# Power Calculations - R lab

Satoshi Koiso

03/08/2022

# Introduction

## Why are power calculations important?

Power influences many design aspects, including what research questions to pursue, how many treatment arms to employ, and even more fundamentally, whether or not to proceed with a potential research project.

For example, it may be that a remedial education program boosts tests scores by 20 percent when comparing treatment and control groups, but due to limited power, the RCT is unable to detect this true effect (with 95% confidence).

However, we can estimate whether a given design is likely to be able to detect a reasonable effect size ex-ante, allowing us to properly manage partner expectations and make the most of limited research resources.

This exercise will cover the conventional parametric method of calculating power for a RCT: 1. The conventional parametric method, and 2. A non-paremetric "simulation" method.

### Questions to consider before running power calculations

1. What is the main specification (e.g. regression) we plan to run? (It doesn't have to be fully baked, but the more "baked" it is, the more precise we can make your power estimates.)
2. What do we expect to be the mean of the outcome in the control group?
3. How about the standard deviation (SD) of the outcome in control group?
4. What sample sizes are feasible?
5. What effect sizes could the intervention reasonably cause?
6. What is the smallest, cost-effective effect size that we are interested in? (We often arrive at a reasonable answer to this question through discussions with partner organizations and literature reviews.)

# Glossary - Ingredients of Power Calculation

- **Power calculations**:
  - Determining the sample size needed to detect the minimum detectable effect (MDE) given other parameters
  - Determining the effect size that can be detected given a set sample size and other parameters.
- **Statistical power**: The sensitivity of an experiment to detect differences between the treatment and the control groups.
  - The probability of rejecting a false null hypothesis (**type II error**: failing to detect an effect when there is one). Formally, power is typically given by $1 - \beta$. To differentiate power from the treatment effect $\beta$, in this resource we will denote power by $1 - \kappa$. That is, maximizing statistical power is to minimize the likelihood of committing a type II error.

---

$$MDE = \left(t_{1-\kappa} + t_{\frac{1}{\alpha}}\right)\sqrt{\frac{1}{P\left(1-P\right)} \times \frac{\sigma^2}{N}}$$

Where:

- **Minimum detectable effect (MDE)**: The smallest effect that, if true, has $(1-\kappa)\%$ chance of producing an estimate that is statistically significant at the $\alpha\%$ level (Bloom 1995). In other words, the MDE is the effect size below which we may not be able to distinguish that the effect is different from zero, even if it is.
- $t_{1-\kappa}, t_{\frac{1}{\alpha}}$: t-value corresponding with the probability, $1-\kappa$ or $1-\kappa$ on the t-distribution with degree of freedom of $2 \times (N-1)$
- **Power** $(1-\kappa)$: Typically set at **0.8**, meaning the probability of falsely failing to reject the null hypothesis is 0.2 or 20%. Power mirrors the significance level, $\alpha$: *as $\alpha$ increases* (e.g., from 1% to 5%), the probability of rejecting the null hypothesis increases, which translates to *a more powerful test.*
- **Significance** $(\alpha)$: The probability of committing a **type I error**(false positive: falsely rejecting the null hypothesis of no effect). It is typically set at **5%**, i.e., $\alpha = 0.05$.
- **Treatment allocation** $(P)$, the proportion of the sample assigned to *the treatment group*. Power is typically maximized with an equal split between treatment arms, though there are instances when an unequal split may be preferred.
- **Variance of outcome variable** $(\sigma^2)$
- **Sample size** $(N)$
- **Intra-cluster correlation coefficient (ICC)**: A measure of the correlation between observations within the same cluster, also often given as $\rho$. *If the study involves clustered randomization* (i.e., when each unit of randomization contains multiple units of observation), you will need to account for the fact that individuals (or households, etc.) within a group such as a town/village/school are more similar to each other than those in different groups. In general, this will increase the required sample size.

---

Let's start with a simple parametric example.

# Example 1. Basic Parametric Example

The `balsakhi` dataset is provided with the package `pwrcalc`. We'll use this dataset to estimate the control mean:

```
library(pwrcalc)
data(balsakhi)

# Let us view the `balsakhi `data that we have just loaded:
View(balsakhi)
```

## 1. Basic Parametric Example

For all parametric power calculations, we'll assume a conventional 95% confidence interval and 80% power.

```
power = 0.8
alpha = 0.05
```

What do we expect the mean and standard deviation of the outcome to be in the baseline group?

