

Analyses of K-Group Designs : Power Analysis & Confidence Intervals

- Power analyses – n , S & N
- Power analyses for pairwise comparisons
- “Practical” *a priori* power analysis
- Alternative to power analysis & their problems
- Confidence Intervals
- Combining the information from r , power & CI

Power analysis is about the interrelationships among four things...

1. the size of the effect - r
2. the sample size of the study -- N
3. Type I Error Rate (False Alarm) -- α , usually .05
4. Type II Error rate (Miss) -- β , usually .20
 - remember that power is $1 - \beta$, usually set at .80

A priori power analyses

- conducted before the study is begun
- start with estimated effect size & desired power and determine the needed $S \rightarrow n \rightarrow N$

Post hoc power analysis

- conducted after retaining H_0 :
- start with obtained effect size & and determine β as $1 - \text{power}$ of the study

Power analyses are dependent upon having accurate effect size estimates !!!!

Determining the power you need ..

For a 2-condition design...

- the omnibus-F is sufficient -- retain or reject, you're done !
- you can easily determine the sample size needed to test any expected effect size with a given amount of power

For a k-condition design ...

- the power of the omnibus-F - isn't what matters !
- a significant omnibus-F only tells you that the “most different” means are significantly different
- follow-up (pairwise) analyses will be needed to test if the pattern of the mean differences matches the RH:
- you don't want to have a “pattern of results” that is really just a “pattern of differential statistical power”
- you need to assure that you have sufficient power for the smallest pairwise effect needed to test your specific RH:

a priori Power Analyses for **k-BG** designs

Important Symbols
 S is the total # of participants in that pairwise comp
 $n = S / 2$ is the # of participants in each condition of that pairwise comparison
 $N = n * k$ is the total number of participants in the study

Example
 We expect the three pairwise comparisons from our study to have effect sizes of about $r = .35$, $r = .40$ & $r = .25$

We should target the smallest effect in our power analysis

- for $r = .25$ and 80% power $S = 120$
- for each of the 2 conditions $n = S / 2 = 120 / 2 = 60$
- for the whole study $N = n * k = 60 * 3 = 180$

a priori Power Analyses for **k-WG** designs

Important Symbols
 S is the total # of participants in that pairwise comp
 For WG designs, every participant is in every condition, so...
 S is also the number of participants in each condition (n) and in the whole study (N) → $S = n = N$

Example
 We expect the three pairwise comparisons from our study to have effect sizes of about $r = .40$, $r = .60$ & $r = .45$ & want 90% power

We should target the smallest effect in our power analysis

- with $r = .40$ and 90% power $S = 58$
- for each condition of a WG design $n = S = 58$
- for the whole study $N = S = 58$

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

Four pairwise comparisons from the same study ($n = 21$) ...

	Informal power analysis	<i>post-hoc</i> power for this study	<i>a priori</i> power for next study
$r = .55, p < .05$	"enough power"	$> .90$ from $S=42$	$S = 20$ for $.80$
$r = .30, p < .05$	"enough power"	$\approx .50$ from $S=42$!!!	$S = 82$ for $.80$
$r = .20, p > .05$	"not enough power"	$\approx .27$ from $S=42$	$S = 191$ for $.80$
$r = .02, p > .05$	"power not 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"??

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 Control = 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 pairwise mean dif \rightarrow smallest pairwise effect size

- for TxOld (50) vs. TxNew (60)
- comp r using MSerror = std² (20² = 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, pairwise effects.

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” \rightarrow common effect sizes for that area are significant using that sample size
- so usually work well -- within their research area on well-known phenomena (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	$\alpha = .05$
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 →	.35
↓ power	
.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 H_0 : 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 H_0), even if the sample we happen to draw has a smaller r-value than the population.

By the way...

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

Confidence Intervals for Pairwise Mean Differences

Note: MS_{Error} and df_{Error} are taken from the omnibus-F

$$CI = \text{pairwise mean difference} \pm t * \sqrt{2 * MS_{error} / n}$$

For CI formula:

t = t-critical for full model df_{Error} and Type I error risk
 n = number of data points in each IV condition (or the average)

You can “adjust” for experiment wise error by changing the p-value used to look up the t in this formula.

You can compute an “HSD CI” by substituting Q for t in this formula.

Note: the +/- value of the 95% CI is the same as the LSD ($p=.05$) minimum mean difference.

Note: The same formula applies for BG and WG designs

The idea behind teaching you these different analyses is that Pairwise NHSTs aren't the only results you should compute and use to evaluate the support for your RH:

Three things you should consider are:

Effect sizes of pairwise comparisons

Post-hoc power analyses of nonsignificant results

Confidence Intervals around mean differences

Using the combination of these techniques – effect sizes, power analyses & confidence intervals -- you will also get useful information about what effects need further study -- perhaps with a modified research design & perhaps just with a larger sample size -- because the results of the present study were inconclusive.

Combing these different types of information ...

	Cx				Tx1			
	mean	M dif	95% CI	r	M dif	95% CI	r	
Cx	20.3							
Tx1	24.6	4.3	(-1.8 to 10.4)	.22				
Tx2	32.1	11.8*	(5.7 to 17.9)	.54	7.5*	(1.4 to 13.6)	.41	

* indicates mean difference is significant based on LSD criterion (min dif = 6.1)

Examining these results...

- The effect size of Cx vs. Tx1 is substantial (Cohen calls .30 "medium and .10 small"), but is not significant (note LSD and CI agree). When we check the power of the study for testing an effect of this size we find that power = .50, so "the null is a power problem"
- Related to this, the size of the Cx vs. Tx1 mean difference could be as large as 10.4 (the large interval also suggests the sample size is small)

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
