Pragmatic and Group-Randomized Trials in Public Health and Medicine
Part 4: Power and Sample Size

David M. Murray, Ph.D.
Associate Director for Prevention
Director, Office of Disease Prevention
National Institutes of Health

A free, 7-part, self-paced, online course from NIH with instructional slide sets, readings, and guided activities
Target Audience

- Faculty, post-doctoral fellows, and graduate students interested in learning more about the design and analysis of group-randomized trials.
- Program directors, program officers, and scientific review officers at the NIH interested in learning more about the design and analysis of group-randomized trials.
- Participants should be familiar with the design and analysis of individually randomized trials (RCTs).
  - Participants should be familiar with the concepts of internal and statistical validity, their threats, and their defenses.
  - Participants should be familiar with linear regression, analysis of variance and covariance, and logistic regression.
Learning Objectives

And the end of the course, participants will be able to…

- Discuss the distinguishing features of group-randomized trials (GRTs), individually randomized group-treatment trials (IRGTs), and individually randomized trials (RCTs).
- Discuss their appropriate uses in public health and medicine.
- For GRTs and IRGTs…
  - Discuss the major threats to internal validity and their defenses.
  - Discuss the major threats to statistical validity and their defenses.
  - Discuss the strengths and weaknesses of design alternatives.
  - Discuss the strengths and weaknesses of analytic alternatives.
  - Perform sample size calculations for a simple GRT.
- Discuss the advantages and disadvantages of alternatives to GRTs for the evaluation of multi-level interventions.
Organization of the Course

- Part 1: Introduction and Overview
- Part 2: Designing the Trial
- Part 3: Analysis Approaches
- Part 4: Power and Sample Size
- Part 5: Examples
- Part 6: Review of Recent Practices
- Part 7: Alternative Designs and References
Power for Group-Randomized Trials

- The usual methods must be adapted for the nested design
  - A good source on power is Chapter 9 in Murray (1998).
  - Other texts include Donner & Klar, 2000; Hayes & Moulton, 2009; Campbell & Walters, 2014; Moerbeek & Teerenstra, 2016.
  - Recent review articles include Gao et al. (2015) and Rutterford et al. (2015).

Power for Group-Randomized Trials

- Power in GRTs is tricky, and investigators are advised to get help from biostatisticians familiar with these methods.
- Power for IRGTs is often even trickier, and the literature is more limited (cf. Pals et al. 2008; Heo et al., 2014; Moerbeek & Teerenstra, 2016).

- Heo M, Litwin AH, Blackstock O, Kim N, Arnsten JH. Sample size determinations for group-based randomized clinical trials with different levels of data hierarchy between experimental and control arms. Statistical Methods in Medical Research. 2014. PMC4329103.
Cornfield’s Two Penalties

- Extra variation
  - Condition-level statistic vs. group-level statistic
  - Greater variation in the group-level statistic
  - Reduced power, other factors constant.

- Limited df
  - df based on the number of groups
  - Number of groups in a GRT is often limited
  - Reduced power, other factors constant

Strategies to Reduce Extra Variation

- Effective strategies
  - Sampling methods
    - Random sampling within groups rather than subgroup sampling
  - Timing of measurement
    - Spring surveys rather than fall surveys for school studies (Murray et al., 1994)
    - Spreading surveys over time where there is a high within-day ICC (Murray, Catellier et al, 2006)

Strategies to Reduce Extra Variation

- Effective strategies
  - Regression adjustment for covariates
    - Fixed covariates in non-repeated measures analyses
    - Time-varying covariates in repeated measures analyses
  - This is one of the most effective methods to reduce intraclass correlation and extra variation (Murray & Blitstein, 2003) and will often reduce the ICC by 50-75%.

Strategies to Increase df

¬ Discounted strategies
  ¬ Individual level df (Murray et al., 1996)
  ¬ Kish’s effective df (Murray et al., 1996)
  ¬ Subgroup df (Murray et al., 1996)
  ¬ Mixed-model ANOVA/ANCOVA with more than 2 time intervals in the model (Murray et al., 1998)

¬ Effective strategies
  ¬ Increased replication of groups and member.


There are seven steps in any power analysis.
- Specify the form and magnitude of the intervention effect.
- Select a test statistic for that effect.
- Determine the distribution of that statistic under the null.
- Select the critical values to reflect the desired Type I and II error rates.
- Develop an expression for the variance of the intervention effect.
- Gather estimates of the parameters that define that variance.
- Calculate sample size, detectable difference or power based on those estimates.
Sample Size, Detectable Difference and Power

- Intervention effects have been defined as 1 df contrasts.
  - A t-test is an appropriate test.
  - The shape of the t-distribution is well known.
  - Critical values are easily obtained given the Type I and II error rates.
- Murray (1998) and other sources provide formulae for the variance of the intervention effect.
- The sixth step...
  - Gather estimates of the parameters that define the variance
  - Best done from data that are similar to the data to be collected (similar population, measures, design, and analysis).
Estimating ICC

- From the literature
- From a one-way ANOVA with group as the only fixed effect:

\[
\text{ICC}_{\text{m, gc}} = \frac{\text{MS}_{\text{between}} - \text{MS}_{\text{within}}}{\text{MS}_{\text{between}} + (m-1) \text{MS}_{\text{within}}}
\]
The seventh step…

