# Module 6: Power Analysis and Sample Design

## Contents

## List of Figures

# 1. INTRODUCTION

In the first section of this course, we introduced you to various statistical and econometrics techniques used when analyzing the impact of a program, and introduced you to R. So far in Section 2, we discussed the "missing data" problem of causal inference that we hope to overcome with different experimental and quasi-experimental methodologies. This module will be slightly different, as we move towards the set-up *a priori* an experiment or evaluation of a prospective project, and how to make sure we have a sufficient sample size to detect an effect if one is actually there.

In the previous modules we have used statistical tests to determine whether the detected difference in mean of an outcome between two groups (treatment and control) is statistically significant. We do so by specifying a null hypothesis that the means are not different from each other, or that the mean difference is zero. The t-statistic and the p-value help us determine whether we can reject the null hypothesis on basis of the role of "chance" or sampling error in the observed difference between the two means. We have covered these concepts in Modules 1.2 and 1.3 already. In the output of any regression analysis in R, the coefficient and its standard error is reported along with the p-value for a two-sided test to evaluate the null hypothesis that the coefficient is 0. The norm is to consider coefficients with p-values less than 0.1, 0.05 or 0.01 as statistically significant at that level. In short, we assessed the probability of detecting a statistically significant impact (difference in the means), or rejecting the null hypothesis when in reality the null hypothesis was the truth. Let's call this latter probability a measurement of "false positive" impact.

What about situations in which we fail to reject the null when in reality an impact occurred? Indeed, this "false negative" estimate is also of concern in statistical testing. The level of acceptable false positive and false negative by which evaluators are willing to abide determines the sample size needed in the study, along with some other population- and intervention-related factors.

It is important to distinguish between a cross-sectional representative sample to measure a population parameter and a sample to detect the difference between the means of two groups of populations. In the former, we mainly worry about estimating the sample size necessary to measure the "expected or assumed" population parameter (e.g. mean income, prevalence of a diseases, etc.) to a predefined level of precision. In the latter, our main concern is to determine a sample size large enough to restrict the prevalence of false positives and false negatives to acceptable minimum levels in measuring "expected" or "assumed" mean difference between two populations.

The objective of this module is to introduce power calculations and sample size determination from an impact evaluation perspective. However, sample design for measuring a population parameter for a given precision is a key precursor to understanding these concepts. Therefore, we urge students to cover the basics of sampling by taking a short online course prepared by Sistemas Integrales (http://lsms.adeptanalytics.org/course/fscommand/session3/Ses3_eng.html)

This module is not designed to provide in-depth knowledge of theoretical issues involved in sample design and power analysis, but to give tools and skills in R that students can use in straightforward situations. In reality, sample design is both a science and art; it has to balance statistical theory with field implementation, logistics, and economics considerations.

The learning objectives of this module are:

- ✓ Understanding the terms minimum detectable effect (MDE), power, and sample size
- ✓ Using R functions to estimate sample size, determine MDE, and assess the power of given sample size for randomized control trials
- ✓ Understanding the effect of sample design assumptions on power, sample size and MDE

# 2. MOTIVATION

To demonstrate how sample size and power analysis are useful, let's conduct a simulation in R. We are going to create an artificial program that randomly assigns college education (a dummy treatment variable equal to 1 for those who are treated) in a population of 3 million individuals. We then evaluate the impact of college education on weekly earnings (measured in logs). Since we are generating this artificial data yourself, we will fix in advance that the true impact of this intervention will be 0.1; in other words, those with college education will have a 10% higher weekly earnings with respect to those without a college education. We also simulate the natural log of the mean of income for the control group as 5 (or 148 dollars per week).

After create a dataset with 3 million observations, we generate an unobservable composite variable `u`, which we assume is distributed as a standard normal curve. Then we create the random treatment variable `edu_random`. Finally, we simulate the natural log of weekly earnings (`lny_random`). The effect of `edu_random` on `lny_random` in the population can be estimated by regression analysis. The R code is as follows.

```
# 3 million participants
set.seed(521)
u_talent <- rnorm(3000000)
aux_random <- rnorm(3000000)

edu_random <- ifelse(aux_random > 0, 1, 0)
rm(aux_random)
lny_random <- 5 + (0.1*edu_random) + u_talent

model <- lm(lny_random ~ edu_random)
summary(model)
```

