion, use a 90% confidence interval, not a 95% confidence interval. His argument is that the estimated effect cannot be small in both directions, so the confidence coefficient is relaxed to provide the same amount of power that would be obtained with a one-sided test. I am not entirely comfortable with this argument, especially after reading the Monte Carlo work by Serlin & Zumbo (2001).

Example from a Correspondent

Correspondent

We performed a 2 x 2 mixed ANOVA with the within subject factor Stimuli (S1 vs. S2) and the between-subject factor Group (Fibromyalgia patients (*n* = 52) vs. healthy controls (*n* = 55). Our interest is to know if there is difference between FM patients and controls in the modulation of brain activity to the second stimuli of a pair of identical stimuli. We obtained the following results:

Main Effect of Stimuli $F(1, 105) = 66.419$; $p <. 001$
Interaction Stimuli x Group $F(1, 105) = 1.672$; $p = .199$
Effect of Group $F(1, 105) = 0.009$; $p = .925$

This is the comment of the referee: "It is always difficult to show that there is no difference between groups. One may always come up with a statistical power argument. This argument is not easily dismissed. I suggest that the authors calculate effect sizes, preferentially Cohen's *d* for independent samples, and report the associated 95% confidence intervals for the relevant terms. That way, readers have more information about whether there was "no evidence for a difference", or there was "evidence for no difference". The difference between both is critical here. This should also be addressed in the general discussion."

Karl

What you really want to do here is show that the difference between the two groups is so small that it might as well be zero. This is exactly what is done when a generic drug manufacturer wants to demonstrate that its generic drug is "biologically equivalent" to the brand name drug. Statistical tests of such bioequivalence have been around for a good while, but are not well known by psychologists. See my document Tests of Equivalence and Confidence Intervals for Effect Sizes.

From the ANOVA stats you provided above, the main effect of Group is clearly the effect for which it can most convincingly be argued that the population effect is close to zero. SAS code for constructing confidence intervals for the standardized difference between two independent means can be found at my SAS Programs Page, under

"Confidence Interval for *d*, Two or More Independent Samples." If you are an SPSS user, see CI-d-SPSS.zip -- Construct Confidence Interval for Cohen's *d*.

You probably will not be able to make as strong a case for the population interaction being very close to zero. It would be possible to use contrast coefficients to code this interaction and then standardize that contrast and put a confidence interval about it. There is, however, a more straightforward method. First you should know that the *F* test of the interaction term in a 2 x 2 Mixed ANOVA is mathematically identical to an independent samples *t* test comparing the two groups on difference scores comparing post with pre (or, in your case, Stimulus 2 with Stimulus 1. This is demonstrated in my document The Pretest-Posttest x Groups Design: How to Analyze the Data.

Given this background, is it now easy to construct *d* with a confidence interval for the interaction term. First, compute the difference score for every case. Second, conduct a pooled-variances independent samples *t* test comparing the groups on those difference scores. You don't even need to actually conduct that *t* test. Simply take the square root of the interaction *F* from the ANOVA, and use the ANOVA error *df* as the *df* for the *t*. From the value of the *t*, you can use one of my programs to get estimated *d* and the confidence interval.

References

- Blume, J. D., D'Agostino McGowan, L., Dupont, W. D., & Greevey, R. A., Jr. (2018). Second-generation p-values: Improved rigor, reproducibility, & transparency in statistical analyses. PLoS ONE 13(3): e0188299.
- Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*, 532-574.
- Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed- and random-effects effect sizes. *Educational and Psychological Measurement*, *61*, 575-604.
- Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement*, *61*, 605-532.
- Serlin, R. C., & Zumbo, B. D. (2001). Confidence intervals for directional decisions. Retrieved from http://edtech.connect.msu.edu/searchaera2002/viewproposaltext.asp?propID=2678 on 20. February 2005.
- Steiger, J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, *9*, 164-182. Retrieved from http://www.statpower.net/Steiger%20Biblio/Steiger04.pdf on 20. February, 2005.

Return to Wuensch's Stats Lessons Page

This document most recently revised on 21-January-2019.