Equivalence Testing

- Consider the common hypothesis test...

\[ H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_A : \mu_1 \neq \mu_2 \]

...and it turns out that the client is content or even happy when the null hypothesis is NOT rejected.

- This should send-up a red flag in terms of the analysis that was done.

In reality, did the researcher NOT want to find significance?

Were they trying to show that the groups were the same?

When we don’t reject, what can be said about the group means?
The Dilemma of the Non-rejected Null

• Fail to reject $\Rightarrow$ There is not statistically significant evidence that the population means are different.

Without more information, we usually consider these groups similar. But this is under the idea that you’re looking for evidence that they are different.

If the researcher wants to claim that they’re ‘similar’, it’s not enough to do the traditional hypothesis test and just ‘not reject’. (If that was the case, just sample 2 people from each group... voila! They’re the same!)

• Under our traditional hypothesis testing, we assume the null is true right from the start. (If you want to SHOW that the null is true, we certainly can’t ASSUME it to be true.)
• If you want to show that the groups are similar, first **ASSUME** that they are different, and then try to gather evidence to the contrary (i.e. evidence that suggests they are the same).

⇒ This is Equivalence Testing

\[ H_0 : \mu_1 \neq \mu_2 \text{ vs. } H_A : \mu_1 = \mu_2 \]

• The difficult question in these tests...

“How close is close enough to be considered ‘the same’”? 

Can we specify a ‘Practical Equivalence’ value?

\[ H_0 : |\mu_1 - \mu_2| > \Delta \text{ vs. } H_A : |\mu_1 - \mu_2| < \Delta \]

\[ \text{Inequivalent} \quad \text{Equivalent} \]
• In equivalence testing, the null hypothesis is a “difference of $\Delta$ or more.”

Restating $H_0$...

$$H_0 : \mu_1 - \mu_2 < -\Delta \quad \text{or} \quad \mu_1 - \mu_2 > \Delta$$

• This leads to the most basic form of equivalence testing, the two one-sided test (TOST) procedure.

$$\frac{(\bar{y}_1 - \bar{y}_2) + \Delta}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > z_{1-\alpha} \quad \text{or} \quad \frac{(\bar{y}_1 - \bar{y}_2) - \Delta}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < -z_{1-\alpha}$$

We declare the two group means equivalent at the $\alpha$ level if, and only if, both are rejected.
• Confidence interval (CI) approach:

Construct the \((1-2\alpha)\%\) CI for the difference in means: \(\bar{Y}_1 - \bar{Y}_2 \pm z_{\alpha} \cdot se(\bar{Y}_1 - \bar{Y}_2)\)

If both one-sided tests are simultaneously rejected, this CI will be contained in the \(\pm \Delta\) interval.

If the CI for the difference is completely contained in the interval with endpoints \(-\Delta\) and \(+\Delta\), then we declare equivalence.

As opposed to classical testing, the researcher WANTS this confidence interval to contain zero.

We want to be able to say the difference is very likely to be zero (beyond random chance).

• In the cheese example, the client chose \(\Delta\) as 2 units (on a 20 point scale).
Even without accounting for multiple testing, none of the new cheeses could be declared equivalent to the milk fat cheese at the $\alpha = 0.05$ level (i.e. controlling the probability of rejecting inequivalence when it was actually true at the 0.05 level).
This procedure was used to compare vaccination rates (as a percentage) among different races/ethnicities and different vaccines in Barker, et al. (2002).

**Figure.** Equivalencies in early childhood immunization coverage (as %) by self-reported race/ethnicity for minority groups versus Whites, National Immunization Survey, 2000. (The researchers have chosen $\Delta$ as 5%).

• When has this come up?
  – New food item meant to be a substitute
  – New generic drug compared to old standard (bioequivalence)

• This process makes more sense logically because more samples gives us more power for detecting ‘equivalence’.

• It may be a subtle difference to pick-up from your client, but sometimes an equivalence test is more appropriate.
I performed an equivalence test for submission for a biology collaborator.

Here is a plot of the length of the Organ of Corti (part of the inner ear) for two different mouse genotypes (wild type and double-conditional knock-out).

They are obviously quite close within each time point, but can we consider them equivalent?
Confidence interval on the difference: Simultaneous confidence intervals
(no multiple comparison adjustment)

Simultaneous confidence intervals (with bonferroni multiple comparison adjustment)
Some references:


