The "NHST Controversy"– Confidence Intervals, Effect Sizes & Power Analyses

- 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
- Effect size estimates for correlation, ANOVA & Chi-square
- Power Analyses – *a priori* & *post hoc*
- Alternatives to Power Analysis
- Considering "Stability" in addition to power
- Putting it all together NHST+

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"
  - 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

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 experts with a mean age of 19.37 (std = 1.837) & 18 novices 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 novices is between .47 lower than the novice mean age and 3.09 lower than the novice 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 experts is between .51 higher than the novices mean age and 4.07 lower than the novice mean age , with a best guess that the experts have a mean age 1.8 years lower than the novices.

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 experts with a mean age of 19.37 (std = 1.837) & 18 novices 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 novices and experts 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 novices and experts 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

Effect Size and Statistical Significance - two useful pieces of info

Statistical Significance Test (Summary) Statistic (t, F and χ^2)

- used primarily as an intermediate step to obtain the p-value for the statistical decision
- the p-value is used to decide "whether or not there is an effect"

Effect size refers to

- the strength or magnitude of the relationship between the variables in the population.
- the extent of departure from the H0: (no relationship)

Their relationship

Significance Test Stat = Effect Size * Size of Study

Effect Size = Significance Test Stat / Size of Study

This formula/relationship tells us

- for any given nonzero effect size, the value of the test statistic (e.g., t, F, X^2) will increase as does the sample size (N)
- for any nonzero effect size, increase in the effect size OR increase in the value of the test statistic will result in a lower p-value, and greater confidence that the population effect size is nonzero

We want to have estimates of effect size/strength that are separable from our inferential test statistic. The key will be to compose these estimates so that the value of the estimate is independent of the size of the study (N).

When we use correlation, **r** is both a summary statistic and an effect size estimate.

- For any given N, df = N-2, and we can look up the critical-r value and decide whether to retain or reject H0:
- Also, we know that the larger **r** is (+ or -), then the stronger is our estimate of the linear relationship between the variables in the population
  - with practice we get very good at deciding whether **r** is "small" (r = .10), "medium" (.30) or "large" (.50)
- We can compare the findings of different studies by comparing the **r** values they found.

Thinking about Effect Sizes, Power Analyses & Significance Testing with Pearson's Correlation

- Dr. Yep correlates the # hours students studied for the exam with % correct on that exam and found r(48) = .30, p. < .05).
- Dr. Nope "checks-up" on this by re-running the study with N=20 finding a linear relationship in the same direction as was found by Dr. Yep, but with r(18) = .30, p > .05).

What's up with that ???

Consider the correlations (effect sizes) … .30 = .30

But, consider the power for each

Dr. Yep -- we know we have "enough power", we rejected H0:
Dr. Nope -- r = .30 with S = 20, power is < .30, so more than a 70% chance of a Type II error

Same correlational value in both studies -- but different H0: conclusions because of very different amounts of power (sample size).

But what if we want to compare the results from studies that used different analyses (because they used quant vs. qual variables)??

- We know we can only compare F-values of studies that have the same sample sizes (Test Stat = Effect Size * Size of Study)
- We know we can only compare X^2-values of studies that have the same sample sizes (Test Stat = Effect Size * Size of Study)
- We can't compare studies that did F-tests with those that did X^2-tests and can't compare either with studies that used r

Unless of course, we had some generalized "effect size measure" that could be computed from all of these statistical tests…
We do ... our old buddy **r**, which can be computed from F or X^2

$$\mathbf{r} = \sqrt{F / (F + df_{error})} \quad \text{and} \quad \mathbf{r} = \sqrt{X^2 / N}$$

By the way, when used this way "r" is sometimes called η (eta).

Also, you want to be sure to distinguish between r/η and r^2/η^2

Now we can summarize and compare the effect sizes of different studies.
Here's an example using two versions of a study using ANOVA...

Researcher #1 Acquired 20 computers of each type, had researcher assistants (working in shifts & following a prescribed protocol) keep each machine working continually for 24 hours & count the number of times each machine failed and was re-booted.

Mean failures PC = 5.7
Mean failures Mac = 3.6
F(1,38) = 10.26, p = .0004

$\sqrt{F / (F + df)} = \sqrt{10.26 / (10.26+38)}$
r = .46

Researcher #2 Acquired 30 computers of each type, had researcher assistants (working in shifts & following a prescribed protocol) keep each machine working continually for 24 hours & measured the time each computer was running.

Mean up time PC = 22.89
Mean up time Mac = 23.48
F(1,58) = 18.43, p = .001

$\sqrt{F / (F + df)} = \sqrt{18.43 / (18.43+58)}$
r = .49

So, we see that these two studies found very similar results – similar → effect direction (Macs better) & effect size !!

Now we can summarize and compare the effect sizes of different studies.
Here's an example using two versions of a study using X^2...

Researcher #1 Acquired 40 computers of each type, had researcher assistants (working in shifts & following a prescribed protocol) keep each machine working continually for 24 hours or until the statistical software froze.

	PC	Mac
Failed	11	3
Not	29	37

$X^2(1) = 5.54$, $p = .03$

$\sqrt{X^2 / N} = \sqrt{5.54 / 80}$

$r = .26$

Researcher #2 Acquired 20 computers of each type, had researcher assistants (working in shifts & following a prescribed protocol) keep each machine working continually for 24 hours or until the graphic editing software froze.

	PC	Mac
Failed	8	3
Not	5	6

$X^2(1) = 1.69$, $p = .193$

$\sqrt{X^2 / N} = \sqrt{1.69 / 22}$

$r = .28$

So, by computing effect sizes, we see that the same effects were found in the two studies – the difference in terms of p-value & "significance" was due to sample size!

What about if we want to compare results from studies if one happened to use a quantitative outcome variable and the other used a "comparable" qualitative outcome variable?

We know we can't only F & X^2 -values from different studies, especially if they have different sample sizes
(Test Stat = Effect Size * Size of Study)

Unless of course, we had some generalized "effect size measure" that could be computed from both F and X^2 s using different DVs & Ns…

We do ... our old buddy **r**, which can be computed from F & X^2

$\mathbf{r} = \sqrt{F / (F + dferror)}$ $\qquad$ $\mathbf{r} = \sqrt{X^2 / N}$

Now we can summarize and compare the effect sizes of different studies.

Here's an example using two versions of a study we discussed last time...

Researcher #1 Acquired 20 computers of each type, had researcher assistants (working in shifts & following a prescribed protocol) keep each machine working continually for 24 hours & count the number of times each machine failed and was re-booted.

Mean failures PC = 5.7, std = 2.1

Mean failures Mac = 3.6, std = 2.1

$F(1,38) = 10.26$, $p = .003$

$\sqrt{F / (F + df)} = \sqrt{10.26 / (10.26+38)}$

$r = .46$

Researcher #2 Acquired 20 computers of each type, had researcher assistants (working in shifts & following a prescribed protocol) keep each machine working continually for 24 hours or until it failed.

	PC	Mac
Failed	15	6
Not	5	14

$X^2(1) = 8.12$, $p < .003$

$\sqrt{X^2 / N} = \sqrt{8.12 / 40}$

$r = .45$

So, by computing effect sizes, we see that these two studies found very similar results, in terms of direction and effect size !!

Families of Effect Size Estimates

r -- variations on the correlation coefficient

-- η is a common variation (range = 0 to 1.0)

r^2 -- variations on the "shared variance" statistics (e.g., η^2)

ω/ ω^2 -- variations on the omega(^{2}) statistic that attempt to correct for the likelihood of overestimating the strength of the population effect size with a large sample

-- have their greatest popularity with ANOVA-types

d -- an index of effect size in terms of the size of the mean difference between two groups expressed as the propotion of a standard deviation (most applicable to analyzes comparing means using t-test & ANOVA)

Computing Effect Size Estimates -- We will focus on the r^2, r, and d estimates (the most common, especially in meta-analysis)

r^2(r) is the most common and most generalizable --

$r^2 = t^2/ (t^2 + df)$ (df = N - 2)

$r = \sqrt{[t^2/ (t^2 + df)]}$

$r^2 = F / (F + df_{error})$ (df_{error} = N - 2)

$r = \sqrt{[F / (F + df_{error})]}$ (2-group designs)

$r^2 = X^2 / N$ (when df = 1)

$r = \sqrt{[X^2 / N]}$

r^2 (proportion of shared variance) "versus" r (size of relationship)

Two "warring camps" (and only part of the argument)

r^2 many psychological "effects" are small (e.g., "significant" clinical effects have typical r^2 = .06) and probably have little impact on daily life, mental health, etc.

r some small effects are very meaningful (r^2 = .04 in a study of jury decision bias means 10 fewer "innocent" people sentenced to death per year)

Keep in mind…

since $r = \sqrt{r^2}$ the discussion is not about the "math" but the "accuracy of representation" … which "expression" will lead to the most people having the "best" understanding of the meaningful size of the effect ??

Computing Effect Size Estimates, cont.

$d = (M_1 - M_2) / s_{pooled}$ (s_{pooled} = pooled std dev)

$$S_{pooled} = \sqrt{\left(\frac{[(n_1 - 1) * S^2_1] + [(n_2 - 1) * S^2_2]}{n_1 * n_2} \right)}$$

n_1 & n_2 = sample sizes

S^2_1 & S^2_2 = sample variances (std²)

$S_{pooled} = \sqrt{MS_{error}}$ MS_{error} is " Within Groups Mean Squares" in SPSS output

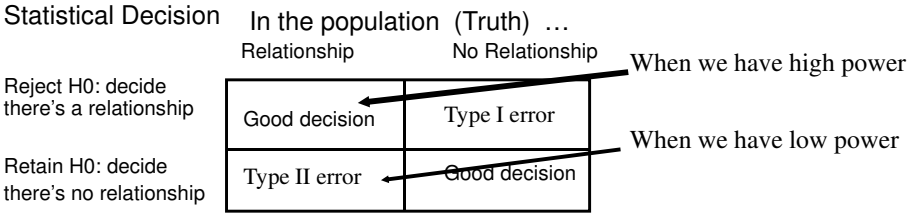
$d = 2t / \sqrt{df}$ (equal-n formula rem: $t = \sqrt{F}$)

$d = [t * (n_1 + n_2)] / [\sqrt{df} * \sqrt{(n_1 * n_2)}]$ (unequal-n)

Again, d is the mean difference between the groups expressed as a proportion of the (pooled) standard deviation

Just a bit of review before discussing Power analysis

Statistical Power (also called sensitivity) is about the ability to reject H0: based on the sample data when there REALLY IS a correlation between the variables in the population



Statistical Power is increased by…

- larger effect (i.e., larger r between the variables)
- larger sample size

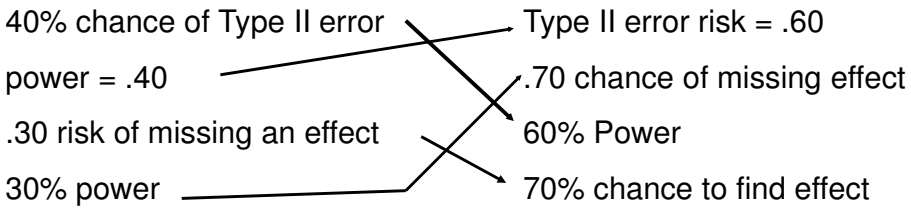
Statistical Power

- The ability to Reject H0: based on the sample data when there really is a correlation between the variables in the population
- Statistical Power is primarily about the sample size needed to detect an "r" of a certain size with how much confidence !!
- Statistical Power tell the probability of rejecting H0:, when it should be rejected.
- On the "next after" page is a "power table" we use for ...
- Two kinds of Power Analyses
  - *a priori* power analyses are used to tell the what the sample size should be to find a correlation of a specified size
  - *post hoc* power analyses are used when you have retained H0:, and want to know the probability that you have committed a Type II error (to help you decide whether or not you "believe" the null result).

But first -- a few important things…

• Power analysis is about Type II errors, "missed effects"
"retaining H0: when there really is a relationship in the population!!
• "Power" is the antithesis of "risk of Type II error"
  • Risk of Type II error = 1 - power
  • Power = 1 - Risk of Type II error

match up the following...



