Power Analysis: Advanced Course in the UCLA Statistical Consulting Series on Power

By

Jason C. Cole, PhD

QualityMetric, Inc.
Senior Consulting Scientist
jcole@qualitymetric.com
310-539-2024

Consulting Measurement Group, Inc.
President & Senior Scientist
jcole@webcmg.com
866-782-8799

Overview

- Brief review on multiplicity control
- Multiplicity control: Round 2
- Longitudinal research: Impact of number & correlation of data points
- The 12(ish) Step Program for power analysis
- Power and psychometrics
- Introduction to Monte Carlo simulation for power

Brief Review on Multiplicity Control

Test	*p*	Standard alpha	Bonferroni	Hochberg alpha	Correlation Adj (r = .3)	Correlation Adj. (r = .6)
	.028	.05 (sig!)	.0125 (not sig!)	.025 (not sig!)	.019 (not sig!)	.029 (sig!)
	.020	.05 (sig!)	.0125 (not sig!)	.0167 (sig!)	.019 (not sig!)	.029 (sig!)
	.011	.05 (sig!)	.0125 (sig!)	.0125 (sig!)	.019 (sig!)	.029 (sig!)
	.038	.05 (sig!)	.0125 (not sig!)	.05 (sig!)	.019 (not sig!)	.029 (not sig!)

Multiplicity Control: Round 2

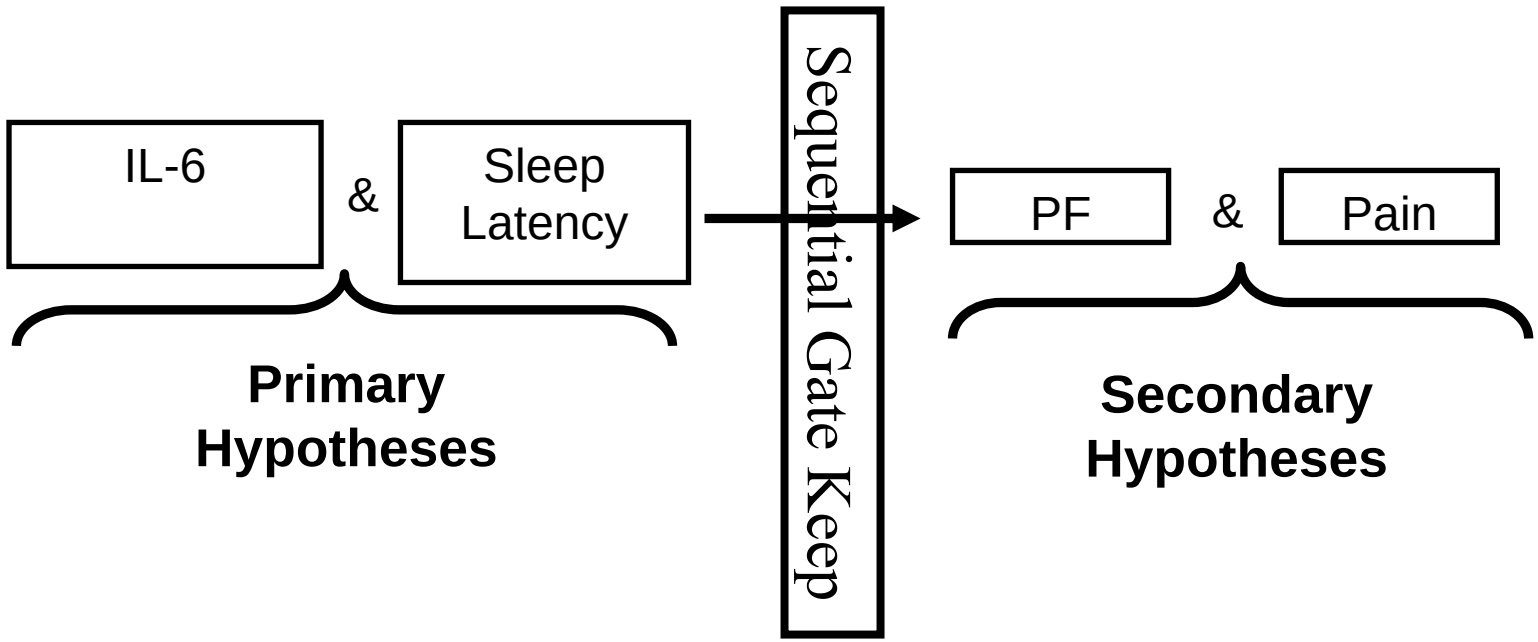
- Modern Approaches
  - Alpha adjustment based on outcome correlations
  - Sequential gate keeping
  - Selective alpha weighting
  - Use of combinatorial outcomes

