Chapter 7: Power analysis

Pre-requisites

- Chapter 1: Intro to STATA
- Chapter 2: Review of Regression
- Chapter 3: Experiments
- Chapter 4: Problems with Experiments
- Chapter 5: Regression Discontinuity Designs
- Chapter 6: DID

Contents

1. Introduction

In previous chapters, we have covered different techniques to analyze the impact of an intervention. In this chapter, we return to the issue of inference. So far, we have been paying attention to the topic of statistical significance. We were interested in defining a tool to evaluate whether an estimated relationship is due to sampling variability. In this chapter, we will focus on a different aspect of inference: power. We will study whether an evaluation design has the ability to detect an impact if this impact actually exists.

Power analysis is a critical element of an evaluation design. For instance, we want to design an evaluation with a sample size large enough to be able to detect a difference between treatment and control. If the sample size is too small, it is hard for the researcher to evaluate whether a non-statistical significant result is due to either a true no effect or lack of power. Since evaluations are generally costly, it is a critical matter to define a sample size consistent with a level of power that minimizes the chance of no finding an effect in the sample when the effect is true in the population.

In power analysis, we are interested in defining a design to be able to compare means of treatment and control groups. This comparison of means is the critical element of power. This differs from the standard sampling design in which the interest is comparing a sampling mean with a population counterpart for a given level of error. Discussing this type of sampling goes beyond the goal of this chapter but it is required for what follows in the course. Therefore, we urge students to cover the basics of sampling using the materials from the online course prepared by Sistemas Integrales, particularly the chapter on sampling. The link is the following:

http://lsms.adeptanalytics.org/course/fscommand/session3/Ses3_eng.html

In this chapter, we will cover the following issues:

- Define the basic elements of power analysis paying special attention to its key components: minimum detectable effect, power and sample size.
- Estimate the components of power analysis for individual randomized design using scalars or implementing them in the official STATA commands.
- Evaluate the consequences of changes in the basic parameters of the power design.
- Estimate the components of power for the case of cluster randomized design.

At the end of this chapter we expect students to be able to:

- Implement the estimation of the elements of power analysis for the individual and randomized design cases.
- Evaluate the sensitivity of the basic results to changes in the basic parameters.

2. Motivation

To motivate power, let's consider a simple simulation. 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 creating this artificial program, we know in advance what the true impact of this intervention is in this population. Let's assume that the true impact is 0.1 or, in other words, that those with college education have a return 10% in weekly earnings with respect to those in the control group. We also assume that the mean of income for the control group (the constant) is equal to 5 (or 148 dollars per week). The goal of this simulation is to evaluate how sample size matters in evaluating whether there is no impact of an intervention. Anytime you find -using a sample of observations- that the coefficient of the treatment variable is not significant, it is possible that this is a reflection of a true no impact in the population but also it can be due to the lack of enough observations in your sample to be able to detect a difference between treatment and control units. Therefore, we create a simple simulation in which a true impact of the intervention is defined for the population and then take samples of different sizes to evaluate whether we are able to recover the true impact of the intervention with these samples.

We first create a dataset with 3 million observations. Then, we proceed by creating the error term `u`, which we assume is standard normal. We create then the treatment variable `edu_random`. Since we want this variable to be randomized, we need first to create an auxiliary variable (`aux_random`) with random number with zero mean and then assign the program to those in the population with positives values. Using the values for `u`, `edu_random` and the constant, we create the value of the outcome `lny_random`, the natural log of weekly earnings. Finally, we use a regression to confirm that the parameters of the simulation are recovered using the information for the whole population. The STATA code is below:

```
* Hypothetical program: 3 million of participants
*-----------------------------------------------

clear all

set seed 521

*Creating a dataset
set obs 3000000

*Creating variables
```

```
generate u = rnormal(0)

gen aux_random=rnormal(0)

generate edu_random=0
replace edu_random=1 if aux_random>0
drop aux_random

generate lny_random = 5 + 0.1*edu_random + u

*Checking true impact: regression

regress lny_random edu_random
```

The regression output is the following:

```
. regress lny_random edu_random

      Source |       SS       df       MS              Number of obs = 3000000
-------------+------------------------------           F(  1,2999998) = 7645.62
       Model |  7639.90493     1  7639.90493           Prob > F      =  0.0000
    Residual |  2997756.072999998  .999252691          R-squared     =  0.0025
-------------+------------------------------           Adj R-squared =  0.0025
       Total |  3005395.982999999  1.00179899          Root MSE      =  .99963

------------------------------------------------------------------------------
  lny_random |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  edu_random |   .1009284   .0011543    87.44   0.000     .0986661    .1031907
       _cons |   4.999916    .000816  6127.05   0.000     4.998317    5.001516
------------------------------------------------------------------------------
```

The results are recovered as expected. The impact of the intervention is 0.1 and the constant 5.

We now take a sample of 3,000 observations from the original population and re-estimate the impact of the intervention from this sample. We use the command `preserve` to made temporary changes in the dataset and the command `restore` to recover the original population. Using the command `sample` we take samples from the population and then proceed with the regression. The code is the following:

```
*Sample size:3000
*----------------

preserve

sample 3000, count

*Testing impact: regression

regress lny_random edu_random
```

```
restore
```

The regression output is the following:

```
. regress lny_random edu_random

      Source |       SS       df       MS              Number of obs =    3000
-------------+------------------------------           F(  1,  2998) =    8.90
       Model |  8.56254841     1  8.56254841            Prob > F      =  0.0029
    Residual |  2883.82628  2998  .961916703           R-squared     =  0.0030
-------------+------------------------------           Adj R-squared =  0.0026
       Total |  2892.38882  2999  .964451092           Root MSE      =  .98077

------------------------------------------------------------------------------
  lny_random |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  edu_random |   .1068677    .035819     2.98   0.003     .0366354    .1771001
       _cons |   5.002025   .0255632   195.67   0.000     4.951902    5.052149
------------------------------------------------------------------------------
```

Notice that we essentially get the same results in terms of the estimated impact. The relationship is statistically significant at 1% confidence level but now it less precise in the sense that the confidence intervals include estimated returns from 0.03 to 0.17. Notice that we were lucky to find a sampling estimate almost the same as the true population parameter but this does need to be the case because of sampling variability.

We can replicate the experiment for a sample size of 300 observations. The code is below:

```
*Sample size:300
*---------------

preserve

sample 300, count

*Testing impact: regression

regress lny_random edu_random

restore
```

The output is the following:

```
. regress lny_random edu_random

      Source |       SS       df       MS              Number of obs =     300
-------------+------------------------------           F(  1,   298) =    0.25
       Model |  .253603988     1  .253603988            Prob > F      =  0.6152
    Residual |  298.464191   298  1.00155769           R-squared     =  0.0008
-------------+------------------------------           Adj R-squared = -0.0025
```

```
       Total |  298.717795   299   .99905617           Root MSE      =  1.0008

------------------------------------------------------------------------------
  lny_random |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  edu_random |   .0582793   .1158176     0.50   0.615    -.1696447    .2862034
       _cons |   4.997034   .0845812    59.08   0.000     4.830582    5.163486
------------------------------------------------------------------------------
```

Now the estimated coefficient is 0.05, statistically insignificant (p-value of 0.615). Using this sample, one improvised research might have concluded that the program has not impact when in fact it does. We know in advance that the program has a true impact, so the problem with this estimation is that the sample size is too small to be able to detect the true impact of 0.1. In other words, the sampling design has not enough **power**. Of course, a typical researcher does not know the true impact of the intervention in the population. Therefore, he can infer that the program has no impact when it does, something known in the statistical literature as **type II error**. In practice, an experienced researcher would simply say that the estimates are too noisy to conclude that the lack of significance can be interpreted as evidence of no impact of the intervention.

We can run one extra simulation assuming a sample size of 50. The code is below:

```
*Sample size:50
*--------------

preserve

sample 50, count

*Testing impact: regression

regress lny_random edu_random

restore
```

The output is the following:

```
. regress lny_random edu_random

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =    1.54
       Model |  1.51977246     1  1.51977246           Prob > F      =  0.2204
    Residual |  47.3252651    48  .985943023           R-squared     =  0.0311
-------------+------------------------------           Adj R-squared =  0.0109
       Total |  48.8450376    49  .996837501           Root MSE      =  .99295

------------------------------------------------------------------------------
  lny_random |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  edu_random |    .355876    .286639     1.24   0.220    -.2204504    .9322023
```

```
       _cons |   4.820268   .1812864    26.59   0.000     4.455767    5.184769
------------------------------------------------------------------------------
```

We still find that the program has no impact but the estimates are noisier. The confidence intervals are too wide to be taken seriously. The estimated coefficient is far from the true impact in terms of magnitude.

3. Conceptual issues

In this section, we cover conceptual issues related to power analysis. We first relate what we already cover about statistical significance and connect this with the concept of power. Then, we provide a graphical explanation of the concept of power. We conclude this section with a more formal treatment.

3.1. Hypothesis testing and power

So far the emphasis in terms of statistical inference was given to the concept of statistical significance. We started by defining a null hypothesis and deriving a distribution for a test statistic consistent with the null hypothesis. Using this test statistic, we compute the probability of observing a null effect by chance (sampling variability). Typically, it is assumed that 95% of this distribution is consistent with the null effect. The other 5% is still consistent with the null hypothesis but, since these values are located in the extremes, the convention is to interpret them as inconsistent with the null. This implies that, anytime we perform hypothesis testing, we are willing to commit a mistake since we interpret an extreme value of the t-statistic as inconsistent with the null hypothesis when in fact can be compatible with it. This is what we call **type I error**. We reject a null effect when in fact is true.

Power is related with an alternative concept of error known as **type II error**. This type of error occurs when we fail to reject the null hypothesis (concluding that there is no difference between treatment and control groups) when in fact is false. This is exactly the case of the example in section 2. We knew that the intervention had an impact in the population (therefore, the null hypothesis was false) but we were not able to reject the hypothesis of no impact due to a low sample size. **Power** is simply the probability of detecting an effect when this effect exists.

