Going Beyond Simple Sample Size Calculation:

a Practitioner's Guide

Brendon McConnell & Marcos Vera-Hernández

September 24, 2015

Motivation

- Authors are aware of the **increases in complexity** that the **RCT's** (Random Controlled Trials) have been facing in the past years.

- Nonetheless this increment did not came with a **proper adjustment** of the **sample size computation** methods.

- **Why we should care about sample size in first place?:**

  1. **Effect size:** Large enough to detect expected differences between control and treatment group.

  2. **Power:** Not enough size could make our experiment unable to detect significant effects: "Usually, Power of 0.8 or 0.9 are considered high enough".

Basic Concepts: *description*

1. **Effect size** or **MDE** (Minimum Detectable Effect):
   - Which is the minimum amount of difference between my control and my treatment group that i want my study to be capable of detect?

2. **Significance level**:
   - Represent how much of "significance by luck" are you willing to admit:
     Probability of **Type I Error** (Reject null Hypothesis when it is true).

3. **Power**:
   - Represents the capacity of proper "significant assignment":
     Complement of the **Type II Error** (Not Reject null Hypothesis when it is false).

Basic Concepts: *description*

	H_0 is true	H_1 is true
Fail to reject null hypothesis	Correct	Type II error
Reject null hypothesis	Type I error	Correct

Table: Decision outcome in Hypothesis testing.

As convention:

$$[\mathbf{H_0} : \mu_T = \mu_C \quad \textit{vs.} \quad \mathbf{H_1} : \mu_T \neq \mu_C]$$

More specifically (for our case):

$$[\mathbf{H_0} : \mu_T - \mu_C = 0 \quad \textit{vs.} \quad \mathbf{H_1} : \mu_T - \mu_C = \delta\,] \rightarrow \textit{MDE}$$

Basic Concepts: *ingredients notation*

1. **Effect size**:
   - $MDE \equiv \delta$

2. **Significance level**:
   - $\textit{Type I Error} = P[\textit{reject } H_0 \mid H_0 \textit{ is True}] \equiv \alpha$

3. **Power**:
   - $\textit{Type II Error} = P[\textit{fail to reject } H_0 \mid H_0 \textit{ is False}] \equiv \beta$
   - $\textit{Power} = P[\textit{reject } H_0 \mid H_0 \textit{ is False}] \equiv 1 - \beta$

4. **Dispersion:**
   - Individual Randomization: $\textit{Variance of outcome} \equiv \sigma^2$
   - Cluster Randomization: $\textit{Intercluster Correlation (ICC)} \equiv \rho$

Overview

1. Continuous Outcome
2. Binary Outcomes
3. Introduction to Cost Minimization
4. Simulation

Continuous Outcome

A reminder:

- Under RCT (which is driven by Random Assignment):

$$\textbf{RCT} \Rightarrow \{y_{0i}, y_{1i}\} \perp T_i$$

- Thanks to this the selection bias disappears, and the Naive Comparison can tell us something meaningfully (**average causal effect**):
  - $E(Y_i|T_i = 1) - E(Y_i|T_i = 0) = \underbrace{E(y_{1i} - y_{0i}|T_i = 1)}_{ATET} + \underbrace{E(y_{0i}|T_i = 1) - E(y_{0i}|T_i = 0)}_{\text{Selection Bias} = 0}$
  - $\text{ATET} = E(y_{1i} - y_{0i} \mid T_i = 1) = E(y_{1i} - y_{0i}) = \text{ATE}$

Continuous Outcome

We have the following regression:

$$Y_i = \alpha + \beta T_i + \epsilon_i$$

where:

$\beta = ATE_{\text{(due to RCT)}} = E(y_{1i} - y_{0i}) \rightarrow$ *OLS estimator.*

$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

Continuous Outcome

Power calculation:

1. Computing Mean and Variance of the Coefficient.
2. Deriving Z-statistic.
3. Computing Power.
4. Optimal sample size given power $(1-\beta)$ and significant level (α).