Note: Since power calculations are usually done prior to a study, we often use baseline/pilot data on the study population, or government statistics for a comparable population, to get an approximate for this outcome in the control group.

The mean at baseline can a good proxy for the control mean if we assume that the outcome for the control group is not different at endline.

```
baseline_mean <- mean(balsakhi$pre_totnorm, na.rm = T) # normalized test score pre
baseline_sd <- sd(balsakhi$pre_totnorm, na.rm = T)

# Before proceeding lets check the values of baseline_mean and baseline_Sd:
baseline_mean
```

```
## [1] 0.003931545
```

```
baseline_sd
```

```
## [1] 1.011013
```

### Calculate sample size given the minimum effect size

---

$$N = \left(t_{1-\kappa} + t_{\frac{1}{\alpha}}\right)^2 \frac{1}{P\,(1-P)} \times \frac{\sigma^2}{MDE^2}$$

One common use of the power calculations is to calculate the minimum required sample size for a given effect size, baseline conditions and research design.

Let's say, based on other studies, that we expect an effect size of a third of a standard deviation. Now let's calculate the sample size given that we know the likely effect size.

```
expected_effect <- baseline_sd/3
treated_mean <- expected_effect + baseline_mean

expected_effect
```

```
## [1] 0.3370044
```

```
treated_mean
```

```
## [1] 0.3409359
```

We can take a look at the package `pwrclc` documentation now to understand the function `twomeans`. The function calculates the sample sizes for two-sample test based on means and standard deviations of the two samples.

The minimum required sample size will also depend on the split between the treatment and the control group. This is specified as the **nratio = treatment size/control size**.

This value is **1** if there are equal number of people in both groups and it increases as we allocate a higher proportion to the treatment group. For instance, if the control group size is 1/2 of the treatment group, the nratio is 2 (=1/(1/2)).

```
nratio = 1
```

Suppose we want to detect a difference of 4 between two groups (e.g., control and treatment). For example, we anticipate the control group mean being 12 and the treatment group mean being 16. In addition, suppose the standard deviation of each group is 5. We can calculate the sample size required with `pwrcalc`:

```
twomeans(m1 = 12, m2 = 16, sd = 5, nratio = 1)
```

```
##
##      Two-sample t-test power calculation
##
##              m1 = 12
##              m2 = 16
##              n1 = 25
##              n2 = 25
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE:
## m1 and m2 are the means of group 1 and 2, respectively.
## n1 and n2 are the obs. of group 1 and 2, respectively.
```

Now, we can calculate the sample sizes for the `balsakhi` dataset:

```
base_model = twomeans(
  m1 = baseline_mean, m2 = treated_mean, sd = baseline_sd,
  nratio=nratio, power=power, sig.level = alpha)

base_model
```

```
##
##      Two-sample t-test power calculation
##
##              m1 = 0.003931545
##              m2 = 0.3409359
##              n1 = 142
##              n2 = 142
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE:
## m1 and m2 are the means of group 1 and 2, respectively.
## n1 and n2 are the obs. of group 1 and 2, respectively.
```

The command outputs the the sample size given the parameters. **n1 and n2 are the control and treatment sizes respectively.** "m1" and "m2" are the control and treatment means respectively.

We need a minimum treatment size of 142 and control size of 142 to detect an effect of 0.3370044, with a probability of 0.8, if the effect is true and the ratio of the treatment and control is 1.

Check the sample number with a different method from the equation.

```
p = nratio/(1+nratio) # prop of the sample in the treatment group

t_power = qt(power, df=Inf) # qt - student t distribution quantile function
t_alpha = qt(1-alpha/2,df=Inf)

min_n <- (t_power+t_alpha)^2*1/(p*(1-p))*baseline_sd^2/(expected_effect)^2
min_n
```

```
## [1] 282.5597
```

### Minimum effect size given the sample size

Say, instead, we knew the sample size and wanted to calculate the Minimum Detectable Effect Size (MDE). You can manually calculate the effect size:

```
sample_n <- 100 # total sample size
nratio <- 1
p = nratio/(1+nratio) # prop of the sample in the treatment group

t_power = qt(power, df = 2*(sample_n-1)) # qt - student t distribution quantile function
t_alpha = qt(1-alpha/2, df = 2*(sample_n-1))

mde <- (t_power + t_alpha) * sqrt(1/(p*(1-p))*sqrt(baseline_sd^2/sample_n))
mde
```

```
## [1] 1.79043
```

Given our sample size of 100 and ratio of treatment and control group as 1, the effect needs to be higher than 1.7904304 for us to detect it with a probability of 0.8.