- Calculate sample size, detectable difference, or power based on those estimates.
- For a one df contrast between two condition means or mean slopes, the detectable difference in a simple RCT is:

\[
\hat{\Delta} = \sqrt{\frac{\sigma^2}{\Lambda}} \left( t_{\text{critical}; \alpha/2} + t_{\text{critical}; \beta} \right)^2
\]

\[
= \sqrt{2 \left( \frac{\sigma^2}{n} \right)} \left( t_{\text{critical}; \alpha/2} + t_{\text{critical}; \beta} \right)^2
\]
The seventh step…

- Calculate sample size, detectable difference, or power based on those estimates.
- For a one df contrast between two condition means or mean slopes, the detectable difference in a simple GRT is:

\[
\hat{\Delta} = \sqrt{\hat{\sigma}^2 \left( t_{\text{critical:}\alpha/2} + t_{\text{critical:}\beta} \right)^2}
\]

\[
= \sqrt{2 \left( \frac{\hat{\sigma}_y^2 \left( 1 + (m-1) \hat{\text{ICC}}_{mg} \right)}{mg} \right) \left( t_{\text{critical:}\alpha/2} + t_{\text{critical:}\beta} \right)^2}
\]
Detectable Difference

- The most influential factors are the ICC and g. (ICC=0.100)
Detectable Difference

- The most influential factors are the ICC and $g$. (ICC=0.010)
The most influential factors are the ICC and \( g \). (ICC=0.001)
The seventh step...

- Calculate sample size, detectable difference, or power based on those estimates.
- For a one df contrast between two condition means or mean slopes, the sample size per condition for a given detectable difference $\Delta$ in a simple RCT is:

$$m = \frac{2\hat{\sigma}_y^2 \left(t_{\text{critical} : \alpha/2} + t_{\text{critical} : \beta}\right)^2}{\hat{\Delta}^2}$$

- In a simple GRT, this expression becomes:

$$g = \frac{2\hat{\sigma}_y^2 \left(1 + (m-1) I \hat{C}_{m:gx}\right) \left(t_{\text{critical} : \alpha/2} + t_{\text{critical} : \beta}\right)^2}{m \hat{\Delta}^2}$$
A Sample Size Example

- Calculate the required sample size per condition for a two-condition RCT, with 5% two-tailed Type I error rate and 80% power for a detectable difference of 0.2 standard deviations.

\[ m = \frac{2\sigma^2_y \left( t_{critical : \alpha/2} + t_{critical : \beta} \right)^2}{\hat{\Delta}^2} \]

- To perform the calculations in standard deviations, set \( \sigma^2_y = 1 \).

- Substitute this expression into the formula for the sample size to determine how many participants must be randomized to each condition.

\[ m = \frac{2(1)(1.96 + 0.84)^2}{(0.2)^2} = 392 \]
A Sample Size Example

- Calculate the required sample size per condition for a two-condition GRT, with 5% two-tailed Type I error rate and 80% power for a detectable difference of 0.2 standard deviations, given an ICC estimate of 0.01 and 100 members per group.

\[
g = \frac{2\sigma_y^2 \left(1 + (m-1)\hat{ICC}_{m:g,c}\right) \left(t_{\text{critical} : \alpha/2} + t_{\text{critical} : \beta}\right)^2}{m \hat{\Delta}^2}
\]

\[
g = \frac{2(1)(1+ (m-1)0.01)(1.96+0.84)^2}{100(0.2)^2} = 7.8
\]
A Sample Size Example

- We cannot stop at this point, because the critical values for $t$ used in this calculation are not matched to the df calculated using the result.
- $df=2(g=1)=2(8-1)=14$.
- The critical values for $t$ based on 14 df are 2.145 and 0.868.
- We repeat the calculation using those values.

$$g = \frac{2(1)(1+(m-1)0.01)(2.145+0.868)^2}{100(0.2)^2} = 9.03$$
A Sample Size Example

- df = 2(g = 1) = 2(9-1) = 16.
- The critical values for t based on 16 df are 2.12 and 0.865.

\[
g = \frac{2(1)(1+(m-1)(0.01))(2.12+0.865)^2}{100(0.2)^2} = 8.86
\]

- We can stop at this point, as the result matches the value used to calculate the critical values for t.
- There will be 80% power for a two-tailed Type I error rate of 5% to detect a 0.2 sd effect given an ICC of 0.01 and m = 100 with 9 groups per condition.
- It would be wise to perform a sensitivity analysis using several values of the ICC and m if those estimates may vary.
Unbalanced Designs

- As long as the ratio of the largest to the smallest group is no worse than about 2:1, the methods presented above are fine.
- Given more extreme imbalance, other methods are required.
  - For a GRT, several recent sources provide alternative methods.
Unbalanced Designs

- As long as the ratio of the largest to the smallest group is no worse than about 2:1, the methods presented above are fine.
- Given more extreme imbalance, other methods are required.
- For an IRGT, see
Summary

- The usual methods for detectable difference, sample size, and power must be adapted to reflect the nested design.
- Power for GRTs and IRGTs is tricky, and investigators are encouraged to collaborate with a biostatistician.
- Both of Cornfield’s penalties must be addressed: extra variation and limited df.
- Failure to do so will result in an inflated Type I error.
- There are effective design and analytic methods to reduce the extra variation.
- The most important factors affecting power in a GRT are the ICC and the number of groups per condition.
- Investigators should seek good estimates for those parameters.
Pragmatic and Group-Randomized Trials in Public Health and Medicine

Visit https://prevention.nih.gov/grt to:

- Provide feedback on this series
- Download the slides, references, and suggested activities
- View this module again
- View the next module in this series:
  Part 5: Examples

Send questions to:
GRT@mail.nih.gov