The "NHST Controversy" & Confidence Intervals

- The controversy
- A tour through the suggested alternative solutions
  - Ban NHST
  - Retain NHST as-is
  - Augment NHST
- How meta-analysis relates to this issue
- Confidence intervals (single means, mean differences & correlations)
- Confidence intervals & significance tests

The "NHST Controversy"

- For as long as there have been NHSTing there has been an ongoing "dialogue" about its sensibility and utility.
- Recently this discussion has been elevated to a "controversy" -- with three "sides" ...
  - those who would eliminate all NHSTing
  - those who would retain NHSTing as the centerpiece of research data analysis (short list & hard to tell from …)
  - those who would improve & augment NHSTing
- Results of this "controversy" have included ...
  - hundreds of articles and dozens of books
  - changes in the publication requirements of many journals
  - changes in information required of proposals by funding agencies

Let's take a look at the two most common positions…

Ban the NHST…

- the "Nill Null" is silly and never really expected
  - the real question is <u>not</u> whether there is a relationship (there almost certainly is) but whether it is large enough to "care about" or "invest in"
  - Nil-null HNST it <u>mis</u>represents the real question of "how large is the effect" as "whether or not there is an effect"
- NHST has been used so poorly for so long that we should scrap it and replace it with "appropriate statistical analyses"

What should we do… (will just mention these -- more to come about each)

- effect size estimates (what is the size of the effect)
- confidence intervals
- NHST using "non-nill nulls"

Keep NHST, but do it better and augment it …

Always perform power analyses (more about actually doing it later)

- Most complaints about NHST mistakes are about Type II errors (retaining H0: there there is a relationship between the variables in the population)
- Some authors like to say "64% of NHST decisions are wrong"
  - 5% of rejected nulls (using p = .05 criterion, as expected)
  - another 59% from Type II errors directly attributable to using sample sizes that are too small

Consider the probabilities involved

- if reject H0: consider the chances it is a Type I error (p)
- if retain H0: consider the chances is it a Type II error (more later)

Consider the effect size, not just the NHST (yep, more later…)

- how large is the effect and is that large enough to "care about" or "invest in"

Consider Confidence intervals (more later, as you could guess…)

- means, mean differences and correlations are all "best guesses" of the size of the effect
- NHST are a guess of whether or not they are "really zero"
- CIs give information about the range of values the "real" population mean, mean difference or r might have

Consider Non-Nill NHST

- it is possible to test for any "minimum difference", not just for "any difference greater than 0"
- there are more elegant ways of doing it but you can…
- if H0: is "TX will improve performance by at least 10 points" ...
  - just add 10 to the score of everybody in the Cx group
- if H0: is "correlation is at least .15" …
  - look up r-critical for that df, and compare it to r - .15

Another "wave" that has hit behavioral research is "meta analysis"

- meta analysis is the process of comparing and/or combining the effects of multiple studies, to get a more precise estimate of effect sizes and likelihood of Type I and Type II errors
- meta analysts need "good information" about the research they are examining and summarizing, which has led to some changes about what journals ask you to report…
  - standard deviations (or variances or SEM)
  - sample sizes for each group (not just overall)
  - exact p-values
  - MSe for ANOVA models
  - effect sizes (which is calculable if we report other things)
- by the way -- it was the meta analysis folks who really started fussing about the Type II errors caused by low power -- finding that there was evidence of effects, but nulls were often retained because the sample sizes were too small

Confidence Intervals

Whenever we draw a sample and compute an inferential statistic, that is our best estimate of the population parameter.

However, we know two things:

the statistic is unlikely to be exactly the same as the parameter

we are more confident in our estimate the larger our sample size

Confidence intervals are a way of "capturing" or expressing our confidence that the value of the parameter of interest is within a specified range.

That's what a CI tells you -- starting with the statistics drawn from the sample, within in what range of values is the related population parameter how likely to be.

There are 3 types of confidence intervals that we will learn about…

1. confidence interval around a single mean
2. confidence interval around a mean difference
3. confidence interval around a correlation

CI for a single mean

Gives us an idea of the precision of the inferential estimate of the population mean

- don't have to use a 95% CI (50%, 75%, 90% & 99% are also fairly common

Eg. … Your sample has a mean age = 19.5 years, a std = 2.5 & a sample size of n=40

50% CI CI(50) = 19.5 +/- .268 = 19.231 to 19.768

We are 50% certain that the real population means is between 19.23 and 19.77

95% CI CI(95) = 19.5 +/- .807 = 18.692 to 20.307

We are 95% certain that the real population means is between 18.69 and 29.31

99% CI CI(99) = 19.5 +/- 1.087 = 18.412 to 20.587

We are 99% certain that the real population means is between 18.41 and 20.59

Notice that the CI must be wider for us to have more confidence.