The regression output is presented in Figure 1. The results show exactly what we would expect by our construction of the simulated data. The coefficient for the treatment is 0.1 and the constant is ~5.

```
> model <- lm(lny_random ~ edu_random)
> summary(model)

Call:
lm(formula = lny_random ~ edu_random)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7914 -0.6749  0.0004  0.6745  4.9888

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.9995104  0.0008163 6124.46   <2e-16 ***
edu_random  0.1006173  0.0011542   87.17   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9996 on 2999998 degrees of freedom
Multiple R-squared:  0.002527,  Adjusted R-squared:  0.002526
F-statistic:  7599 on 1 and 2999998 DF,  p-value: < 2.2e-16
```

**Figure 1. Regression using entire population to demonstrate true population level impact**

We now take a random sample of 3,000 observations from the original population and re-estimate the impact of the intervention from this sample. We use the `sample` function to take samples from the population, and then proceed with the regression. The code is as follows:

```
# 3000 sample
edu_random_3000 <- sample(edu_random, 3000)
lny_random_3000 <- sample(lny_random, 3000)

model_3000 <- lm(lny_random_3000 ~ edu_random_3000)
summary(model_3000)
```

The regression output is given in Figure 2. We again find that the impact, or the average treatment effect (ATE), is approximately 0.1 and the constant term is approx. 5 as expected. Note that, because the sample is random, the sample estimated parameters are unbiased expectations of the population level parameters. However, we find that the standard errors are larger relative to those from our analysis using the full population. We discussed in Module 1.3 how standard errors shrink as sample size increases.

```
> model_3000 <- lm(lny_random_3000 ~ edu_random_3000)
> summary(model_3000)

Call:
lm(formula = lny_random_3000 ~ edu_random_3000)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6054 -0.6566 -0.0127  0.6804  3.6650

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.00427    0.02489 201.053  < 2e-16 ***
edu_random_3000  0.10168    0.03551   2.863  0.00422 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9723 on 2998 degrees of freedom
Multiple R-squared:  0.002728,  Adjusted R-squared:  0.002395
F-statistic: 8.199 on 1 and 2998 DF,  p-value: 0.004219
```

**Figure 2. Regression using a sample (n = 3000) of population to estimate sample level impacts**

We can repeat above analysis but with a sample size of 50 as follows. The results of regression analysis are given in Figure 3. The ATE is now 0.36, which is higher than the true ATE of 0.1. However, the standard errors are large so that the ATE is statistically not significantly different from 0 (or 0.1, the true ATE) at conventional levels. Therefore, this analysis would not have provided evidence that the true ATE is greater the zero (rejecting the null hypothesis that it's equal to zero), despite the fact that it is! Therefore, we would commit a 'false negative' error in our inference.

```
# Sample size : 50
edu_random_50 <- sample(edu_random, 50)
lny_random_50 <- sample(lny_random, 50)
model_50 <- lm(lny_random_50 ~ edu_random_50)
summary(model_50)
```

```
> model_50 <- lm(lny_random_50 ~ edu_random_50)
> summary(model_50)

Call:
lm(formula = lny_random_50 ~ edu_random_50)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8636 -0.8363  0.0436  0.6933  2.4486

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.8266     0.2291   21.06   <2e-16 ***
edu_random_50  0.1384     0.3009    0.46    0.648
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.05 on 48 degrees of freedom
Multiple R-squared:  0.004389,  Adjusted R-squared:  -0.01635
F-statistic: 0.2116 on 1 and 48 DF,  p-value: 0.6476
```

**Figure 3. Regression using a sample (n = 50) of population to estimate sample level impacts**

# 3. DEFINITION OF KEY TERMS

We now understand the terms hypothesis testing and statistical significance. We have discussed that 'false positive' errors can happen when we infer that the difference between two means is statistically significantly different when in reality both groups have the same mean. The rate or probability of observing false-positive errors is called the rate of **Type I error**. The most common convention is to keep the rate of Type I error to 0.05 or less. However, the values of 0.1 and 0.01 are also sometimes used in literature.

'False negative' errors happen when we statistically fail to detect a difference between two means when in fact the difference exists. The rate or probability of observing 'false negative' error is called the rate of **Type II error**. The typical acceptable value of Type II Error used in literature is 0.1, or as high as 0.2.

The **Power of the test** is a concept closely related to the Type II error rate. Power is defined as 1 minus the Type II error rate. For example, for a Type II error rate of 0.2 we have (1 – 0.2 =) 80% power. Power is therefore the probability of detecting an effect when this effect truly exists.