Continuous Outcome

Mean and Variance of the Coefficient:

- $[\hat{\beta} = \bar{Y}_1 - \bar{Y}_0]$
- $E[\hat{\beta}] = E[\bar{Y}_1 - \bar{Y}_0] = \mu_1 - \mu_0 \underset{\text{Under } H_0}{\Rightarrow} 0$
- $\text{Var}[\hat{\beta}] = \text{Var}[\bar{Y}_1 - \bar{Y}_0] = \text{Var}[\bar{Y}_1] + \text{Var}[\bar{Y}_0] - \underbrace{2\text{Cov}(\bar{Y}_1, \bar{Y}_0)}_{=0 \text{ by independence}}$;

$$Var[\hat{\beta}] = \text{Var}[\tfrac{1}{n_1}\textstyle\sum_{i=1}^{n_1} Y_i] + \text{Var}[\tfrac{1}{n_0}\sum_{i=1}^{n_0} Y_i];$$
$$Var[\hat{\beta}] \to \left[\begin{smallmatrix}\text{as all}\\ Y_1, Y_i, \ldots, Y_n\\ \text{are independent}\end{smallmatrix}\right] = \tfrac{1}{n_1^2}\textstyle\sum_{i=1}^{n_1} \text{Var}(Y_i) + \tfrac{1}{n_0^2}\sum_{i=1}^{n_0} \text{Var}(Y_i);$$
$$Var[\hat{\beta}] = \tfrac{n_1}{n_1^2} Var(Y_i) + \tfrac{n_0}{n_0^2}\text{Var}(Y_i) = \tfrac{\sigma^2}{n_1} + \tfrac{\sigma^2}{n_0}$$

Continuous Outcome

Z-statistic:

- $Z = \frac{\bar{Y}_1 - \bar{Y}_0}{\sigma\sqrt{\frac{1}{n_0}+\frac{1}{n_1}}}$
- Given this statistic the null Hypothesis will be rejected at a significant level of α whenever:

$$[Z \geq z_{\alpha/2} \mid Z \leq -z_{\alpha/2}]$$

where $(-)z_{\alpha/2}$ is the outcome of a standard normal distribution that leaves an area of $\alpha/2$ to the (left) right.

Continuous Outcome

Power:

- $1-\beta = P(\underbrace{\text{reject } H_0}_{[Z \geq z_{\alpha/2} | Z \leq -z_{\alpha/2}]} \mid H_1 \text{ true}) = P(Z \geq z_{\alpha/2} \text{ or } Z \leq -z_{\alpha/2} \mid H_1 \text{ true})$

  $1-\beta = P(Z \geq z_{\alpha/2} \mid H_1 \text{ true}) + P(Z \leq -z_{\alpha/2} \mid H_1 \text{ true})$

- The thing is now we are not in the H_0 space (H_1 is true), so is not true anymore that our z-statistic follows a standard normal distribution. If we want to compute those probabilities we need to recalculate the mean (and subtract it to restore standardization).

Continuous Outcome

Power:

- $E(Z) = E(\frac{\bar{Y}_1 - \bar{Y}_0}{\sigma\sqrt{\frac{1}{n_0}+\frac{1}{n_1}}}) = \frac{1}{\sigma\sqrt{\frac{1}{n_0}+\frac{1}{n_1}}} \cdot \underbrace{E(\bar{Y}_1 - \bar{Y}_0)}_{\textit{Under } H_1; = \delta} = \frac{\delta}{\sigma\sqrt{\frac{1}{n_0}+\frac{1}{n_1}}}$
- $\text{Var}(Z) = \text{Var}\left(\frac{\overline{Y}_1 - \overline{Y}_0}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_0}}}\right) = \text{Var}(\overline{Y}_1 - \overline{Y}_0) \cdot \frac{1}{\left(\sqrt{\frac{\sigma^2}{n_0}+\frac{\sigma^2}{n_1}}\right)^2} = 1$
- This switch to the right of the mean also vanish $P(Z \leq -z_{\alpha/2} \mid H_1 \text{ true})$ almost to zero, and allow us to get rid of it. So we end up with:

$$1 - \beta = P\left(Z - \frac{\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_0}}} \geq z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_0}}} \,\middle|\, H_1 \text{ true}\right) \sim \mathcal{N}(0,1)$$

Continuous Outcome

Power:

- Now we want to get the cdf (cumulative distribution function), so:

$$\cancel{1-}\beta = \cancel{1-}P\left(Z - \frac{\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_0}}} < z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_0}}} \,\middle|\, H_1 \text{ true}\right) \quad \textit{By Standard Normal symmetry}$$

$$\beta = \Phi\left(z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_0}}}\right) \quad \textit{where } \Phi(\cdot) \textit{ is the cdf of Standard Normal distribution}$$

- Using the properties of the inverse function:

$$z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_0}}} = \Phi^{-1}(\beta)$$

