Sample Size Calculations for Group Randomized Trials with Unequal Sample Sizes through Monte Carlo Simulations

Ben Brewer

Duke University

March 10, 2017

Introduction

- Group randomized trials (GRTs) are widely used.
  - Cancer Prevention
  - Health Promotion Research involving an intervention
  - … and more!
- In GRTs, groups of subjects serve as the randomization units.
  - NOT the subjects themselves.

Introduction

- Responses within a group can be similar.
- Quantified by the Intraclass Correlation Coefficient (ICC).
  - $0 \leq ICC \leq 1$
- Measures how strongly units within a group resemble each other.
- Equal to fraction of total variation in data attributable to that unit of assignment.

Starting Out

- Lots of things to consider at the design stage.
- One of the most important considerations…

What sample size do I need?!

$N =$ 

Current Practice

- Minimal required sample size for two-arm GRT with continuous outcome:

$$N = \frac{2[z_{1-\alpha/2} + z_{1-\beta}]^2[1 + (n-1)\rho]}{\Delta^2}$$

- N = subjects per arm
- n = subjects per group
- ρ = ICC
- Δ = effect size

Groups per arm:

$$m = \frac{N}{n}$$

Issues

- Problem: the previous formula assumes equal group size!
  - Not always true.
- Formulas for other outcome types do exist…
  - These *also* assume equal group sizes.

Quick Fix?

- Replace the fixed group size n by an estimate of the average group size: $\bar{n}$.
  - Works in situations with low to moderate variance in group sizes.
- Sadly, not all real life situations feature low to moderate variance.
- When variance is great, the “equal group size” assumption fails.

Consequences

- Estimates involving sample size and statistical power are **unreliable**.
  - Could waste money.
  - Could waste time.
  - Could lead to imprudent conclusions.
  - … None of these things are good.

More Comprehensive Solutions

- Formulas do exist that adjust for variation.
  - Assume that mean are variance of group sizes are known in advance.
  - Performance can vary.
  - Most are for continuous outcomes.

Alternative Approach

- Monte-Carlo Simulation!
  - Sidesteps around study-specific assumptions.

1. Repeat data generating process.
2. Obtain empirical distribution of test statistic.
3. Use that empirical distribution to estimate power of statistical test.

- For our purposes, we do require that data generating process be fully detailed.

Roadmap

1. General framework
   I. Monte-Carlo applied to GRTs
2. Continuous Outcomes
   I. Simulated Example
3. Binary Outcomes
4. Real-Life Example!

Refocus

- Goal: Estimate sample size for balanced two-arm GRT study to evaluate an intervention.
- What do we need?

- An effect size, Δ
- A prespecified type I error, α
- A prespecified type II error, β
- $f_{n_j}(*)$ = the probability distribution of n_j
- $f_{\rho}(*)$ = the probability distribution of the ICC

Remarks

- We assume group sizes are random variables from a known probability distribution $f_{n_j}(*)$.
- We need some prior information about n_j…
  - May be approximately known by design stage.
  - May figure it out from literature review or pilot study.
  - May be able to use sensitivity analysis based on different prior assumptions of group size distributions.
    - This will give you a range: conservative is better!

Remarks

- Assuming the ICC is from a known probability distribution $f_\rho(*)$ allows for uncertainty in ρ!
- Also requires prior information.
  - May be determined by previous slide methods.
- If we have NO prior information, could use some fixed value.

STEP 1

- Since m is the number of groups per arm, $2m$ is the total number of groups in the study.
- Sample $2m$ n_j's from $f_{n_j}(*)$, and sample a single ρ from $f_\rho(*)$.

Brief Interjection

- The initial m can be chosen as the minimum feasible number of groups.
- Could also be guided by the sample calculation from way back in slide 4.
- Could be bounded above by something else…

STEP 2

- Based on the statistical model G used in the analysis, the effect size Δ, and the ICC ρ, simulate data, y.

STEP 3

- Fit the simulated data, y, using G and calculate the test statistic T.
  - Can use any fully specified parametric model.
  - We will use generalized linear mixed effects models, where groups are handled as random effects.
    - This accounts for the ICC!
- Perform the hypothesis test relevant to your study.
- Reject or fail to reject based on α.

STEP 4

- Repeat the steps 1-3 M times.
- Let K be the number of rejected hypotheses.
- Then you can estimate the power using $^{K}/_{M}$.
- If $^{K}/_{M} \approx (1-\beta)$, then m is the required number of groups per arm.
- If $^{K}/_{M} < (1-\beta)$, then increase m and re-iterate through the steps again.

Comments on Selecting M

- Want M large enough to get meaningful results.
- Don't want M so large that the results of your simulation go on your epitaph.

- If the standard deviation of the estimated power is $\sqrt{\frac{p(1-p)}{M}}$, you can get an estimate for M.
  - Just prespecify a standard deviation (like 1%) and solve!

Word of Warning

- The above method has some dodgy results when m is very small.
  - Greater variation in estimated power.
- Although this is uncommon in practice, still good to know!

Continuous Outcome (Setup)

- Prespecify:
  - $\mu_1 =$ intervention arm mean
  - $\mu_2 =$ control arm mean
  - ${\sigma_y}^2 =$ overall variance
- Then the effect size is just $\frac{\mu_1 - \mu_2}{{\sigma_y}^2}$.

Continuous Outcome (Model)

- Our linear mixed effect model is:

$$Y_{ij} = \beta_0 + \beta_1 X_j + b_j + \epsilon_{ij}$$

- $Y_{ij} =$ outcome for subject i nested within group j
- $X_j =$ binary variable indicating treatment arm
- $b_j =$ random effect of group j, $b_j \sim N(0, {\sigma_b}^2)$.
- $\epsilon_{ij} =$ random effect of subject i, group j, $\epsilon_{ij} \sim N(0, {\sigma_\epsilon}^2)$.
- We can now calculate $\rho = \frac{{\sigma_b}^2}{{\sigma_b}^2 + {\sigma_\epsilon}^2} = \frac{{\sigma_b}^2}{{\sigma_y}^2}$.

STEP 1

- Sample $2m$ n_j's from $f_{n_j}(*)$, and sample a single ρ from $f_\rho(*)$.

STEP 2

- Calculate ${\sigma_b}^2$ and ${\sigma_\epsilon}^2$.

STEP 3

- Sample b_j and ϵ_{ij} from their respective distributions.
- Simulate data Y_{ij} according the prespecified model.
  - Specifically, $\beta_1 = \mu_1 - \mu_2$.
  - Also, β_0 is the average outcome of the control arm.
- Fit the model and perform the hypothesis test for your study.

STEP 4

- Repeat steps 1-3 M times, and let K be the number of times the hypothesis was rejected.
- If ${}^{K}/_{M} \approx (1-\beta)$, then m is the required minimum number of groups per arm.
  - Total number of subjects required in the study is estimated by the average of $\sum_{i=1}^{2m} n_j$.
- If ${}^{K}/_{M} < (1-\beta)$, then increase m and re-iterate through the steps again.

Simulated Example

- Goal: calculate sample size for balanced two-arm GRT study with a continuous outcome.
- Required parameters are pre-specified:
  - $\mu_1 - \mu_2 = 5$
  - ${\sigma_y}^2 = 20$
  - So $\Delta = \frac{\mu_1 - \mu_2}{{\sigma_y}^2} = \frac{5}{20} = 0.25$
  - Desired power = 80%
  - Put $\rho = 0.02$

Simulated Example (con.)

- As a reference, assume all group sizes are equal (n = 50).
- Calculate the total sample size using:

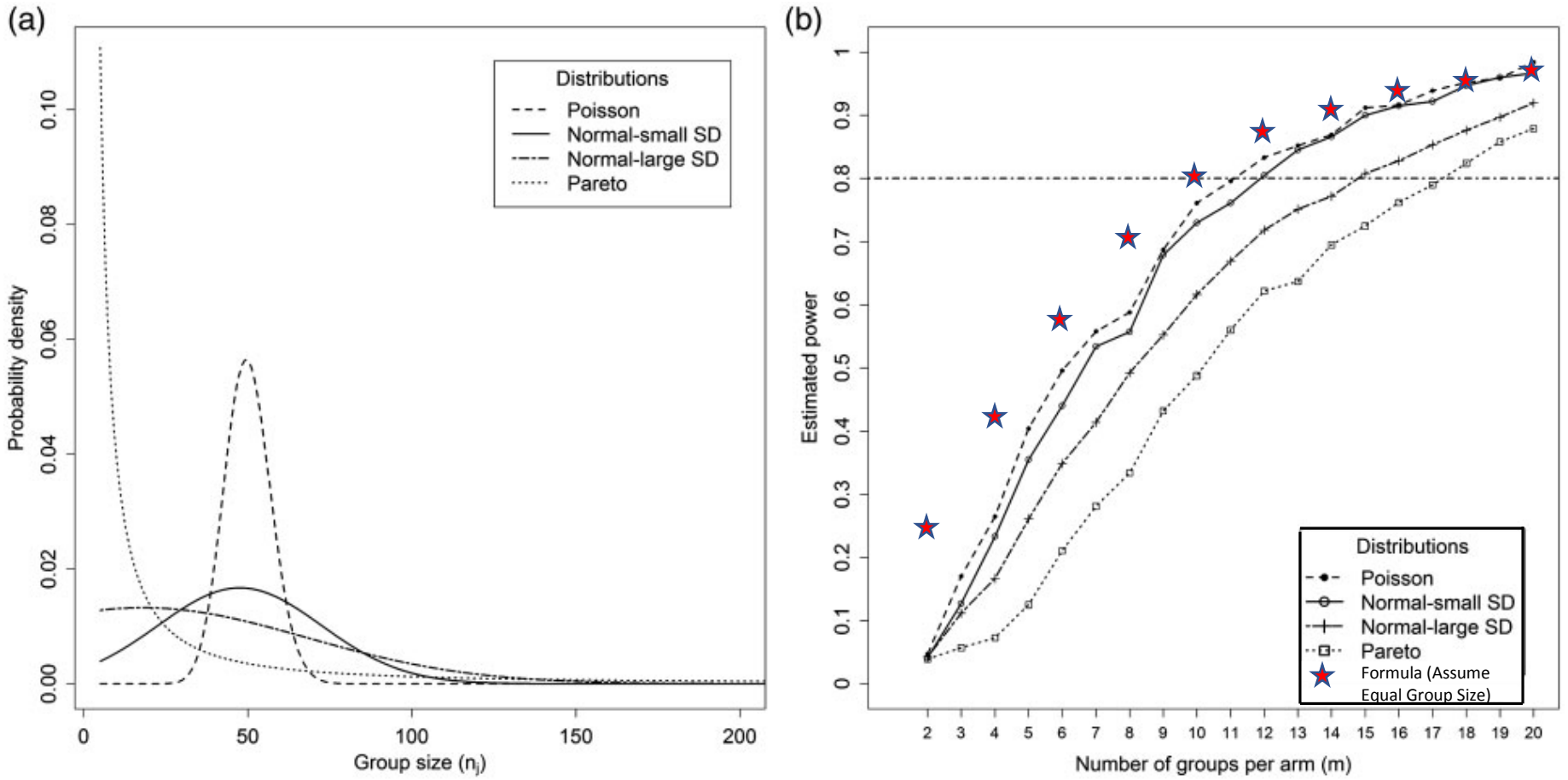
$$N = \frac{2[z_{1-\alpha/2} + z_{1-\beta}]^2[1 + (n-1)\rho]}{\Delta^2}$$

- This is the “current practice” method.

Simulated Example (con.)