**Minimum Detectable Effect (MDE)** is the smallest difference in means we can measure between two groups for the given sample size, power, and Type I error rate. If the estimated difference is less than the MDE, then we may not be able to find statistical evidence of impacts when they truly exist.

The **sample size** is the number of observations or participants in the study. Sometimes sample size is reported for each group being compared, and sometimes the total sample size is reported along with the proportion in each of the groups.

Although R and other statistical software can do all calculations related to power analysis and sample design, we list a few formulae for your reference.

Consider the following simple regression to evaluate the impact of a ***randomized*** intervention (D):

$$Y_i = \alpha + \beta D_i + \varepsilon_i;$$

where $D_i$ is the treatment variable and $\varepsilon_i$ is an error term, both defined for individual $i$. $Y_i$ is the outcome of interest and $\beta$ is the impact of the intervention.

If we assume that the observations are independent of each other and are identically distributed, then the variance of the treatment effect can be written:

$$Var(\hat{\beta}) = \frac{1}{P(1-P)}\frac{\sigma^2}{N}$$

where $\sigma^2$ is the variance of the outcome of interest, $N$ is the sample size and $P$ is the fraction of the sample assigned to the treatment group.

The MDE is estimated as

$$MDE(k,\alpha,N,P) = (t_{(1-k)} + t_{\alpha}) * \sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}$$

where $t_{(1-k)}$ is the t-statistic associated to the inverse of the power (Type II error rate) and $t_{\alpha}$ is the t-statistic associated with the significance level (Type I error rate). For a level of power of 80% and a significance level of 5%, these values are 0.84 and 1.96 respectively for large sample; for smaller samples, we have to use a student's t distribution with N-1 degrees of freedom instead of the normal distribution. To interpret the MDE, it is important to compare it to a given "standard" or reference value. This standard or reference value is usually derived from the theory of change and the indicator for success. For example, for a new cash transfer program the funders or implementers might decide that the anticipated treatment effect will be 100 US dollars in annual income. If the MDE for a given sample is $150, then we will not be able to statistically detect differences less than $150, and thus risk finding no statistical evidence of a positive treatment effect despite the effect's existence!

Power can be computed as,

$$t_{(1-k)}(N,\alpha,\beta_E,P) = \frac{\beta_E}{\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}} - \tau_{\alpha};$$

where $\beta_E$ is the effect size, or the expected change in the outcome as a consequence of the intervention. Notice that the MDE is also an effect size, but it is the minimum effect size for a given level of power and sample size. We distinguish between these two just to emphasize that the MDE is an estimate, whereas the effect size is a parameter (an assumption we make at the baseline).

The sample size can be estimated as,

$$N(k,\alpha,\beta_E,P) = \left[\frac{\sigma * (t_{(1-k)} + t_{\alpha}) * \sqrt{\frac{1}{P(1-P)}}}{\beta_E}\right]^2;$$

where all the terms have been previously defined.

# 4. POWER ANALYSIS AND SAMPLE DESIGN IN R

Sometimes we may work with secondary data and have to determine the power for given "fixed" sample size. Other times, we mainly need to recommend a sample size before conducting a survey. We will start an R session but we don't need to load any data. We will demonstrate some features of the `power.t.test` command in R in this section. We will only demonstrate basic features and options; students are encouraged to read the help files for more advanced options.

**MDE for comparing two means**

The command below is used to estimate the MDE (called 'delta' in `power.t.test` command output) for given sample size (N), power and alpha or Type I error rate. We use the `power.t.test` command because we are using a t-test for a comparison of two means, there are also other commands for different tests. Figure 4 presents the command and the output. Note that we have specified the power as 90% and Type I error rate or alpha as 0.05 (or 5%). The output suggests that we can detect a difference of approx. 203 units in the mean between the treatment and control groups.

```
> power.t.test(n = 1000, delta = NULL, sd = 1400, sig.level = 0.05,
+              power = 0.9,
+              type = "two.sample",
+              alternative = "two.sided",
+              strict = FALSE, tol = .Machine$double.eps^0.25)

     Two-sample t test power calculation

              n = 1000
          delta = 203.0486
             sd = 1400
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group
```