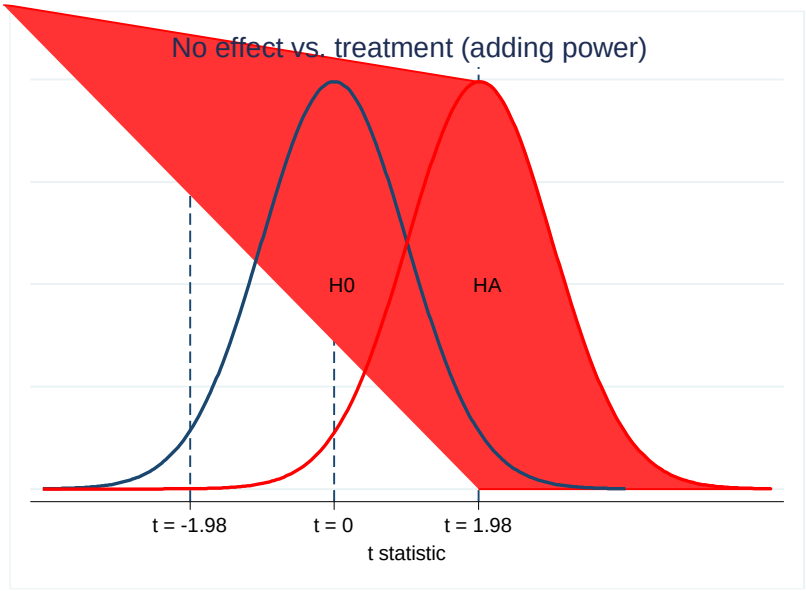
3.2. Graphical analysis

We can illustrate the concept of power using a graph (Figure 1). We plot two distributions: one consistent with the null (no impact of the intervention) and one consistent with a positive impact of the program (also known as the alternative hypothesis). We add to the graph of the null distribution of the values for the t-statistic for the 95% confidence interval. Since we assumed a sample size of 100 units, these are defined by the values for the t-statistic 1.98 and -1.98.

Imagine a difference between treatment and control groups consistent with a t-statistics of 1.5. That particular difference can belong to either distribution in Figure 1. More importantly, it is part of the 95% confidence interval. Therefore, we cannot reject the null. Now consider a t-statistic of 1.99. This value is still consistent with the null but, as explained above, we are willing to tolerate the possibility of committing a type I error for these extreme values. We reject the null which implies that it more likely

that this difference in means belong to the distribution of a positive impact. Notice that at 1.98 the distribution consistent with the alternative hypothesis is divided in two areas: the one below 1.98 is consistent with the null hypothesis; therefore all the differences in means below that value won't be statistically significant. Above 1.98, all the values would be consistent with a rejection of the null, implying that all the difference in means in this area will be statistically significant. The area in red above 1.98 in Figure 1 represents power since in all these points we are able to reject the null when in fact there is an effect since these points belongs to the distribution of the alternative hypothesis.

Figure 1: Power



As in the case of hypothesis testing, we need to set a standard about a minimum level of power for our research design. The most common choices are 80% and 90%. This means that, if we were able to replicate the experiment a large number of times, we will be able to detect an impact when this exists in 80% (90%) of the cases.

3.3. Formal treatment: Minimum detectable effect, power and sample size

To formally address power, we start by a simple regression to evaluate the impact of an intervention that followed an individual randomized design. The equation is the following:

(1) $Y_i = \alpha + \beta D_i + \varepsilon_i;$

where D_i is the treatment variable and ε_i is an error term, both defined for individuals. Y_i is the outcome of interest and β recovers the impact of the intervention. If it is assumed that the observation are independent and with the same distribution, the variance of the treatment effect can be written as follows:

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

Where σ^2 is the variance of the outcome of interest is, N is the sample size and P is the fraction of the sample assigned to the treatment group. Since the level of power is the area above that falls to the right of 1.98, it can be shown that β has to be higher than the standard error multiplied by the t-statistics for the statistical significance and the inverse of power (Duflo et al 2008), which defines what it is known as the **minimum detectable effect** (MDE). The MDE is defined in the following way:

$$(3)\ MDE(k,\alpha,N,P) = (t_{(1-k)} + t_{\alpha}) * \sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}$$

where $t_{(1-k)}$ is the t-statistic associated to the inverse of the level of power and t_{α} is the t-statistic associated to the significance level. For a level of power of 80% and a significance level of 5%, these values are 0.84 and 1.96 using a normal approximation to a t-distribution.

To interpret the MDE, it is important to compare it to a given standard. For instance, assume that after comparing similar programs in Latin America, you conclude that your new cash transfer program should have an impact of 100 Mexican pesos. Then, for a given sample size and level of power, you estimate a MDE of 150. This implies that your design won't able to detect an impact because your estimated MDE implies that only impacts of 150 or higher can be captured with the current sample size and level of power. On the other hand, if your analysis of similar or previous programs tells you that at least an impact of 200 Mexican pesos is expected, then your current design will be able to detect a difference between treatment and control units since the expected change is higher than the estimated MDE.

Notice that power and sample size are parameters in the computation of the MDE. We can reorganize the terms in (3) to obtain expressions for the level of power and sample size. For instance, **power** can be computed using the following expression:

$$(4)\ t_{(1-k)}(N,\alpha,\beta_E,P) = \frac{\beta_E}{\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}} - \tau_{\alpha};$$

where β_E is the effect size. The effect size is just 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 make the difference between these two just to emphasize that the MDE is an estimate whereas the effect size is a parameter.