Continuous Outcome

Power:

- After solving this $\Phi^{-1}(\beta)$ we reach the final equation:

$$P(Z > z_{1-\beta}) = 1 - \beta$$

$$\cancel{1} - P(Z \leq z_{1-\beta}) = \cancel{1} - \beta$$

$$\Phi(z_{1-\beta}) = \beta \rightarrow \Phi^{-1}(\beta) = z_{1-\beta} = -z_{\beta}$$

$$\left[-z_{\beta} = z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}} \right]$$

Continuous Outcome

- Before reaching the main equation we need to do one more transformation:

  We don't observe the population standard deviation, only the sample one, so we can not directly use the Normal distribution, instead we will be using a **t-student** distribution (with $v = n_1 + n_0 - 2)_{\textit{degrees of freedom}}$ and so:

$$\left[-t_\beta = t_{\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}} \right]$$

Continuous Outcome

- Isolating for *effect size* (δ) we get our **main equation**:

MDE for a 1- β power and a significant level of α

$$\delta = \left(t_\beta + t_{\alpha/2}\right) \sigma \sqrt{\frac{1}{n_0} + \frac{1}{n_1}}$$

Making some assumptions or modifications over this equation is how we are going to compute the demanded **optimal sample size**.

Continuous Outcome

- For example assuming that number of components of both groups (control and treatment) are the same ($\mathbf{n_1} = \mathbf{n_0}$):

$$\delta = \left(t_\beta + t_{\alpha/2}\right) \sigma \sqrt{\frac{2}{n}}$$

$$\frac{1}{\sqrt{n}} = \frac{\delta}{(t_\beta + t_{\alpha/2}) \cdot \sigma \cdot \sqrt{2}}$$

$$\left[\mathbf{n}^* = 2(t_\beta + t_{\alpha/2})^2 \frac{\sigma^2}{\delta^2} \right]$$

Continuous Outcome

- Also if the variance differs from treatment to control group ($\sigma_0^2 \neq \sigma_1^2$), we have:
  $\delta = (t_\beta + t_{\alpha/2})\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}$

$$N^* = (t_\beta + t_{\alpha/2})^2 \frac{1}{\delta^2}\left(\frac{\sigma_0^2}{\pi_0^*} + \frac{\sigma_1^2}{\pi_1^*}\right)$$

where:
$\pi_0^* = \frac{\sigma_0}{\sigma_0+\sigma_1}$ y $\pi_1^* = \frac{\sigma_1}{\sigma_0+\sigma_1}$, y $n_0^* = \pi_0^* N^*$ y $n_1^* = \pi_1^* N^*$
π_i represents some how the proportion of each group in the overall.

This case will beacome more relevant in the binary outcome case.

Cluster Randomisation

When should you cluster?

- Concerns over spillover effects.
- Concerns over unobserved characteristics

The equation will take the following form:

$$Y_{ij} = \alpha + \beta T_j + \underbrace{\nu_j}_{\text{Cluster error term}} + \underbrace{\epsilon_{ij}}_{\text{Individual error term}}$$

- Now, we define $var(\nu_j) = \sigma_c^2$ & $var(\epsilon_{ij}) = \sigma_p^2$
  The total variance will be then: $\sigma^2 = \sigma_c^2 + \sigma_p^2$

Cluster Randomization

- To compute sample size, now we will need **Intracluster correlation** (ICC), which is:

$$\rho = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_p^2}$$

Intuition

The larger the fraction of the σ^2 accounted for by the between cluster variance (σ_c^2), the more similar are outcomes within the cluster, so the less information is extracted from adding extra individuals to them.

Cluster Randomisation

- So now our cluster-level-randomisation-modified main equation is:

$$\delta^2 = (t_{\alpha/2} + t_{\beta})^2 2 \left(\frac{m\sigma_c^2 + \sigma_p^2}{mk} \right)$$

- $mk = (t_{\alpha/2} + t_{\beta})^2 \cdot 2(m\sigma_c^2 + \sigma_p^2) \cdot \frac{1}{\delta^2}$

given that $\rightarrow m\sigma_c^2 + \sigma_p^2 = (1 + (m-1) \cdot \underbrace{\frac{\sigma_c^2}{\sigma_c^2 + \sigma_p^2}}_{\rho}) \cdot \underbrace{(\sigma_c^2 + \sigma_p^2)}_{\sigma^2}$:

$$\left[\mathbf{n}^* = \mathbf{mk}^* = (t_{\alpha/2} + t_{1-\beta})^2 \cdot 2 \cdot \frac{\sigma^2}{\delta^2} \cdot \underbrace{(1 + (m-1)\rho)}_{VIF} \right]$$

where:
$m \equiv$ *individuals per cluster*
$k \equiv$ *number of cluster*

Cluster Randomisation

- Here we introduce the term **Variance Inflation Factor** (VIF) or design effect.