Here's the power table we'll use most often…

Power, Effect Size & Sample Size*

r ? / ? power	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70
.20	124	32	21	15	14	13	11	9	7	5			
.30	208	93	53	34	24	18	14	11	9	8	7	6	5
.40	296	132	74	47	33	24	19	15	12	10	8	7	6
.50	382	170	95	60	42	30	23	18	14	12	9	8	7
.60	488	257	143	90	62	45	34	24	20	16	13	11	9
.70	613	300	167	105	72	52	39	29	23	28	15	12	10
.80	781	343	191	120	82	59	44	33	26	20	16	13	11
.90	1045	459	255	160	109	78	58	44	34	27	21	17	13

* "**S**" values given for $\alpha = .05$

Values taken from (Friedman, 1982 & Cohen, 1988), with some interpolation.

a priori Power Analyses -- r

You want to be able to reject H0: if r is as large as .30

- pick the power you want
  - probability of rejecting H0: if there is a relationship between the variables in the population (H0: is wrong)
  - .80 is "standard" -- 80% confidence will reject H0: if there's an effect
- go to the table
  - look at the column labeled .30 (r = .30)
  - look at the row labeled .80 (power = .80)
  - you would want S = 82
- What about… necessary sample size (S)
  - r = .40 with power = .90 ???
  - r = .15 with power = .80 ???
  - r = .20 with power = .70 ???

The *catch* here is that you need some idea of what size correlation you are looking for!!! Lit review, pilot study, or "small-medium-large" are the usual solutions -- but you must start *a priori analyses* with an expected r !!!

How do you really do an *a priori* Power Analysis ???

The basis for a worthwhile *a priori* power analysis is a good set of effect size estimates – one for each of the pairwise comparisons needed to test the RH: (especially for the smallest effect we want to "chase" !)

But from where do we get the estimates?

Most studies are a combination of replication comparisons and new comparisons

- get the effects sizes for the replication comparisons from the lit
- get the effects sizes for the new comparisons indirectly …
  - do you expect your new conditions to yield larger or smaller pairwise effects than the replications? How much so ?
  - use the std or MSerror from earlier studies to help compute r

How do you really do *a priori* Power Analyses ???

Example

Two conditions in the study are replications – one is new
- based on lit rev we expect means of Cx = 30 & TxOld = 50
- that lit also shows std for these conditions ≈ 20
- we expect our TxNew to have a mean of about 60

The smallest mean dif ➔ smallest pairwise effect size
- for TxOld (50) vs. TxNew (60)
- comp r using MSerror = std^2 (20^2 = 400) giving r = .24

Now we can do the a priori power analysis
- with r = .25 and 80% power S = 120
- for each of the 2 conditions n = S / 2 = 120 / 2 = 60
- for the whole study N = n * k = 60 * 3 = 180

With enough power for this smallest effect, we'll have ample power for the other larger effects.

post hoc Power Analyses -- r

You obtained r(30)=.30, p > .05, and decided to retain H0:
- What is the chance that you have committed a Type II error ???
- Compute S = df + 2 = 30 + 2 = 32
- go to the table
  - look at the column labeled r = .30
  - look down that column for S = 32 ➔ 24/33
  - read the power from the left-most column (.30-.40)
- Conclusion?
  - power of this analysis was .30-.40
  - probability that this decision was a Type II error (the probability we missed an effect that really exists in the population) = 1 - power = 60-70%
  - NOT GOOD !! If we retain H0: there's a 60-70% chance we're wrong and there really is a relationship between the variables I the population We shouldn't trust this H0: result !!

post hoc "vs." *a priori* power -- big enough sample?!?

Four analyses from the same study (n = 21) …

	Informal power analysis	*post-hoc* power for this study		*a priori* power for next study
r =.55, p<05	"enough power"	>.90 from S=42		S = 20 for .80
r =.30, p<.05	"enough power"	≈.50 from S=42	!!!	S = 82 for .80
r =.20, p>.05	"not enough power"	≈.27 from S=42		S = 191 for .80
r =.02, p>.05	"not power problem"	<.01 from S=42	!!!	S>3000 for .80

Caveats:

"Enough" post-hoc N might not be "enough" a priori N !!!