We can also obtain an expression in terms of the **sample size**. The equation is the following:

$$(5)\ N(k,\alpha,\beta_E,P) = \left[\frac{\sigma * (t_{(1-k)} + t_{\alpha}) * \sqrt{\frac{1}{P(1-P)}}}{\beta_E}\right]^2;$$

where all the terms have been previously defined. It is important to note that these 3 elements are critical components of power analysis and that computing one of them requires the other two to be given. Hence, computing the MDE implies to previously define the level of power and the sample size. Computing power requires a sample size and the effect size. Finally, estimating the sample size is not possible without specifying the level of power and the effect size. Of course, there are other parameters like the significance level or the proportion of the sample allocated to treatment and control groups. We will analyze how the key elements of the power analysis (MDE, power and sample size) are affected by the choice of values for these parameters.

4. Power analysis for individual randomized designs

We discuss in this section the implementation of power analysis for individual randomized designs using the conceptual framework covered in the previous section. We first upload the dataset:

```
use "C:\Users\Stanislao\Dropbox\Teaching\1. Current\Econometrics\4. Handbook\2. Data\ENCEL 2007\Final\DataFinal_power.dta", clear
```

We have two options to implement this in STATA. The first one consists in using scalars and directly computes each element of the formulas for the MDE, power and sample size. The second option uses directly the `power` command.

Assume we are interested in evaluating the impact of a cash transfer on household income and you are required to perform the power analysis for an evaluation design in which households are randomly assigned to the program. The new program will be implemented in 2014 and you have access to information of a similar program for 2007. As it was previously discussed, we need information about the mean and the variance of the outcome of interest, so we compute that from the 2007 survey and record the results in scalars we can use later for the estimation of each element of the formula for the MDE, power or sample size, depending on the case. The STATA code is below:

```
* 1. MEANS AND VARIANCE OF IMPACT INDICATOR

* Variance and standard deviation

sum IncomeLabHH1, detail

scalar mean_income = r(mean)              /*Mean*/

scalar var_income = r(Var)                /*Variance*/

scalar stddev_income = r(sd)              /*Standard deviation*/
```

The results are below:

```
. sum IncomeLabHH1, detail

              HH Primary Monthly Income: Pesos
-------------------------------------------------------------
```

```
      Percentiles      Smallest
 1%          100             1
 5%          600             4
10%         1050             4       Obs                17593
25%         1600             4       Sum of Wgt.        17593

50%         2400                     Mean           2585.186
                     Largest         Std. Dev.      1402.329
75%         3000            9600
90%         4400            9600     Variance        1966528
95%         5400            9600     Skewness       1.141613
99%         7500            9600     Kurtosis       5.433854
```

The mean household income is 2,585 Mexican pesos with a standard deviation of 1,402. We use these values to represent the mean and standard deviation of the control group. Recall that, if you want to compute power or sample size, you need to specify the effect size. Therefore, we assume that the intervention would have an impact of 0.2 standard deviations or 280 Mexican pesos. Then, the mean for the treatment group would be 2,865 pesos. The code is below:

```
* Means for treatment and control units (Assumption: standardized effect size
of 0.2)

scalar control_income=mean_income

scalar treat_income=control_income+0.2*stddev_income

local control_m = control_income
local treat_m = treat_income
local sigma_m = stddev_income
```

Notice that we also use locals to save the same information. The reason is that we want to use this information to implement the canned programs and they don't allow the use of scalars.

If the goal is to compute the MDE or power, we need to specify the standard error and that requires defining the sample size. We assume that 1,000 households can be interviewed. The code is below:

```
* 2. COMPUTING STANDARD ERRORS: (sigma^2/N)^(1/2)

scalar n_sample = 1000  /*Assuming known sample size and we need to compute
power or MDE*/

scalar st_error = (var_income/n_sample)^(1/2)
```

We also need to define the proportion of treatment and control units in the sample. We assume that the 50% of the sample is composed by treated households. The code is below:

```
* 3. PROPORTION OF TREATMENT AND CONTROL UNITS: (1/(P(1-P)))^(1/2)

local p=0.5                          /*P: proportion of treatment units*/
```

```
scalar p_exp=(1/(`p'*(1-`p')))^(1/2)
```

Finally, we need to define the levels for t-statistics for alpha and beta. We use a normal approximation rather than the exact t-values for simplicity. Given the sample sizes we usually use on impact evaluation, this decision does not affect the results and simplifies the computation. These values assume a significance level of 5% and a level of power of 90%. The code is below:

```
* 4. T-VALUES FOR ALPHA (SIGNIFICANCE LEVEL) AND 1-K (POWER) *

scalar t_alpha=invnormal(0.975)           /*Note: 0.95 if one-sided test*/

scalar t_beta=invnormal(0.90)

scalar t_alphaplusbeta=t_alpha+t_beta
```

Using these parameters, we are in the position of computing the different components of the equations discussed above. Before that, let's display the values of the most relevant parameters. The code is below:

```
* Parameters

display control_income
display treat_income

display p_exp

display t_alpha
display display t_beta
```

The results are below:

```
. * Parameters

. display control_income
2585.1857

. display treat_income
2865.6516

. display p_exp
2

. display t_alpha
1.959964

. display t_beta
1.2815516
```

We already presented the means for treatment and control units. In the case of the expression associated to the proportion of treatment and controls in the sample, the value is 2. The values for t-statistics for the significance and power levels are 1.96 and 1.28.