VIF : $(1 + (m-1)\rho) \geq 1$

This term is a consequence of the clustered treatment allocation, and will lead systematically to **larger required sample sizes**.

Take into account:

$$n_i^* = 2(t_\beta + t_{\alpha/2})^2 \frac{\sigma^2}{\delta^2}$$
$$n_c^* = 2(t_\beta + t_{\alpha/2})^2 \frac{\sigma^2}{\delta^2} \cdot (1 + (m-1)\rho)$$

But **only difference is not VIF**, now degrees of freedom for t-student have changed and so t_j. Now $v = 2 \cdot$ (k - 1) (not $2 \cdot (n-1)$ as before).

Cluster Randomisation

- In the case there are unequal number of clusters:

$$k_1 = \frac{(t_{\alpha/2} + t_{\beta})^2 \sigma^2 \left(\frac{1+(m-1)\rho}{m} \right)}{\delta^2 - (t_{\alpha/2} + t_{\beta})^2 \sigma^2 \left(\frac{1+(m-1)\rho}{mk_0} \right)}$$

$$m_1 = \frac{(t_{\alpha/2} + t_{\beta})^2 \sigma^2 \left(\frac{(1-\rho)}{k} \right)}{\delta^2 - (t_{\alpha/2} + t_{\beta})^2 \sigma^2 \left(\frac{1+(2m_0-1)\rho}{m_0 k} \right)}$$

Role of Covariates

- We are in the Randomized world:

  That means covariates are irrelevant for control out differences between groups. Nonetheless they have something to say in sample size requirement due to their **indirect effect on output variance**.

  Adding covariates $\rightarrow$ lower variance $\rightarrow$ **lower simple size requirement**

- Adding covariates leeds to:

$$n^* = mk^* = (t_{\alpha/2}+t_{1-\beta})^2 \cdot 2 \cdot \frac{\sigma_x^2}{\delta^2} \cdot (1+(m-1)\rho_x)$$

The only difference is now we have **conditional variance** (residual variance once we controll for covariates).

Role of Covariates

- This equation is fine if we have proper estimators for these conditional parameters. But this is not always the case. An **alternative** power calculation with conditional parameters is:

$$n^* = m^* k^* = \left(t_{\alpha/2} + t_\beta\right)^2 \frac{2\sigma^2}{\delta_2^2}[\ (1 + (m-1)\rho)\ -\ \underbrace{\left(R_p^2 + \left(mR_c^2 - R_p^2\right)\rho\right)}_{\text{covariates impact on the design effect}}\ ]$$

Presented by , Hedges and Rhoads (2010).

In particular this equation may be useful if R_p^2 *and* R_c^2 are reported in existing research, and the conditional parameters of the previous are not.

where:
$R_c^2 \equiv$ proportion of the **cluster level** variance component explained by the covariates.
$R_p^2 \equiv$ proportion of the **individual level** variance component explained by the covariates.

Difference-in-Differences and Lagged Outcome as a Covariate

- Data on the outcome variable **prior (baseline) and post treatment**.
- Our regression:

$$Y_{ijt} = \beta_0 + \beta_1 T_j + \beta_2 POST_t + \beta_3 (POST_t \times T_j) + v_j + v_{jt} + \epsilon_{ij} + \epsilon_{ijt}$$

where:

$v_j, v_{jt} \equiv$ **time invariant** and **time variant cluster** level errors (respectively).
$\epsilon_{ij}, \epsilon_{ijt} \equiv$ **time invariant** and **time variant individual** level errors (respectively).