**Figure 4. Power command to estimate MDE**

**Power for detecting given treatment effect**

Figure 5 presents an additional R command and the output. Here, we want to detect a 150 unit difference in means, but the sample size is adequate to have power of only 67%. This is expected, because the MDE for a sample size of 1000 was 203 (see Figure 4) and in detecting a smaller difference we are bound to lose some power.

```
> power.t.test(n = 1000, delta = 150, sd = 1400, sig.level = 0.05,
+              power = NULL,
+              type = "two.sample",
+              alternative = "two.sided",
+              strict = FALSE, tol = .Machine$double.eps^0.25)

     Two-sample t test power calculation

              n = 1000
          delta = 150
             sd = 1400
      sig.level = 0.05
          power = 0.6680995
    alternative = two.sided

NOTE: n is number in *each* group
```

**Figure 5. Power command to estimate power of a given sample**

**Sample size estimation**

Figure 6 presents a power command used to estimate the necessary sample size of an intervention. We find that, in order to detect a treatment effect of 150 units, with our other assumption about parameters, we need more than 1800 participants in our survey.

```
> power.t.test(n = NULL, delta = 150, sd = 1400, sig.level = 0.05,
+              power = 0.9,
+              type = "two.sample",
+              alternative = "two.sided",
+              strict = FALSE, tol = .Machine$double.eps^0.25)

     Two-sample t test power calculation

              n = 1831.588
          delta = 150
             sd = 1400
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group
```

**Figure 6. Power command to estimate the sample size**

*Exercise 6.1*

*(a) Estimate the minimum detectable effect for a two sided comparison of means with sample size of 440, standard deviation of 10.2, significance level of 0.05, and power of 80%. What happens to the MDE if you increase the sample size?*

*(b) Estimate the statistical power for a two sided comparison of means test with significance level of 0.10, sample size of 20500, and standard deviation of 1010, and MDE of 250. What happens to the statistical power if you increase the standard deviation?*

*(c) Estimate the required sample size for significance level of 0.05, power of 80% standard deviation of 0.65 and MDE of .03. What happens to the required sample size when you increase the power?*

**Effect of Power and Type I error assumptions**

It can be useful to graph the sample size for different delta and power values to understand cutoffs for our experiments. Figure 7 is produced by the following command:

```
nvals <- seq(500, 10000, 500)
powervals <- sapply(nvals, function (x) power.t.test(n=x, delta=150, sd=1400, sig.level = 0.05)$power)
plot(nvals, powervals, xlab="n", ylab="power",
     main="Power curve for\n t-test with  delta=150, sd = 1400, sig-level = 0.05",
     lwd=2, col="red", type="l")
```