Multiplicity Control: Round 2

- Sequential Gatekeeping:
  - Primary hypotheses are tested at their appropriately adjusted alpha level
  - If, **and only if**, all primaries are significant then the testing of secondary hypotheses is conducted
  - Secondary hypotheses are NOT adjusted for alpha based on the first set because the SGK protects the alpha between the two sets

Westfall, P. H., & Krishen, A. (2001). Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *Journal of Statistical Planning and Inference, 99*, 25-40.

Multiplicity Control: Round 2



Multiplicity Control: Round 2

- Selective Alpha Weighting
  - Certain hypotheses will have more difficulty obtaining significance; when known ahead of time you can account for this difference
  - Make an a priori partition of the alpha to help the weaker results by taking from stronger parts
    - i.e. alpha at .07 for the weak outcome of two results and .03 for the stronger outcome of the pair

Westfall, P. H., & Krishen, A. (2001). Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *Journal of Statistical Planning and Inference, 99*, 25-40.

Westfall, P. H., & Young, S. S. (1989). p-Value adjustments for multiple tests in multivariate binomial models. *Journal of the American Statistical Association, 84*, 780-786.

Multiplicity Control: Round 2

- Combinatorials
  - Example, several markers of affective disorders could be demonstrated to be influenced by a single second-order latent variables, and therefore appropriate to use as a combined single score.
  - Can be difficult to demonstrate appropriateness of using a total score, but may be considered if
    - many outcomes need to be acquired to capture the essence of a complex condition
    - previous research has demonstrated psychometric efficacy of a combinatorial score

Moyé, L. A. (2003). *Multiple analyses in clinical trials: Fundamentals for investigators*. New York: Springer-Verlag.

Multiplicity Control: Round 2

Known expected correlation between outcomes

	Test	*p*	Standard alpha	Bonferroni	Hochberg alpha	Corr Adj (r = .3) for each SGK	Corr Adj (r = .6) for each SGK
SGK #1	Tender joint	.028	.05 (sig!)	.025 (not sig!)	.05 (sig!)	.0311 (sig!)	.0381 (sig!)
	Swollen joints	.020	.05 (sig!)	.025 (sig!)	.025 (sig!)	.0311 (sig!)	.0381 (sig!)
SGK #2	HAQ	.011	.05 (sig!)	.025 (sig!)	.025 (sig!)	.0311 (sig!)	.0381 (sig!)
	PCS	.038	.05 (sig!)	.025 (not sig!)	.05 (sig!)	.0311 (not sig!)	.0381 (sig!)

Multiplicity Control: Round 2

Unknown expected correlation between outcomes

	Test	*p*	Standard alpha	Bonferroni	Hochberg alpha	Tukey adjustment (alpha/p)
SGK #1	Tender joint	.028	.05 (sig!)	.025 (not sig!)	.05 (sig!)	.03562 / .03937 (not sig!)
	Swollen joints	.020	.05 (sig!)	.025 (sig!)	.025 (sig!)	.03562 / .02817 (sig!)
SGK #2	HAQ	.011	.05 (sig!)	.025 (sig!)	.025 (sig!)	.03562 / .01552 (sig!)
	PCS	.038	.05 (sig!)	.025 (not sig!)	.05 (sig!)	.03562 / .05331 (not sig!)