$t \equiv \{0, 1\} \rightarrow \{\text{prior treatment ,post treatment}\}$

Difference-in-Differences and Lagged Outcome as a Covariate

- For the power calculation we will need the **autocorrelation over time** at individual and cluster level:

$$\rho_{p,\ \textit{(individual)}} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{pt}^2} \quad \text{and} \quad \rho_{c,\ \textit{(cluster)}} = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_{ct}^2}$$

where:
$\text{var}(v_j) = \sigma_c^2, \quad \text{var}(v_{jt}) = \sigma_{ct}^2, \quad \text{var}(\epsilon_{ij}) = \sigma_p^2, \quad \text{and} \quad \text{var}(\epsilon_{ijt}) = \sigma_{pt}^2$

- Then our IIC will be:

$$\rho = \frac{\sigma_c^2 + \sigma_{ct}^2}{\sigma_c^2 + \sigma_{ct}^2 + \sigma_p^2 + \sigma_{pt}^2}$$

Difference-in-Differences and Lagged Outcome as a Covariate

- Now we introduce the key parameter for this case:

  $r \equiv$ fraction of total variance composed by time invariant components.

$$r = \frac{\sigma_c^2 + \sigma_p^2/m}{\sigma_c^2 + \sigma_{ct}^2 + \sigma_p^2/m + \sigma_{pt}^2/m}$$

  After some modifications we end up with:

$$r = \frac{m\rho}{1 + (m-1)\rho}\rho_c + \frac{1-\rho}{1 + (m-1)\rho}\rho_p$$

Difference-in-Differences and Lagged Outcome as a Covariate

- The sample size using baseline data as covariates:

$$n^* = m^* k^* = \underbrace{(1 - r^2)}_{\textit{baseline data effect}} \quad (t_{\alpha/2} + t_{\beta})^2 \frac{2\sigma^2}{\delta_2^2} (1 + (m-1)\rho)$$

- How useful is this?

  Notice that $(1 - r^2) < 1$ *and* $(1 - r^2) < 2(1 - r)$ (where $2(1 - r)$ is the effect on sample size of mere Dif-in-Dif). $\rightarrow$ if possible, **always control** for baseline data.

  But if r is very close to 0, maybe is better strategy to devote resources to other things like increasing post-treated sample size.

Binary Outcome Case

- Now, we move to the binary case, i.e. the outcome variable is binary (the person is working or not, graduate or not...)
- We will deal with this using differences in probability of success ($\equiv \delta$).
- One big difference between the continuous case is that here, the variance is always known.
  In the binary case, the outcome follows a Bernoulli distribution, so if you know p, the variance is $p(1-p)$

Binary Outcome Case

- We are going to use a logistic model, where y_i is binary, so:

$$p_i = \text{Prob}(y_i = 1 | T_i) = \frac{e^{\beta_0 + \beta_1 T_i}}{1 + e^{\beta_0 + \beta_1 T_i}}$$

- the effect size δ can be written as

$$\delta = \underbrace{Prob(y_i = 1 | T_i = 1)}_{p_1} - \underbrace{Prob(y_i = 1 | T_i = 0)}_{p_0}$$

Binary Outcome Case

Now, following a procedure similar to the continuous case, we can arrive to:

$$N^* = \left(\frac{p_1(1-p_1)}{\pi} + \frac{p_0(1-p_0)}{1-\pi} \right) \frac{(z_\beta + z_{\alpha/2})^2}{(p_1 - p_0)^2}$$

- π: Proportion of the sample that is treated
- $n_1^* = \pi N^*$
- $n_0^* = (1-\pi) N^*$
- In this case, optimal allocation to treatment is: $\pi^* = \frac{\sqrt{\frac{p_1(1-p_1)}{p_0(1-p_0)}}}{1+\sqrt{\frac{p_1(1-p_1)}{p_0(1-p_0)}}}$

Binary Outcome Case

- In general, π^* will differ from 0.5, but in the case that $p_0 = 1 - p_1$, then there will be an even split between treatment and control status, so we can rewrite the optimal sample size as

$$n^* = (p_1(1 - p_1) + p_0(1 - p_0)) \frac{(z_\beta + z_{\alpha/2})^2}{\delta^2}$$

Binary Outcome Case: Cluster level

- We will follow now a Generalised Estimating Equation (GEE), where the clustering is accounted for in the Variance Covariance Matrix, using ρ.
- The probability of success for individual i in cluster j is

$$p_{ij} = \text{Prob}(y_{ij} = 1 | T_j) = \frac{e^{\beta_0 + \beta_1 T_j}}{1 + e^{\beta_0 + \beta_1 T_j}}$$

Binary Outcome Case: Cluster level

For cluster j, the $m \times m$ variance covariance matrix is: $V_j = A_j^{1/2} R(\rho) A_j^{1/2}$