4.1. Minimum detectable effect (MDE)

We start by computing the MDE for this case. Recall that computing the MDE requires making assumptions on the level of power and the sample size. We assume a level of power of 90 and a sample size of 1,000 observations. All the other parameters are as the ones defined above.

With this information, we directly compute each element of the MDE formula. The code is below:

```
* 5. OPTION 1: COMPUTING MDE (N and power given) *

/* FORMULA OF MDE:

MDE={t_(1-k)+t_alpha}*{(1/(P(1-P)))^(1/2)}*{(sigma^2/N)^(1/2)}

*/

scalar mde_abs=t_alphaplusbeta*p_exp*st_error

display mde_abs
```

The result is the following:

```
. display mde_abs
287.49358
```

Assuming a power level of 90% and a sample size of 1,000 observations, we estimate a MDE of 287 Mexican pesos. This means that, with given sample size and level of power, we are able to distinguish any impact of at least 287 pesos or more. If the true impact is lower than 287 pesos, we won't be able to detect an impact with the given sample size and power level.

An alternative is to use the STATA `power` command. We use the option `twomeans` because we are using a t-test for a comparison of two means. We need to specify the mean of the outcome and the standard deviation for the control group. Since we have already saved that information in locals, we call these locals. We also specify the level of power with option `power( )` and the sample size with option `n( )`. The STATA code is the following:

```
*Using POWER command (STATA 13)

power twomeans `control_m', sd(`sigma_m') alpha(0.05) power (0.9) n(1000)
```

The STATA output is the following:

```
. power twomeans `control_m', sd(`sigma_m') alpha(0.05) power (0.9) n(1000)

Performing iteration ...
```

```
Estimated experimental-group mean for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1; m2 > m1

Study parameters:

        alpha =    0.0500
        power =    0.9000
            N =      1000
  N per group =       500
           m1 = 2585.1857
           sd = 1402.3294

Estimated effect size and experimental-group mean:

        delta =  287.7706
           m2 = 2872.9563
```

STATA reports the MDE under the name of delta. As you see, the outcome is the same as the one we got using our manual approach. A MDE of 287 Mexican pesos was estimated.

4.2. Power

Assume now that you want to estimate the level of power. We know that the sample size and the effect size are given in this case. Assuming a sample size of 1,000 households and effect size of 280 Mexican pesos (0.2 standard deviations), we use scalars to compute power using the power formula above. The STATA code is below:

```
* 6. OPTION 2: COMPUTING POWER (N and effect size given) *

/*FORMULA OF POWER:

t_(1-k)=[delta/{(1/(P(1-P)))^(1/2)}*{(sigma^2/N)^(1/2)}]-t_alpha

*/

*Assume effect size (delta=280 mexican pesos or 0.2 std deviation)

scalar delta_n=0.2

scalar delta_a=delta_n*stddev_income

display delta_a

scalar t_power= (delta_a/(p_exp*st_error))- t_alpha

local t_power=t_power
```

```
scalar power=normal(`t_power')

display power
```

The STATA output is below:

```
. display power
.88537899
```

The result implies that, with a sample size of 1,000 observations and an effect size of 280 Mexican pesos, we have a level of power of 88% or 0.88. We compare this level with the standards usually applied in the literature, typically 80% or 90%. Therefore, if we set the standard in 80%, we would have a decent level of power. On the other hand, if we set power in 90%, we will marginally below the desired level of power for our design.

We can implement the same analysis using the `power` command in STATA. The effect size is captured by adding the information of the mean for treatment and control group. The rest of the code is the same as above. The STATA routine is the following:

```
*Using POWER command (STATA 13)

power twomeans `control_m' `treat_m', sd(`sigma_m') alpha(0.05) n(1000)
```

The result is the following:

```
. power twomeans `control_m' `treat_m', sd(`sigma_m') alpha(0.05) n(1000)

Estimated power for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1

Study parameters:

        alpha =    0.0500
            N =      1000
  N per group =       500
        delta =  280.4659
           m1 = 2585.1857
           m2 = 2865.6516
           sd = 1402.3294

Estimated power:

        power =    0.8848
```

We confirm the results.

4.3. Sample size

We focus now the estimating the sample size for a given level of power and effect size. We assume a level of power of 90% and effect size of 280 Mexican pesos. We just use scalars to compute the elements of the formula for sample size. The STATA code is below:

```
* 7. OPTION 3: DETERMINATION OF SAMPLE SIZE (for a given effect size and
power) *

/*FORMULA OF SAMPLE SIZE (n):

n = [sigma*{t_(1-k)+t_alpha}*{(1/(P(1-P)))^(1/2)}/delta]^2

*/

scalar n_sample=[stddev_income*t_alphaplusbeta*p_exp/delta_a]^2

display n_sample
```

The output is below:

```
. display n_sample
1050.7423
```

We have estimated a sample size of 1,050 observations to be able to detect a difference between treatment and control households of 280 Mexican pesos with a level of power of 90%. The same result is obtained using the `power` command. The code is below:

```
*Using POWER command(STATA 13)

power twomeans `control_m' `treat_m', sd(`sigma_m') alpha(0.05) power(0.9)
```

The STATA output is the following:

```
. power twomeans `control_m' `treat_m', sd(`sigma_m') alpha(0.05) power(0.9)

Performing iteration ...

Estimated sample sizes for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1

Study parameters:

        alpha =    0.0500
        power =    0.9000
        delta =  280.4659
           m1 = 2585.1857
           m2 = 2865.6516
```

```
            sd = 1402.3294

Estimated sample sizes:

             N =      1054
  N per group =       527
```

Notices that the numbers are not exactly the same due to the use of decimals but they are pretty close each other.