Longitudinal Research: Impact of Number & Correlation of Data Points

- Power and the number of assessments included in a longitudinal analysis are positively correlated
- Power and the correlation between outcomes over time are positively correlated (i.e., the more an outcome is correlated to itself at different assessment times, the greater our power)

Murphy, K. R., & Myors, B. (2004). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests*. Mahwah, NJ: Lawrence Erlbaum.

Longitudinal Research: Impact of Number & Correlation of Data Points

- The impact between the analysis plan and number of assessments on power can be marked.
  - For example, if one is using a change score analysis (low in power), than only 2 assessments are needed.
  - Alternatively, if one has measured an outcome 5 times then the analysis should incorporate all 5 times (for example, RM ANOVA, mixed modeling, latent growth curve modeling)

Longitudinal Research: Impact of Number & Correlation of Data Points

- How do we get the correlation between time points?
  - Consider a study with a baseline, 2 month and 6 month assessment. Now, a prior study has found that the correlation between repeated measurements on our outcome at baseline and 2 months is .84.
  - Correlations tend to degrade autoregressively over time, so squaring the correlation for each doubling of the time is a decent approximation of the correlation between other time points.
  - Hence, if a 2-month correlation is .84, then a four-month correlation should be near .84 * .84 = .7056 (the correlation between 2-months and 6-months), and six-month correlation should be .84 * .84 *.84 = .5927.

Longitudinal Research: Impact of Number & Correlation of Data Points

- OK, I've got my list of correlations, now what?
  - This depends on your power software. You can either (A) enter each pairwise correlation into the power analysis or (B) enter a single average correlation.
  - If using the single average, make sure to (A) use a Fisher's *z* transformation and (B) consider a 90% CI on the correlation to examine the impact of a more conservative correlation
  - For the previous example, the mean correlation is .729

Longitudinal Research: Impact of Number & Correlation of Data Points

Necessary sample size per group (alpha = .05, ES is nearly large)

# Assessment	*r* not included (*r* = 0)	*r* = .729	Each *r* noted in power
2 (change-score)	323	76	132
3 (RM ANOVA)	81	55	21
5 (RM ANOVA)	41	36	SMALL!

Longitudinal Research: Impact of Number & Correlation of Data Points

- Let's see how this looks in STATA
  - sampsi 7.57 0, sd1(20.98) sd2(20.98) method(change) pre(1) post(1) power(.80) r1(.1)
  - sampsi 7.57 0, sd1(20.98) sd2(20.98) method(change) pre(1) post(4) power(.80) r1(.1)
  - sampsi 7.57 0, sd1(20.98) sd2(20.98) method(change) pre(1) post(1) power(.80) r1(.5)
  - sampsi 7.57 0, sd1(20.98) sd2(20.98) method(change) pre(1) post(4) power(.80) r1(.5)

The 12(ish) Step Program for Power Analysis

1. Organize list of all study hypotheses
2. Determine the best approach for your purposes: comprehensive or Monte Carlo (will most assume comprehensive herein, though all could be used for Monte Carlo)
3. Review the literature generally to examine which manuscripts have enough information for ESs, previously used power levels, and if current budget will allow for general sample size estimates seen in current literature

The 12(ish) Step Program for Power Analysis

4. Determine the goal power (usually .80 or .90, but it is worthwhile to consider others)
5. Extract information from previous literature (assuming no pilot study to get estimate)
   a. Use papers that have somewhat similar design: either by sample, methods, instruments, or other important factors
   b. Use meta-analysis techniques to determine which approach is best to extract ES given available information from the manuscript (usually unweighted ES estimates are best) – not all papers provide the best effect size estimates even if they are included!
   c. Organize into an Excel sheet with other pertinent information (prior example!)
   d. Convert ES into d (I like better than r because of no limit of top end of scale)
   e. Note information on missingness and nonnormality from previous literature