- A_j is a diagonal matrix with diagonal elements $\underbrace{p_{ij}(1 - p_{ij})}_{\text{variance of a Bernoulli}}$
- $R(\rho)$ is a correlation matrix with diagonal elements taking the value of 1, and off-diagonal the value of ρ
- Therefore, $cov(y_{ij}, y_{km}) = \rho$ when j = m and $= 0$ when $j \neq m$

Important

Note that $R(\rho)$ has no subscript, because we are taking a GLS approach, and the same correlation is assume across clusters

Binary Outcome Case: Cluster level

The sample size equation can be written as:

$$N^* = \left(\frac{p_1(1-p_1)}{\pi} + \frac{p_0(1-p_0)}{1-\pi} \right) \frac{(z_{\alpha/2} + z_\beta)^2}{\delta^2} (1 + (m-1)\rho)$$

and, if the treatment is equally allocated, we can write it as:

$$n^* = mk^* = (p_1(1-p_1) + p_0(1-p_0)) \frac{(z_\beta + z_{\alpha/2})^2}{\delta^2} (1 + (m-1)\rho)$$

- As we can see, the design effect is the main difference between the cluster level and the individual level.

Binary Outcome Case: Cluster level

Table 4: Sample Size Requirements for Binary Outcomes Under Cluster Randomisation

		Total Sample Size Requirements (N*)				Number of Clusters (2k*)			
		Control Group Success Rate (p0): 0.1							
		numbers of individuals per cluster (m)				*numbers of individuals per cluster (m)*			
		10	30	60	100	10	30	60	100
ICC	0	392	392	392	392	39	13	7	4
	0.01	428	506	624	781	43	17	10	8
	0.03	498	734	1087	1558	50	24	18	16
	0.05	569	961	1550	2335	57	32	26	23
	0.1	746	1531	2708	4278	75	51	45	43
	0.2	1099	2669	5023	8163	110	89	84	82
		Control Group Success Rate (p0): 0.3							
		numbers of individuals per cluster (m)				*numbers of individuals per cluster (m)*			
		10	30	60	100	10	30	60	100
ICC	0	706	706	706	706	71	24	12	7
	0.01	770	911	1123	1406	77	30	19	14
	0.03	897	1321	1957	2804	90	44	33	28
	0.05	1024	1731	2790	4203	102	58	47	42
	0.1	1342	2755	4874	7700	134	92	81	77
	0.2	1978	4804	9042	14693	198	160	151	147
		Control Group Success Rate (p0): 0.5							
		numbers of individuals per cluster (m)				*numbers of individuals per cluster (m)*			
		10	30	60	100	10	30	60	100
ICC	0	769	769	769	769	77	26	13	8
	0.01	838	992	1223	1531	84	33	20	15
	0.03	977	1438	2131	3054	98	48	36	31
	0.05	1115	1885	3038	4577	112	63	51	46
	0.1	1461	3000	5307	8384	146	100	88	84
	0.2	2154	5230	9846	15999	215	174	164	160

Effect size is set to .1 and treatment is evenly allocated (π=.5).

$$N^* = \left(\frac{p_1(1-p_1)}{\pi} + \frac{p_0(1-p_0)}{1-\pi} \right) \frac{(z_{\alpha/2} + z_\beta)^2}{\delta^2} (1 + (m-1)\rho)$$

Binary Outcome Case: Cluster level

In the case that there are unequal number of clusters:

$$k_1 = \frac{\frac{p_1(1-p_1)}{m}(z_{\alpha/2} + z_\beta)^2(1 + (m-1)\rho)}{\delta - \left(\frac{p_0(1-p_0)}{mk_0}\right)(z_{\alpha/2} + z_\beta)^2(1 + (m-1)\rho)}$$

Binary Outcome Case: Covariates

- Now, we are back to individual treatment, but we allow for covariate X_i, that is discrete but not necessarily binary.

And what if it is continuous?

In the case that the covariate X_i is continuous, then we should discretise the variable

Here, we write p_i as:

$$p_i = \text{Prob}(y_i = 1 | T_i, X_i) = \frac{e^{\beta_0 + \beta_1 T_i + \beta_2 X_i}}{1 + e^{\beta_0 + \beta_1 T_i + \beta_2 X_i}}$$

- We will need extra inputs into the sample size equation, depending on the success probabilities change according to the covariate values.