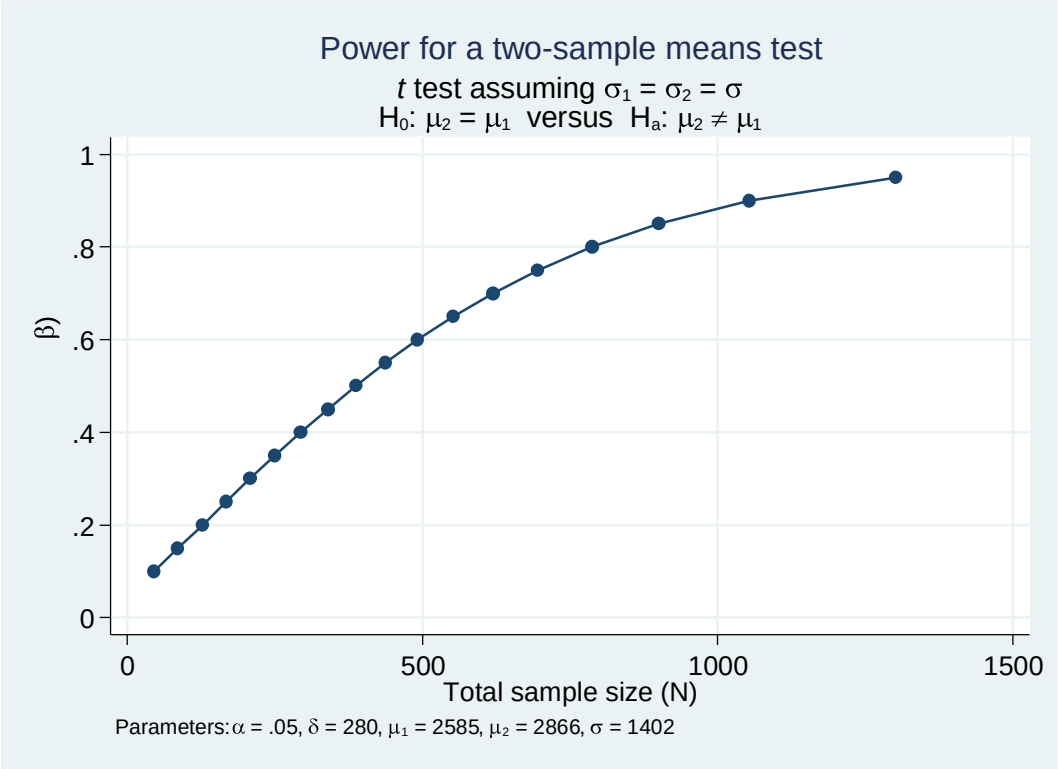
4.4. Power curves

Before finishing this section, it is worth to introduce power curves, a graphical tool that related the level of power with sample size. The `power` command includes the option `graph(y(power))`, which create the graph. Since we want to estimate the sample size for several levels of power, holding constant the effect size, we need to tell STATA to do so. Therefore, we use the option power in a different manner. Instead of using just one level of power, we accommodate several values in the following way:

```
* 8. POWER CURVE *

power twomeans `control_m' `treat_m', sd(`sigma_m') alpha(0.05)
power(0.1(0.05)0.95) graph(y(power))
```

Notice that we introduce the values for power with the option `power(0.1(0.05)0.95)`. The graph is the following:



We have now a tool to evaluate how the sample size changes as we change the level of power.

5. Factors affecting power analysis

We discuss in this section how the basic results discussed in the previous section change when the assumptions are changed. We focus on power for simplicity and to avoid repetition, but the same logic applies to MDE and sample size. Only when we use power as a parameter in the analysis, we will look to its impact on sample size.

We repeat here the estimation of the mean and the standard deviation for the control group and replicate the baseline results for the MDE. The code is below:

```
scalar control_income=mean_income
display control_income

scalar treat_income02=control_income+0.2*stddev_income
display treat_income02

local control_m = control_income
local treat_m02 = treat_income02

local sigma_m = stddev_income

*Baseline case
power twomeans `control_m' `treat_m02', sd(`sigma_m') alpha(0.05) n(1000)
```

The results are the following:

```
. power twomeans `control_m' `treat_m02', sd(`sigma_m') alpha(0.05) n(1000)

Estimated power for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1

Study parameters:

        alpha =    0.0500
            N =      1000
  N per group =       500
        delta =  280.4659
           m1 = 2585.1857
           m2 = 2865.6516
           sd = 1402.3294

Estimated power:

        power =    0.8848
```

Therefore, using the parameters from the previous section, we have a level of power of 0.88. This is the benchmark we will use to evaluate how results changes as a consequence of a modification in the values of the parameters of the power analysis.