It is becoming increasingly common to include "whiskers" on line and bar graphs. Different folks espouse different "whiskers" …

- standard deviation -- tells variability of population scores around the estimated population mean
- SEM -- tells the variability of sample means around the true population mean

CI -- tells with what probability/confidence the population is within what range/interval around the estimate from the sample

Things to consider…

- SEM and CI, but not std, are influenced by the sample size
- The SEM will always be smaller ("look better") than the std
- 1 SEM will be smaller than CI
  - but 2 SEMs is close to 95% CI (1.96*SEM = 95% CI)
- Be sure your choice reflects what you are trying to show
  - variability in scores (std) or sample means (SEM) or confidence in population estimates estimate (CI)

CI for a mean difference (two BG groups or conditions)

Gives us an idea of the precision of the inferential estimate of the mean difference between the populations.

- Of course you'll need the mean from each group to compute this CI!
- You'll also need either…

The Std and n for each group or the MSerror from the ANOVA

Eg. … Your sample included 24 females with a mean age of 19.37 (std = 1.837) & 18 males with a mean age of 21.17 (std = 2.307). Using SPSS, an ANOVA revealed F(1,40) = 7.86, p = .008, MSe = 4.203

95% CI CI(95) = 1.8 +/- 1.291 = .51 to 3.09

We are 95% certain that the real population mean age of the females is between .47 lower than the male mean age and 3.09 lower than the male mean age, with a best guess that the mean difference is 1.8.

99.9% CI CI(99.9) = 1.8 +/- 2.269 = -.47 to 4.069

We are 99.9% certain that the real population mean age of the females is between .51 higher than the male mean age and 4.07 lower than the male mean age , with a best guess that the females have a mean age 1.8 years lower than the males.

Confidence Interval for a correlation

Gives us an idea of the precision of the inferential estimate of the correlation between the variables.

- You'll need just the correlation and the sample size
- One thing – correlation CIs are not symmetrical around the r-value, so they are not expressed as " r +/- CI value"

Eg. … Your student sample of 40 had a correlation between age and #credit hours completed of r = .45 (p = .021).

95% CI CI(95) = .161 to .668

We are 95% certain that the real population correlation is between .16 and .67, with a best estimate of .45.

99.9% CI CI(99.9) = -.058 to .773

We are 99.9% certain that the real population correlation is between -.06 and .77, with a best estimate of .45.

NHST & CIs

The 95% CI around a single mean leads to the same conclusion as does a single-sample t-test using p = .05 …

- When the 95% CI does not include the hypothesized population value the t-test of the same data will lead us to reject H0:
  - from each we would conclude that the sample probably did not come from a population with the hypothesized mean
- When the 95% CI includes the hypothesized population value the t-test of the same data will lead us to retain H0:
  - from each we would conclude that the sample might well have come from a population with the hypothesized mean

1-sample t-test & CI around a single mean

From the earlier example -- say we wanted a sample from a population with a mean age of 21

1-sample t-test

- with H0: = 21, M=19.5, std = 2.5, n = 41
  - t = (21 - 19.5) / .395 = 3.80
- looking up t-critical gives t(40, p=.05) = 2.02
- so … reject H0: and conclude that this sample probably did not . come from a pop with a mean age less than 21

CI around a single mean

- we found 95% CI = 19.5 +/- .807 = 18.692 to 20.307
- because the hypothesized/desired value is outside the CI, we would conclude that the sample probably didn't come from a population with the desired mean of 21

Notice that the conclusion is the same from both "tests" -- this sample probably didn't come from a pop with a mean age of 21

BG ANOVA & CI around a mean difference

Your sample included 24 females with a mean age of 19.37 (std = 1.837) & 18 males with a mean age of 21.17 (std = 2.307).

BG ANOVA

- F(1,40) = 7.86, p = .008, MSe = 4.203
- so … reject H0: and conclude that the populations of men and women have different mean ages

CI around a mean difference

- we found 95% CI = 1.8 +/- 1.291 = .51 to 3.09
- because a mean difference of 0 is outside the CI, we would conclude that the populations of men and women have different mean ages

Notice that the conclusion is the same from both "tests" – these sample probably didn't come from populations with the same mean age

r significance test & CI around an r value

Your student sample of 40 had a correlation between age and #credit hours completed of r = .45 (p = .021).

r significance test

- p < .05, so would reject H0: and conclude that variables are probably correlated in the population

CI around an r-value

- we found 95% CI = .161 to .668
- because an r-value of 0 is outside the CI, we would conclude that there probably is a correlation between the variables in the populations

Notice that the conclusion is the same from both "tests" – these variables probably are correlated in the population