Binary Outcome Case: Covariates

- First, assume X_i can take any value in $\{x_1, ..., x_Q\}$
- Now, define:
  $\theta_q = \text{Prob}(X_i = x_q)$ for $q \in \{1, \ldots, Q\}$, with $(0 < \theta_q < 1)$ and $\sum_q \theta_q = 1$
- Now, we can compute the local probabilities as:
  $p_{0q} = \text{Prob}(Y_i = 1 | T_i = 0, X_i = x_q)$ and $p_{1q} = \text{Prob}(Y_i = 1 | T_i = 1, X_i = x_q)$.
- Now, we can define the effect size for specific value of q as $\delta_q = p_{1q} - p_{0q}$ and the overall effect size as $\delta = \sum_q \theta_q \delta_q$.

Binary Outcome Case: Covariates

Now, the sample size equation would be:

$$N^* = \left(\mathbf{g}\mathbf{M}^{-1}\mathbf{g}\right) \frac{\left(z_\beta + z_{\alpha/2}\right)^2}{\delta^2}$$

where:

$$\mathbf{M} = \begin{bmatrix} m_1 & m_2 & m_3 \\ m_2 & m_2 & m_4 \\ m_3 & m_4 & m_5 \end{bmatrix}, \ \mathbf{g} = [g_{11}, g_{12}, g_{13}]$$

Binary Outcome Case: Covariates

$$m_1 = \sum_q \{\pi\theta_q p_{1q}(1-p_{1q}) + (1-\pi)\theta_q p_{0q}(1-p_{0q})\}$$

$$m_2 = \sum_q \{\pi\theta_q p_{1q}(1-p_{1q})\}$$

$$m_3 = \sum_q x_q \{\pi\theta_q p_{1q}(1-p_{1q}) + (1-\pi)\theta_q p_{0q}(1-p_{0q})\}$$

$$m_4 = \sum_q x_q \{\pi\theta_q p_{1q}(1-p_{1q})\}$$

$$m_5 = \sum_q x_q^2 \{\pi\theta_q p_{1q}(1-p_{1q}) + (1-\pi)\theta_q p_{0q}(1-p_{0q})\}$$

Binary Outcome Case: Covariates

$$\mathbf{g}[1,1] = \sum_q \theta_q \left[p_{1q}(1-p_{1q}) - p_{0q}(1-p_{0q})\right]$$

$$\mathbf{g}[1,2] = \sum_q \theta_q \left[p_{1q}(1-p_{1q})\right]$$

$$\mathbf{g}[1,3] = \sum_q x_q \theta_q \left[p_{1q}(1-p_{1q})\right]$$

Binary Outcome Case: Covariates & Cluster

- Now, we consider a cluster randomised treatment in the presence of a cluster level covariate. As before, the probability of success can be expressed as:

$$p_{ij} = \text{Prob}(y_{ij} = 1 | T_j, X_j) = \frac{e^{\beta_0 + \beta_1 T_j + \beta_2 X_j}}{1 + e^{\beta_0 + \beta_1 T_j + \beta_2 X_j}}$$

The Variance Covariance matrix is very similar to the one without covariate, but using the conditional ICC ρ_x instead of the general one, so:

$$V_j = A_j^{1/2} R(\rho_x) A_j^{1/2}$$

Binary Outcome Case: Covariates & Cluster

The sample size calculation for this section would be:

$$N^* = 2m^*k^* = (\mathbf{g}\mathbf{M}^{-1}\mathbf{g}^\top)\frac{(z_\beta + z_{\alpha/2})^2}{\delta^2}(1 + (m-1)\rho_x)$$

Table 5: Number of Clusters Required for Binary Outcomes Under Cluster Randomisation With A Binary Covariate

Control group success rates for X_j=0/X_j=1	ICC=.05			ICC=.1		
	Impacts for X_j=0/X_j=1			Impacts for X_j=0/X_j=1		
	.1/.1	.05/.15	.03/.17	.1/.1	.05/.15	.03/.17
.45/.55	50	49	49	88	86	85
.4/.6	49	47	47	85	83	81
.3/.7	42	40	39	74	70	68
.2/.8	32	29	27	56	50	47

Number of individuals per cluster, m, is set at 60. The overall base rate in this table is set to .5, with the overall impact set to .1. Treatment is evenly allocated (π=.5), and θ= P(Xj=1)=.5.

Introduction to Cost Minimization

- In cases where costs depends **solely on total number** of the sample: **Matching** number of subjects in the **Control** with the ones in the **Treatment** group $\rightarrow$ **Maximize power and minimize costs**.