- Now, allow the group sizes n_j to vary across all groups. Consider the following four distributions:
  - TruncPoisson(50, range = [5, 500]).
  - TruncNormal(47.54, 25, range = [5, 500]).
  - TruncNormal(17.99, 50, range = [5, 500]).
  - TruncPareto(5, 20, range = [5, 500]).
- NOTE: mean of TruncNormal is $\mu^* = \mu + \frac{\varphi(a) - \varphi(b)}{\Phi(a) - \Phi(b)}$

Some Graphs


(a)
Distributions
Poisson
Normal-small SD
Normal-large SD
Pareto
Probability density
0.00
0.02
0.04
0.06
0.08
0.10
0
50
100
150
200
Group size (n_j)
(b)
Estimated power
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Number of groups per arm (m)
Distributions
Poisson
Normal-small SD
Normal-large SD
Pareto
Formula (Assume Equal Group Size)

Additional Notes

- The minimum number of groups for each arm is in relatively good agreement under some circumstances.
- Notice how variance can make a big difference in required sample size!
- Alternative formula for estimating M: if the mean of $f_{n_j}(*)$ is known, then use $2m\bar{n}_f$.

Binary Outcome (Setup)

- Prespecify:
  - $p_1 =$ proportion for intervention arm
  - $p_2 =$ proportion for control arm

Binary Outcome (Model)

- The logistic mixed effect model is as follows:

$$Y_{ij} = Bern(p_{ij})$$

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 X_j + b_j$$

- $Y_{ij} =$ outcome (0 or 1)
- $p_{ij} =$ probability of 1 for subject i nested within group j
- X_j and b_j are defined as in the continuous outcome model.

Finding the ICC

- For this logistic mixed-effects model, the within group variance is $\frac{\pi^2}{3}$.
- Thus, the ICC is $\rho = \frac{{\sigma_b}^2}{{\sigma_b}^2 + \frac{\pi^2}{3}}$

STEP 1

- Sample $2m$ n_j's from $f_{n_j}(*)$, and sample a single ρ from $f_\rho(*)$.
- You should be feeling some déjà vu by this point!

STEP 2

- Calculate ${\sigma_b}^2$.

STEP 3

- Sample b_j and simulate data Y_{ij} according to the logistic model.
  - $\beta_1 = \ln(\frac{p_1/(1-p_1)}{p_2/(1-p_2)})$ = log-odds ratio between two arms.
  - $\beta_0 = \ln(\frac{p_1}{1-p_1})$ = log-odds of the proportion of the control arm.
- Fit the model and perform the hypothesis test for your study.

STEP 4

- As in the previous model, repeat your simulation M times.
- Same conclusions as before! Posted below in case you forgot:
- If $^{K}/_{M} \approx (1-\beta)$, then m is the required minimum number of groups per arm.
  - Total number of subjects required in the study is estimated by the average of $\sum_{i=1}^{2m} n_j$.
- If $^{K}/_{M} < (1-\beta)$, then increase m and re-iterate through the steps again.

Again… Why do we Care?

- Consider a study performed in 2012.
  - Goal: Examine the efficacy of the Patient Navigation Research Program.
  - Primary outcome: Receiving a definitive diagnosis within a six-month follow-up period post-intervention.
- The study employed a GRT design; there were 12 clinics involved in the study.
  - These were the units of randomization.
  - Some clinics received “treatment;” others did not.

Okay, so…

- Researchers determined that they needed 1576 total patients for this study.
  - Assumption: group sizes are equal.
- However, group sizes were NOT equal.
  - Smallest group size: 24
  - Largest group size: 379

What happened?

- The study did not find a significant difference between the two groups.
- The study also had a whopping power of $\approx 30\%$.
- These results have to be taken with a grain of salt; that sample size calculation was **unreliable**.
- What if the intervention helped expedite the health care process? What about the time and resources used to conduct the study?

Analysis

- The authors took this study to be “prior information” and performed a sample size calculation of their own.
- They considered three distributions for $f_{n_j}(*)$:
  - Truncated Poisson
  - Truncated Normal
  - Truncated Pareto
- Each of these led to a total sample size between 5,000 and 6,000.

Punch Line

- Not assuming that group sizes were equal would have led to a more realistic sample size calculation, thereby potentially saving money *and* time.

Conclusion

- Monte Carlo Simulation can provide an alternative framework for determining sample size when group sizes are not assumed to be equal.
  - More versatile; not constrained to continuous data!
- Something to ponder: what changes would need to be made when the logistic model has proportions close to 0 or 1?

Questions???