5.1. Effect size

We start discussing the role of effect size. In our previous example, we considered an effect size of 280 Mexican pesos (a change of 0.2 standard deviations). We consider instead a change of 140 Mexican pesos. The new mean for the treatment group is 2,725 Mexican pesos. We keep all the other parameters in their original values. The STATA code is below:

```
* A. EFFECT SIZE

scalar treat_income01=control_income+0.1*stddev_income
display treat_income01
local treat_m01 = treat_income01
```

The output is the following:

```
. display treat_income01

2725.4186
```

Instead of using scalar, we use the `power` command for simplicity. The code is the following:

```
power twomeans `control_m' `treat_m01', sd(`sigma_m') alpha(0.05) n(1000)
```

The resulting output is below:

```
. power twomeans `control_m' `treat_m01', sd(`sigma_m') alpha(0.05) n(1000)

Estimated power for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1

Study parameters:

        alpha =    0.0500
            N =      1000
  N per group =       500
        delta =  140.2329
           m1 = 2585.1857
           m2 = 2725.4186
           sd = 1402.3294

Estimated power:

        power =    0.3520
```

The new estimated power is 0.35 or 35%, way below the usual standard (80%). This an indication that power is highly sensitive to changes in the effect size. A reduction in 0.1 standard deviations in the outcome of interest leads to change in the level of power from 88% to 35%.

5.2. Sample size

We proceed in a similar way with the case of power. We use all the values of the parameters for the benchmark case except the sample size. We consider a reduction in the sample size of 500 households. The STATA code is below:

```
* B. SAMPLE SIZE

power twomeans `control_m' `treat_m02', sd(`sigma_m') alpha(0.05) n(500)
```

We obtain the following results:

```
. power twomeans `control_m' `treat_m02', sd(`sigma_m') alpha(0.05) n(500)

Estimated power for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1

Study parameters:

        alpha =    0.0500
            N =       500
  N per group =       250
        delta =  280.4659
           m1 = 2585.1857
```

```
           m2 = 2865.6516
           sd = 1402.3294

Estimated power:

        power =    0.6071
```

A reduction of 50% of the sample size leads to reduction from 88% to 60%.

5.3. Power level

Consider now the role of power. Of course, we need to evaluate its role with respect sample size or the MDE. We choose sample size to illustrate the point. We replicate the benchmark case and study what is the impact of reducing the level of power to 80%. The code is the following:

```
* C. POWER LEVEL

power twomeans `control_m' `treat_m02', sd(`sigma_m') alpha(0.05) power(0.9)

power twomeans `control_m' `treat_m02', sd(`sigma_m') alpha(0.05) power(0.8)
```

The results are reported below:

```
.   power   twomeans   `control_m'   `treat_m02',   sd(`sigma_m')   alpha(0.05)
power(0.9)

Performing iteration ...

Estimated sample sizes for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1

Study parameters:

        alpha =    0.0500
        power =    0.9000
        delta =  280.4659
           m1 = 2585.1857
           m2 = 2865.6516
           sd = 1402.3294

Estimated sample sizes:

            N =      1054
  N per group =       527

.   power   twomeans   `control_m'   `treat_m02',   sd(`sigma_m')   alpha(0.05)
power(0.8)
```

```
Performing iteration ...

Estimated sample sizes for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1

Study parameters:

        alpha =    0.0500
        power =    0.8000
        delta =  280.4659
           m1 = 2585.1857
           m2 = 2865.6516
           sd = 1402.3294

Estimated sample sizes:

            N =       788
  N per group =       394
```

Originally, the sample size required to be able to detect a difference of 0.2 standard deviations with a power level of 90% was 1,054 households. Since we are reducing the level of power to 80%, we being less strict regarding the probability of committing type II error, and therefore less sample size is required. We only need 788 households to get the same effect size with a power of 80%.

5.4. Variance of the outcome

We have covered the critical parameters for power analysis. We paid special attention to MDE, power and sample size as the key elements of power analysis. We now discuss with some detail how the results of the exercise are modified when the other parameters are changed.

We start with the variability of the outcome of interest. We know from the conceptual section that more dispersion is associated with less power. We consider an increase of the standard deviation by a factor of 1.5, keeping the values of all the other parameters as in the benchmark case. The STATA code is the following:

```
* D. OUTCOME VARIANCE

local sigma_m2 = stddev_income*1.5

power twomeans `control_m' `treat_m02', sd(`sigma_m2') alpha(0.05) n(1000)
```

The results are presented below:

```
. power twomeans `control_m' `treat_m02', sd(`sigma_m2') alpha(0.05) n(1000)

Estimated power for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1
```

```
Study parameters:

        alpha =    0.0500
            N =      1000
  N per group =       500
        delta =  280.4659
           m1 = 2585.1857
           m2 = 2865.6516
           sd = 2103.4941

Estimated power:

        power =    0.5581
```

We find that power reduces when the outcomes has more variability. In this case, the new level of power is 55%.

5.5. One-sided and two-sided test

We assumed that a two-sided test for statistical significance. We can use a one-sided test instead. In STATA, it is very straightforward to implement this using the option `onesided`. The code is the following:

```
* E. ONE-SIDED VS TWO-SIDED TEST

power twomeans `control_m' `treat_m02', sd(`sigma_m') alpha(0.05) n(1000)
onesided
```

The result is the following:

```
. power twomeans `control_m' `treat_m02', sd(`sigma_m') alpha(0.05) n(1000)
onesided

Estimated power for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 > m1

Study parameters:

        alpha =    0.0500
            N =      1000
  N per group =       500
        delta =  280.4659
           m1 = 2585.1857
           m2 = 2865.6516
           sd = 1402.3294

Estimated power:
```

```
        power =    0.9351
```

We observe an increase in the level of power from 88% to 93%.