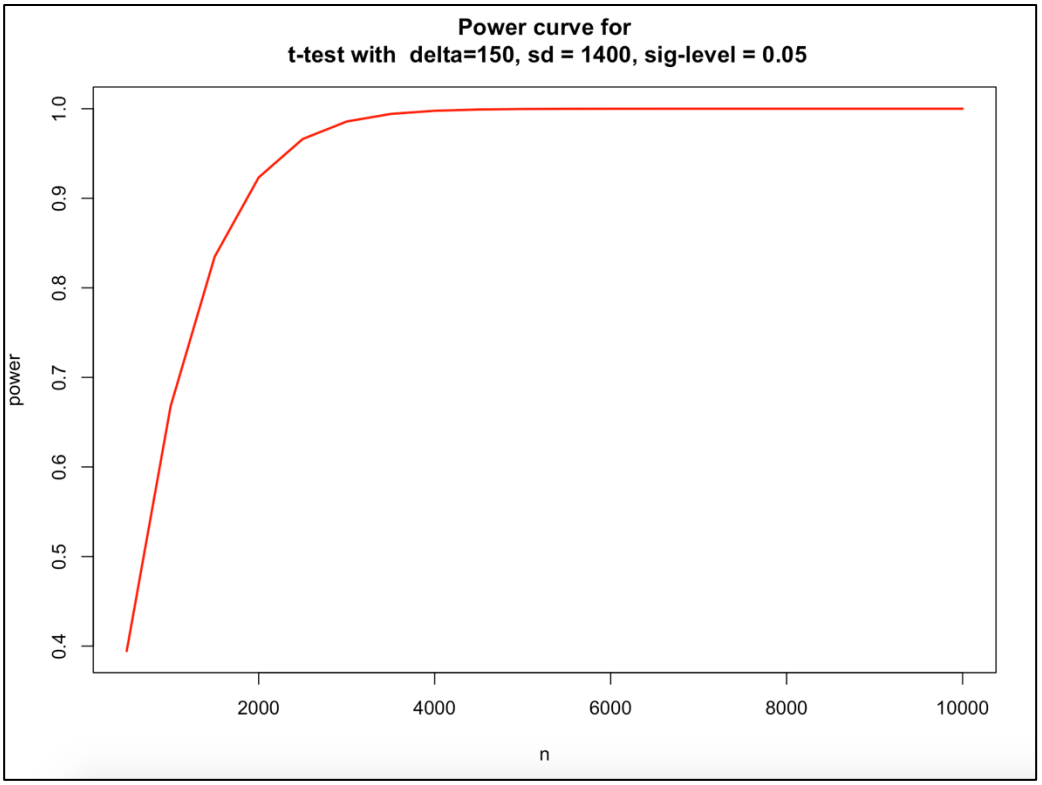


**Figure 7. The relationships between power and sample size**

Figure 8 presents the minimum value of the difference in means in the treatment group (MDE) that we could detect with sample sizes ranging from 500 to 10,000. Both figures show that the estimated sample size is a function of assumptions we make.

```
deltas <- c(100, 150, 200, 250)
plot(nvals, seq(0,1, length.out=length(nvals)), xlab="n", ylab="power",
     main="Power Curve for\nt-test with varying delta", type="n")
for (i in 1:4) {
  powvals <- sapply(nvals, function (x) power.t.test(n=x, delta=deltas[i], sd=1400, sig.level = 0.05)$power)
  lines(nvals, powvals, lwd=2, col=i)
}
legend("bottomright", lwd=2, col=1:4, legend=deltas)
```

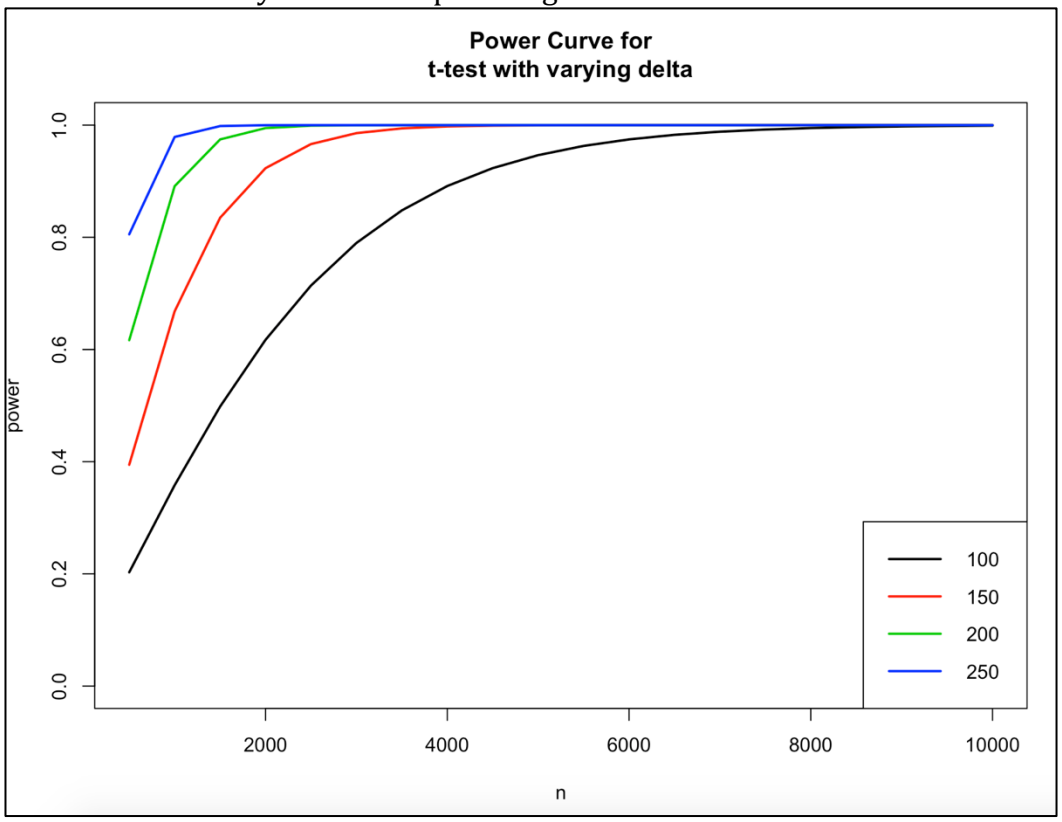


**Figure 8. The relationship between MDE and sample size**

In addition to the parameters described above, the following parameters affect the sample size, power and MDE:

- ✓ Variance of the outcome of interest in treatment and control groups (larger variances imply higher sample sizes)
- ✓ Whether we need one-sided and two-sided test (one sided test needs a smaller sample)
- ✓ Proportion of sample in treatment and control groups (50-50 allocation results in minimum sample size)

**Standardized effect size**

In the R demonstrations above we used a case of assumed means and standard deviations of outcomes in the treatment and control group. What if we don't know the means or the standard deviations? In such situation, we can take advantage of "standardized" outcome variables. A variable $Y$ can be standardized such that it has a mean 0 and standard deviation of 1 as,

$$Y_{std} = \frac{Y - mean(Y)}{SD(Y)}$$

Then, for the control group, the mean is 0 and the common standard deviation of the outcome in both groups is 1. Cohen (1988) provides some guidelines for "standardized" effect sizes based on a meta-analysis, given estimates of the mean and standard deviation of the outcome. For example, an intervention with a change of 0.2 standard deviations is considered a "small" impact, a 0.5 standard deviation change is considered medium impact, and a 0.8 standard deviation change is considered a large impact. Based on these guidelines, assumptions, or secondary data analysis, we can express the mean outcome in the treatment group in terms of a standard deviation change from the mean of the

control group (=0). For example, to design a sample able to identify weak impacts (0.2 SD changes) from the control mean of 0, we will specify following command in R:

```
power.t.test(n = 1000, delta = 0.2, sd = 1, sig.level = 0.05,
             power = NULL,
             type = "two.sample",
             alternative = "two.sided",
             strict = FALSE, tol = .Machine$double.eps^0.25)
```

# 5. POWER ANALYSIS FOR CLUSTER RANDOMIZED CONTROL TRIALS

So far, we have assumed that the examined intervention is randomized at the individual levels. However, often interventions are randomized at a cluster level and all eligible individuals in the cluster receive the intervention. For example, PROGRESA was implemented at—and thus randomized at—the village level. The main implication of this type of designs is that observations are no longer ***independent and identically distributed.*** Since individuals within a group are exposed to the same treatment, it is very likely that their outcomes are going to be correlated within the group or cluster. Individuals in the same cluster may also share other common facilities and infrastructure, and their socio-economic and cultural factors could be correlated.

Consider two drastic examples:

- ✓ The outcomes and covariates for all individuals in a given cluster are correlated with correlation coefficient of almost 1. In this case, whether we measure 1 or 100 individuals from each village, we get the same information.
- ✓ If outcomes and covariates for individuals in a village are not at all correlated, then each individual adds "information" to our sample.

Therefore, the amount of variability we observe in our sample depends on the degree of similarity between individuals within each cluster and between different clusters. This is captured in a parameter called ***Intra-Cluster-Correlation*** (ICC). Higher ICCs imply higher required sample sizes. We calculate the increase in the sample size required, compared to the sample size estimated in the individual RCT, as:

$$Design\ Effect\ (DE) = 1 + \sqrt{(m-1)\rho}$$

where $m$ is the number of individuals included in the sample per cluster and $\rho$ is the ICC. We then

multiply the sample size estimated for individual RCT by DE and divide by $m$ to obtain the number of clusters required. For example, if we decide that $m$ is 40 per village (there exist guidelines about how to select $m$, but this is often left up to the researcher), then for different $\rho$, we will get different DEs as follows:

- ✓ $\rho = 0.05$ → DE = 1.72
- ✓ $\rho = 0.5$ → DE = 4.53
- ✓ $\rho = 0.95$ → DE = 6.17

Higher correlation within clusters thus leads to higher required sample sizes. We will not cover this topic further in this module, but the computation of DE becomes more complicated when we select the sample in multiple stages (such as by districts, by villages, and then by households.

# 6. BIBLIOGRAPHY/FURTHER READINGS

1. Bloom, Howard (1995), "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs," *Evaluation Review*, 19(5), 547-556, October.
2. Cohen, Jacob (1988). *Statistical Power for the Behavioral Sciences*. Second Edition. Lawrence Erlbaum Publishers.
3. Duflo, Esther; Rachel Glannester and Micheal Kremer (2008). "Using Randomization in Economic Development Research: A Toolkit," *Handbook of Development Economics*, Vol. 4, Elservier Science.
4. Gertler, Paul J., Sebastian Martinez, Patrick Premand, Laura B. Rawlings, and Christel MJ Vermeersch. "Impact evaluation in practice." World Bank Publications, 2011.
5. Power Analysis in R: http://www.evolutionarystatistics.org/document.pdf