Some other questions to answer before calculating power:

1. How do the sample size and MDE change when the different components of the power command change?
2. Will our main specification include controls?
3. Will this study be cluster-randomized?
4. Do we expect only part of the treatment group to take-up the intervention?

## 2. Relationship between power and its components

Now, let us get a better intuition on how a larger or smaller sample size $N$ affects our power to pick up an effect.

Say *our anticipated effect size is smaller* than originally thought; how much larger would we need to make the sample in order to still pick up an effect?

Let's try an effect size that is *half* as large:

```
smaller_expected_effect <- expected_effect/2
smaller_treatet_mean <- smaller_expected_effect + baseline_mean
twomeans(m1=baseline_mean, m2=smaller_treatet_mean, sd=baseline_sd, nratio = nratio, power=power, sig.l
```

```
##
##      Two-sample t-test power calculation
##
##               m1 = 0.003931545
##               m2 = 0.1724337
##               n1 = 566
##               n2 = 566
##        sig.level = 0.05
##            power = 0.8
##      alternative = two.sided
##
## NOTE:
## m1 and m2 are the means of group 1 and 2, respectively.
## n1 and n2 are the obs. of group 1 and 2, respectively.
```

OR:

```
min_n_smallefff <- (t_power+t_alpha)^2*1/(p*(1-p))*baseline_sd^2/(smaller_expected_effect)^2
min_n_smallefff
```

```
## [1] 1141.46
```

```
min_n_smallefff/min_n
```

```
## [1] 4.039712
```

*Observation: The minimum sample required is four times as large.*

**Remember: If our MDE decreases by a factor of X, the required sample size increases by the square of X!**

You can verify this from the other side i.e. look at the impact on the MDE of increasing your sample size by a factor of X. Say X is 4:

```
large_sample <- 4*(sample_n)

new_mde <- (t_power + t_alpha) * sqrt(1/(p*(1-p))*sqrt(baseline_sd^2/large_sample))
new_mde/mde
```

```
## [1] 0.7071068
```

## 3. Parametric Power calculations with controls

Now, say we plan to control for baseline covariates in our main specification. The inclusion of these controls will improve our power, since they explain some of the variance in our outcome.

For example, including data on baseline characteristics like sex or age may explain some of the variance in the outcome. Note that we may not need/want to include covariates if the treatment and the control are randomly allocated. To see how potential controls affect power, we would ideally have access to a sample data set (e.g. historical or pilot data).

With these data, we would want to **regress the outcome on the covariates to evaluate how much variance is explained by the set of covariates we plan to include.**

From this regression, we are interested in the residual standard deviation of the outcome variables, or the variance of the outcome that is NOT explained by controls.

**This residual SD becomes the new SD** we include in our parametric power calculations.

Using `balsakhi` data, this would be:

```
# We are using the math and verbal scores of the students at baseline as covariates

fit <- lm(pre_totnorm~pre_math+pre_verb, data = balsakhi, subset=bal==0)
summary(fit)
```

```
##
## Call:
## lm(formula = pre_totnorm ~ pre_math + pre_verb, data = balsakhi,
##     subset = bal == 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5964 -0.3316 -0.1227  0.2977  0.9087
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.3455498  0.0091805 -146.57   <2e-16 ***
## pre_math     0.0364971  0.0007000   52.14   <2e-16 ***
## pre_verb     0.0485796  0.0008088   60.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3736 on 5205 degrees of freedom
## Multiple R-squared:  0.8604, Adjusted R-squared:  0.8604
## F-statistic: 1.605e+04 on 2 and 5205 DF,  p-value: < 2.2e-16
```

```
res_baseline_sd <- sd(summary(fit)$residuals, na.rm = T)
res_baseline_sd
```

```
## [1] 0.3735381
```

If we knew the effect size and wanted to know the sample size needed:

```
cov_model <- twomeans(m1 = baseline_mean, m2 = treated_mean, sd = res_baseline_sd, nratio=nratio,
                      power=power, sig.level = alpha)
```

We need a minimum treatment size of 20 and control size of 20 to detect an effect of 0.3370044 with a probability of 0.8 if the effect is true and the ratio of the treatment and control is 1.

## 4. Parametric power calculation with partial take-up

In randomized designs, it is common that there is partial take-up of the intervention. For example, in the Oregon Health Insurance Experiment, the offer to apply for health insurance was associated with only a 25 percentage point increase in take-up of health insurance. When take-up is not 100 percent, researchers are often interested in the answers to second stage questions, such as what the average effect of becoming insured is on health care utilization.