The 12(ish) Step Program for Power Analysis

6. Determine alpha level for each hypothesis
   a. Consider complete organization of hypotheses and previous ES estimates to consider use of sequential gate keeping, selective alpha weighting, alpha adjustments based on outcomes correlations, and combinations thereof
   b. Ascribe a specific alpha level for each hypothesis and subhypotheses based on alpha adjustments
7. Determine statistical plan for each hypothesis, considering use of multiple times and correlated outcomes enhance power
8. Determine if using a single or pooled ES and the need for using a SE of the ES estimate is needed

The 12(ish) Step Program for Power Analysis

9. Determine if any sample weightings will be applied (e.g., 2:1 ratio of experimental vs. placebo patients)
10. Estimate sample size requirements for several levels of power, for point estimate and lower bound ES, for varied levels of missingness, and perhaps for various statistical approaches using either a statistics program of Monte Carlo estimation
11. Plot Power curves
12. Find the sample size which maximizes the power for the study at hand
13. Share results, be happy, strive for world peace

Power and Psychometrics

- When validating a test, it is not sufficient to just get a significant correlation (i.e., with another convergent measure)
- Similar to other statistics (e.g., *r*), many classical psychometric statistics (e.g., coefficient α) are large-sample statistics.

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement, 11*, 93-103.

Charter, R. A. (1999). Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology, 21*, 559-566.

Charter, R. A. (2003). Study samples are too small to produce sufficiently precise reliability coefficients. *The Journal of General Psychology, 130*, 117-129.

Power and Psychometrics

- Despite the classical measurement concerns, most modern validations also involve a test of the factor structure.
- Confirmatory factor analysis is what typically drives the sample size of a validation study given that when CFA has sufficient power, most classic psychometrics have sufficient power.

Power and Psychometrics

- CFA power is determined by one of three sophistications.
  - Easiest and least accurate:
    - N:q hypothesis: one needs 10 participants per free parameter in the CFA.
    - In a unidimensional model, the number of free parameters is often equal to 2 x the number test items
    - Bentler & Chou (1987) recommended at least 10 participants per free parameter

Bentler, P. M., & Chou, C. (1987). Practical issues in structural modeling. *Sociological Methods and Research, 16*, 78-117.

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N:q hypothesis. *Structural Equation Modeling, 10*, 128-141.

Power and Psychometrics

- Medium difficulty and accuracy
  - Based on RMSEA and formulas from MacCallum et al. (1996).
  - Balances between free parameters and sample size
- Difficult and exacting
  - Monte Carlo simulation of power estimates based on formulas from Muthén and Muthén

Hancock, G. R. (2006). Power analysis in covariance structure modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 69-118). Greenwich, CT: Information Age Publishing.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance modeling. *Psychological Methods, 1*, 130-149.

Muthén, L. K., & Muthén, B. O. (2002). How to use Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9*, 599-620.

Introduction to Monte Carlo Simulation

- Monte Carlo is a technique wherein data are generated from a population of hypothesized parameters (typically based on previous research).
- A large number of synthetic samples (typically in the thousands) are used to estimate model parameters, and parameters values and their SEs are averaged over the samples.
- From these values, sample size estimates can be derived for a given power.

Introduction to Monte Carlo Simulation

- Mplus software and SAS can both calculate Monte Carlo estimates.
- Mplus has an easy engine which can accommodate nonnormality, varied missingness, growth, multilevel models, and more – see Muthén & Muthén (2002).
  - Mplus Monte Carlo uses three criteria when determining sufficient sample size is appropriate for a given power:
    - 1. Parameter & SE bias < 10%
    - 2. SE bias for parameter being tested < 5%
    - 3. Coverage remains between .91 and .98 (proportions of samples where 95% CI contains true value)

Muthén, L. K., & Muthén, B. O. (2002). How to use Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9*, 599-620.