How small of an effect can you afford to "chase"??

Power analysis with r is simple, because…

- r is the "standard" effect size estimate used for all the tests
- the table uses r

- when working with F and X^2 we have to "detour" through r to get the effect sizes needed to perform our power analyses
  - here are the formulas again

$$\mathbf{r} = \sqrt{F / (F + df_{error})} \quad \text{and} \quad \mathbf{r} = \sqrt{X^2 / N}$$

- as with r, with F and X^2
  - we have a priori and post how power analyses
  - for a priori analyses we need a starting estimate of the size of the effect we are looking for

post hoc Power Analyses -- F

You obtained F(1, 28) = 3.00, p > .05, and decided to retain H0:

- What is the chance that you have committed a Type II error ???
- Compute $r = \sqrt{F / (F + df_{error})} = \sqrt{3 / (3 + 28)} = .31$
- Compute $S = df_{error}$ + #IV cond = 28 + 2 = 30
- go to the table
  - look at the column labeled .30 (closest to r = .31)
  - look down that column for S = 30 (33 is closest)
  - read the power from the left-most column (.40)
- Conclusion?
  - power of this analysis was .40
  - probability that this decision was a Type II error (the probability we missed an effect that really exists in the population) = 1 - power = 60% -- NOT GOOD !! We won't trust this H0: result !!

What if you plan to replicate this study -- what sample size would you want to have power = .80? What would be your risk of Type II error?

S = 82 - 41 in each cond. Type II error Risk = 20%

post hoc Power Analyses -- X^2

You get $X^2(1) = 3.00$, $p > .05$ based on N=45, and decided to retain H0:

- What is the chance that you have committed a Type II error ???
- Compute $r = \sqrt{X^2 / N} = \sqrt{3 / 45} = .26$
- Compute S = N = 45
- go to the table
  - look at the column labeled .26
  - look down that column for S = 45 (33 is closest)
  - read the power from the left-most column (.40)
- Conclusion?
  - power of this analysis was .40
  - probability that this decision was a Type II error (the probability we missed an effect that really exists in the population) = 1 - power = 60% -- NOT GOOD !! We won't trust this H0: result !!

What if you plan to replicate this study -- what sample size would you want to have power = .80? What would be your risk of Type II error?

S = 120 - 60 in each cond. Type II error Risk = 20%

Now we can take a more complete look at types of statistical decision errors and the probability of making them ...

Statistical Decision	In the Population: H0: True	In the Population: H0: False
Retain H0:	Correctly Retained H0: Probability = 1 - α	Incorrectly Retained H0: Type II error Probability = β
Reject H0:	Incorrectly Rejected H0: Type I error Probability = α	Correctly Rejected H0: Probability = 1 - β

How this all works …

Complete stat analysis and check the p-value

If reject H0: …

- Type I & Type III errors possible
- p = probability of Type I error
- Prob. of Type III error not estimable
- MUST have had enough power (rejected H0: !)

If retain H0:

1. Need to determine prob. of Type II error
   - Compute effect size → r
   - Compute S
   - Determine power
   - Type II error = 1 - power
2. Likely to decide there's a power problem -- unless the effect size is so small that even if significant it would not be "interesting"

Applyiung these probabilities !!

Imagine you've obtained r(58) = .25, p = .05

If I decide to reject H0:, what's the chance I'm committing a Type I error ?

This is α (or p) = 5%

If I decide to reject H0:, what's the chance I'm committing a Type III error ?

"not estimable"

If I decide to reject H0:, what's the chance I'm committing a Type II error ?

0% -- Can't possibly commit a Type II error when you reject H0:

If I decide to retain H0:, what's my chance of committing a Type I error ?

0% -- Can't commit a Type I error when you retain H0:

If I decide to retain H0:, what's my chance of committing a Type III error ?

0% -- Can't commit a Type III error when you retain H0:

If I decide to retain H0:, what's the chance I'm committing a Type II error ?