Below, we provide code that adjusts command to take into account that only some individuals in the treatment group take up the intervention. The measure of take-up that we care about is "effective take-up", or **the percentage of individuals in the treatment group that takes up the intervention MINUS the percentage of individuals in the control group that takes up**.

Let us say 90% take up in the treatment group, and 10% do so in the control group. We then have an effective take-up rate of 80%:

```
takeup_treat <- 0.9
takeup_control <- 0.1

tu <- takeup_treat - takeup_control #effective take-up
effect_tu <- expected_effect * tu #effect size after adjusting for take-up
treat_tu <- baseline_mean + effect_tu #treatment mean after adjusting for take-up

partial_model <- twomeans(m1 = baseline_mean, m2 = treat_tu, nratio=nratio, sd=baseline_sd, power=power
partial_model
```

```
##
##      Two-sample t-test power calculation
##
##              m1 = 0.003931545
##              m2 = 0.2735351
##              n1 = 221
##              n2 = 221
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE:
## m1 and m2 are the means of group 1 and 2, respectively.
## n1 and n2 are the obs. of group 1 and 2, respectively.
```

Here we assume that the standard deviation does not change with the treatment but you can also specify different standard deviations for the control and treatment groups. We need a higher sample size to have the same power **because the expected effect has decreased by the take-up rate**.

## 5. Parametric power calculation for cluster RCTs

Many designs randomize at the group level instead of at the individual level. For such designs, we need to adjust our power calculations so that they incorporate the fact that individuals within the same group may be subject to similar shocks, and thereby have correlated outcomes. Duflo et al. "Using Randomization in Development Economics Research: A Toolkit." presents a modified parametric approach, which takes into account **the intra-cluster correlation (ICC)** that arises from randomization at the group level.

We can think of cluster RCTs as follows: - When ICC = 0, then our N is effectively the number of individuals in the study. - When ICC = 1, then our N is effectively just the number of clusters. - Usually the ICC lies somewhere between 0 and 1, requiring that we adjust our power calculations to account for this.

Below we adjust power estimates based on Duflo et al.'s model.

---

$$
\begin{aligned}
MDE &= \left(t_{1-\kappa} + t_{\frac{1}{\alpha}}\right)\sqrt{\frac{1}{P\left(1-P\right)} \times \frac{\sigma^2}{N} \times \left(1+\rho\left(m-1\right)\right)} \\
&= \left(t_{1-\kappa} + t_{\frac{1}{\alpha}}\right)\sqrt{\frac{1}{P\left(1-P\right)J}} \times \sqrt{\frac{\sigma^2}{m} \times \left(1-\rho\right)+\rho}
\end{aligned}
$$

Where:

- $J$: the number of clusters
- $m$: the number of individuals in each cluster (ave. cluster size)

---

Note: This model assumes that all clusters in a treatment arm are of the same size and have the same number of individuals. It's usually okay if this is violated in reality, but you would not want to use these adjustments if groups are dramatically different in size (e.g. group one has 10 individuals, group two has 1,000 individuals.) More on this model is explained in Duflo et al.'s article "Using Randomization in Development Economics Research: A Toolkit."

First, let's calculate the intra-cluster correlation (ICC) which measures how correlated the error terms of individuals in the same cluster are:

Here the ICC is calculated from the baseline dataset but it can also be manually defined based on historical data, other studies etc

```r
baseline_subset <- subset(balsakhi, !is.na(pre_totnorm))
cluster_var_subset <- as.factor(baseline_subset$divid) #divid - cluster variable
outcome_subset <- baseline_subset$pre_totnorm

icc <- ICCest(cluster_var_subset, outcome_subset, data=baseline_subset)
rho <- icc$ICC
rho
```

```
## [1] 0.1355969
```

### Part 1: Calculating MDE

```
# Number of individuals in  each cluster
cluster_size <- 53
# number of clusters (as documented in the Balsakhi experiment):
total_clusters <- length(unique(baseline_subset$divid))

# assuming 95% confidence intervals and 80% of power
t_stat <- t_alpha + t_power
# assuming 1:1 treatment and control group of study population
p <- 0.5

mde <- t_stat * sqrt(1/(p*(1-p)*total_clusters))*sqrt((baseline_sd^2)*(1-rho)/cluster_size+rho)

# total sample size and treated individuals
n <- total_clusters * cluster_size
treated <- n * p
```