- **BUT** if **cost depends also on other parameters** this is not the case anymore. (e.g Usually treatment people are more expensive due to the treat).

  Under this situations we find ourself in a push-pull situation where we have that imbalance induce **losses in power**, but this could be **compensated by an increase in the overall sample** (possible thanks to the savings).

Introduction to Cost Minimization: Simplest Case

- Individual level Randomization where control and treatment groups have different costs.

- Minimization Problem:

$$\min(\delta) = (t_\beta + t_{\alpha/2}) \cdot \sigma \cdot \sqrt{\frac{1}{n_0} + \frac{1}{n_1}}$$

$$s.t\ C = c_0 n_0 + c_1 n_1$$

Introduction to Cost Minimization: Simplest Case

- Lagrangian:

$$\mathcal{L}(n_0, n_1, \lambda) = (t_\beta + t_{\alpha/2}) \cdot \sigma \cdot \sqrt{\frac{1}{n_0} + \frac{1}{n_1}} + \lambda(C - c_0 n_0 - c_1 n_1)$$

- F.O.C:

$$\frac{\partial \mathcal{L}}{\partial n_0} = (t_\beta + t_{\alpha/2}) \cdot \sigma \cdot \left(-\frac{1}{2} \cdot \frac{1}{n_0^2 \cdot \sqrt{\frac{1}{n_0} + \frac{1}{n_1}}} \right) - \lambda c_0 = 0 \quad (1)$$

$$\frac{\partial \mathcal{L}}{\partial n_1} = (t_\beta + t_{\alpha/2}) \cdot \sigma \cdot \left(-\frac{1}{2} \cdot \frac{1}{n_1^2 \cdot \sqrt{\frac{1}{n_0} + \frac{1}{n_1}}} \right) - \lambda c_1 = 0 \quad (2)$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = C - c_0 n_0 - c_1 n_1 = 0 \quad (3)$$

Introduction to Cost Minimization

- (1) with (2):

$$\frac{\frac{(t_\beta + t_{\alpha/2}) \cdot \sigma}{2n_0^2 \cdot \sqrt{\frac{1}{n_0} + \frac{1}{n_1}}}}{\frac{(t_\beta + t_{\alpha/2}) \cdot \sigma}{2n_1^2 \cdot \sqrt{\frac{1}{n_0} + \frac{1}{n_1}}}} = \frac{c_0}{c_1} \rightarrow \frac{n_1^2}{n_0^2} = \frac{c_0}{c_1} \rightarrow \frac{n_1}{n_0} = \sqrt{\frac{c_0}{c_1}} \rightarrow n_1 = n_0 \cdot \sqrt{\frac{c_0}{c_1}} \quad (4)$$

- (4) in (3):

$$C = n_0 \left(c_0 + c_1 \sqrt{\frac{c_0}{c_1}} \right) = n_0 \left(c_0 + \sqrt{c_1 c_0} \right) \quad (5)$$

Introduction to Cost Minimization

- From (5) and knowing (4) we get the syst:

$$n_0^* = \frac{C}{c_0 + c_1\sqrt{\frac{c_1}{c_0}}} = \frac{C}{c_0 + \sqrt{c_1 c_0}}$$

$$n_1^* = \frac{C}{c_0 + \sqrt{\frac{c_1}{c_0}}} \cdot \sqrt{\frac{c_1}{c_0}} = \frac{C\sqrt{c_0}}{c_0\sqrt{c_1} + \sqrt{c_1^2 c_0}} = \frac{C\sqrt{c_0}}{c_0\sqrt{c_1} + c_1\sqrt{c_0}} = \frac{C\cancel{\sqrt{c_0}}}{\cancel{\sqrt{c_0}}\left(\frac{c_0}{\sqrt{c_0}}\sqrt{c_1} + c_1\right)}$$

$$= \frac{C}{c_1 + \sqrt{c_0 c_1}}$$

- Solving for the effect size equation: $\delta^* = (t_\beta + t_{\alpha/2})\sigma\sqrt{\frac{1}{n_0^*} + \frac{1}{n_1^*}}$

$$\delta^* = (t_\beta + t_{\alpha/2})\frac{\sigma}{\sqrt{C}}(\sqrt{c_0} + \sqrt{c_1}); \quad C^* = (t_\beta + t_{\alpha/2})^2\frac{\sigma^2}{\delta^2}(\sqrt{c_0} + \sqrt{c_1})^2$$

The End