For r =.25, S=60, power = 50%
So I have a 50% chance of Type II error

Alternatives to Power Analyses

"Rules of Thumb"

• usually based on the idea that "if you can't find a significant effect with "this sample size", then the effect probably isn't large enough to care about

• most common in areas that don't use effect sizes or power analysis – when you do these, you often discover that the rule " works" → common effect sizes for that area are significant using that sample size

• so usually work well -- within their research area on well-known phenomena (design, task/stim & DV combinations)!!!

  • but be careful about "transplanting" rules-of-thumb across content areas or to new phenomena

Alternatives to Power Analyses, cont.

"Selecting S for significance"

• estimate the pairwise effect size, say r = .35

• using the correlation critical-value table, select a sample size for which that effect size will be significant

• r = .35 will be significant if df = 30 or S=32

Partial critical-r Table

df	α = .05
20	.42
25	.38
30	.35
35	.33
40	.30
45	.29
50	.27
60	.25

What's the power of this sample size ??

For r = .35 & S=30,

Power is only 50%

So, this approach leads to very low power !

r → **↓ power**	**.35**
.20	13
.30	18
.40	24
.50	30
.60	45
.70	52
.80	59
.90	78

Why do these two approaches differ so much ?

The difference in "suggested S" is because the power analysis takes into account that the r-value of a sample drawn from a population with r = .361 might, by chance, be smaller than .361 !!!

Remember that we are testing RH: and making inferences about the population correlation !!!!

So, we want to be able to correctly decide that there is a correlation in the population (i.e., reject H0:), even if the sample we happen to draw has a smaller r-value than the population.

By the way…

For a given r → the sample size for 80% power is about 2X the sample size for which that r will be significant (p = .05)

NHST Power "vs." Parameter estimate stability

NHST power → what's the chances of rejecting a "false null" vs. making a Type II error?

Statistical power is based on…

- size of the effect involved ("larger effects are easier to find")
- amount of power (probability of rejecting H0: if effect size is as expected or larger)

Stability → how much error is there in the sample-based estimate of a parameter (correlation, regression weight, etc.) ?

Stability is based on …

- "quality" of the sample (sampling process & attrition)
- sample size

Std of r = 1 / √ (N-3), so …

N=50 r +/- .146 N=100 r +/- .101 N=200 r +/- .07
N=300 r +/- .058 N=500 r +/- .045 N=1000 r +/- .031

The power table only tells us the sample size we need to reject H0: r=0!! It does not tell us the sample size we need to have a good estimate of the population r !!!!!

Partial Power Table (taken & extrapolated from Friedman, 1982)

r	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70
power												
.30	93	53	34	24	18	14	11	9	8	7	6	5
.40	132	74	47	33	24	19	15	12	10	8	7	6
.50	170	95	60	42	30	23	18	14	12	9	8	7
.60	257	143	90	62	45	34	24	20	16	13	11	9
.70	300	167	105	72	52	39	29	23	18	15	12	10
.80	343	191	120	82	59	44	33	26	20	16	13	11
.90	459	255	160	109	78	58	44	34	27	21	17	13

"Sufficient power" but "poor stability"

How can a sample have "sufficient power" but "poor stability"?
Notice it happens for large effect sizes!!
e.g., For a population with r = .30 & a sample of 100 …

- Poor stability of r estimate → +/- 1 std is .20-.40
- Large enough to reject H0: that r = 0 → power almost .90

So, what do you get out of all these analyses ???

effect size estimates

- mean -- most basic description/inference but…
  difference - DV scale can be difficult to generalize
  - does not account for variability around the means or sample size
- F-value -- integrates effect size, variability and sample size, but (without practice) is most useful to obtain p-value
- d, r, etc. -- tells "how big" is the effect considering variability, but without considering sample size/power - easy to interpret metrics (r & d), but tells nothing about the likelihood of α or β

assessing statistical conclusion error

- CI -- expresses mean difference taking variability and sample size (α) into account -- allows testing of non-nil H0: ("practical significance"
- p-value -- probability that a rejected H0: is a Type I error
- post-hoc power analysis - prob that a retained H0: is a Type II error