### Part 2: The number of clusters given cluster size and effect size

For R, please refer to the `pwrclc` [documentation: http://pwrcalc.readthedocs.io/en/latest/?badge=latest) for the function `clustered` that adjust for the number of individuals per cluster:

```
cluster_number <- twomeans(m1=baseline_mean, m2=treated_mean, sd=baseline_sd, nratio=nratio, sig.level=a
  clustered(obsclus = cluster_size, rho = rho)

minimum <- cluster_number$`Minimum number of clusters`
minimum
```

```
## [1] 44
```

Given the size of each cluster as 53, and ratio of the number of units in the treatment to control as 1, we need a minimum of 44 clusters to detect an effect of 0.3370044, with a probability of 0.8 if the effect is true.

Adjusted n1 and n2 indicate the sample size in the control and the treatment group respectively. Sample size is the total number of units across the clusters.

## Part 3. Cluster size given the number of clusters and effect size

```
cluster_size_model <- twomeans(m1=baseline_mean, m2=treated_mean, sd=baseline_sd, power=power, nratio=n
  clustered(numclus=total_clusters, rho=rho)

ave_cluster <- cluster_size_model$`Average per cluster`
```

Given, 193 clusters, and the ratio of units in the treatment and the control as 1, the minimum size of each cluster should be 2 for us to detect an effect of 0.3370044 with a probability of 0.8 if the effect is true.

Note that the above calculations assume a distribution of the average effect size. Non-parametric power simulations do better than parametric power calculations when we have access to good data (historical, baseline, or pilot) on our study population. From these data, we can simulate a fake dataset that assumes the treatment has no effect and then see what effects we are powered to detect, by looking at the simulated 95% confidence interval around our null effect. We should expect that any effect greater than this confidence interval would be detected by our study.

In particular, power simulations do not require the assumption that the sampling distribution of your Beta coefficient(s) of interest takes a normal distribution in your (finite) sample. You may be particularly worried about this assumption (of parametric power calculations) if your sample is very small.

To do non-parametric power simulations we need to create a (reasonable) "fake" or simulated dataset. For example, if you have baseline data for the 3 months prior to a 12 month trial, then a reasonable way to expand this dataset would be to simply randomly draw days with replacement until you have 365 days in your dataset. Similarly, you could use historical data from the two years prior to the study to estimate the confidence interval around a null effect in the past year, with data on your outcome variable from two years ago serving as controls.

---

# Questions:

### 3.1. What percent of the variance of study period test scores is explained by the covariates, pre_verb and pre_math? (Hint: Look at the R^2 statistic of the regression.)

```
fit <- lm(pre_totnorm~pre_math+pre_verb, data = balsakhi, subset=bal==0)
summary(fit)
```

```
##
## Call:
## lm(formula = pre_totnorm ~ pre_math + pre_verb, data = balsakhi,
##     subset = bal == 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5964 -0.3316 -0.1227  0.2977  0.9087
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.3455498  0.0091805 -146.57   <2e-16 ***
## pre_math     0.0364971  0.0007000   52.14   <2e-16 ***
## pre_verb     0.0485796  0.0008088   60.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3736 on 5205 degrees of freedom
## Multiple R-squared:  0.8604, Adjusted R-squared:  0.8604
## F-statistic: 1.605e+04 on 2 and 5205 DF,  p-value: < 2.2e-16
```

**A. About 86 percent of the variation in pre_totnorm is explained by pre_verb and pre_math**

### 3.2. How does this affect our sample size (compared to not including controls)?

```
1-cov_model$n1/base_model$n1
```

```
## [1] 0.8591549
```

**A. Including controls reduces sample size by about 85 percent.**

### 3.3. How about our MDE?

**A. Reduces MDE to a lesser extent (~63%).**

### 6.1. Why do we have to adjust power for clustering when running a cluster RCT?

**A. Assignment is only random at the cluster level; thus we must cluster our standard errors in our main specification. To this end, our power calculations must also take this into account, since, by clustering in our main specification, we will lose all precision gained from intra-cluster correlation in outcomes.**

### 6.2. Assuming ICC>0, does adding a new cluster of 5 individuals or adding 5 individuals to already-existing clusters give us more power to detect effects?

**A. Adding 5 individuals to existing clusters will increase power by a lesser amount. In the limiting case of ICC=1, adding 5 individuals to previous clusters would have no effect on power, while adding 5 individuals in a new cluster would increase our effective N by 1.**