5.6. Proportion of sample in treatment and control groups

We assumed that the sample size was equally split between treatment and control units. Although this is the most efficient case, there are situations in which equal samples are not possible. Assume that 75% of your sample is assigned to treatment group. STATA allows the user to specify the number of units to be assigned to each group. The code is below:

```
* F. PROPORTION OF TREATMENT AND CONTROLS IN THE SAMPLE

power twomeans `control_m' `treat_m02', sd(`sigma_m') alpha(0.05) n1(750)
n2(250)
```

The results are the following:

```
. power twomeans `control_m' `treat_m02', sd(`sigma_m') alpha(0.05) n1(750)
n2(250)

Estimated power for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1

Study parameters:

        alpha =    0.0500
            N =      1000
           N1 =       750
           N2 =       250
        N2/N1 =    0.3333
        delta =  280.4659
           m1 = 2585.1857
           m2 = 2865.6516
           sd = 1402.3294

Estimated power:

        power =    0.7811
```

We obtain a lower level of power 78%.

6. Standardized effect size

So far, we paid attention to the case in which the effect size was defined in absolute values. For instance, we were interested in an effect size of 280 Mexican pesos for a new cash transfer program. How we justify choosing 280 pesos over, for instance, 300 or 400? Choosing the effect size implies to

have an idea about the impact of similar programs around the world but this information is not presented in terms of Mexican pesos. Therefore, for comparison purposes, the impact of similar interventions can be expressed in terms of standard deviations, so researchers can use information from similar interventions in other parts to get a reasonable estimate of the effect size of a new program. This is especially the case when no pre-treatment data is available or when the program is launched for a first time in a particular setting. A standardized effect δ can be expressed in the following manner:

(6) $$\delta = \frac{\beta}{\sigma} = \frac{Y_T - Y_C}{\sigma}$$

The formula for the MDE for the standardized case is the following:

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

Cohen (1988) has proposed standards for effect sizes based on meta-analysis. According to this author, an intervention with 0.2 standard deviations change is considered a "small" impact. If the change is 0.5 standard deviations, then the intervention has a "medium" impact, whereas a 0.8 standard deviation change is a "large" impact. Therefore, if the researcher lacks of pre-treatment data, these guidelines can be useful to perform power analysis.

To implement this in STATA, we set the standard deviation to be equal to 1. Then, the mean for control group is set to 0 and the mean for treatment group is set to the number of standard deviations we expect the program to change the outcome. In this case, let's assume a change of 0.2 standard deviations in the household income. The STATA code is below:

```
*-------------------------*
* STANDARIZED EFFECT SIZE *
*-------------------------*

power twomeans 0 0.2, sd(1) alpha(0.05) n(1000)
```

The result is the following:

```
. power twomeans 0 0.2, sd(1) alpha(0.05) n(1000)

Estimated power for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1

Study parameters:

        alpha =    0.0500
            N =      1000
  N per group =       500
        delta =    0.2000
           m1 =    0.0000
```

```
            m2 =    0.2000
            sd =    1.0000

Estimated power:

         power =    0.8848
```

We got the same level of power as before (88%), which is not surprising since we define the effect size in absolute terms (280 Mexican pesos) after adding 0.2 standard deviations in household income to the mean of the outcome.

7. Power analysis for cluster randomized designs

So far we have assumed that the intervention is implemented at individual level. However, it is very common that social programs are randomized at a group level. For instance, instead of individuals/households, PROGRESA was assigned at village level. Most cash transfers shared these characteristics. An implication of this type of designs is that observations are no longer independent and identically distributed. Since individuals/units within a group are exposed to the same shock (treatment), it is very likely that their outcomes are going to be correlated within the group or cluster. As a consequence, less variability will be observed since individuals from the same cluster are more similar, which reduces power.

7.1. Conceptual issues

Let's start again defining the equation of interest:

(8) $Y_{ij} = \alpha + \beta D_{ij} + \upsilon_j + \omega_{ij}$

The standard error can be written in the following way:

(9) $SE(\beta) = \sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{m\tau^2 + \sigma^2}{mJ}};$

Where m is the number of units (individuals/households) per cluster, J is the number of clusters, τ^2 is the variance across clusters and σ^2 is the variance across units. The MDE for a cluster randomized design follows the same structure as in the individual case. We will need to multiply the expression $(t_{(1-k)} + t_{\alpha})$ in order to have an expression for it.

An alternative is to express the cluster randomized case as a variant of the individual case by comparing the standard errors. In particular, the ratio between the standard errors is known as the **Design effect (D)**. The expression is the following:

$$D = \frac{SE(\beta)_{cluster}}{SE(\beta)_{individual}} = \frac{\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{m\tau^2+\sigma^2}{mJ}}}{\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\tau^2+\sigma^2}{mJ}}} = \sqrt{1+(m-1)\rho};$$

where ρ is the **intra-cluster correlation (ICC)**. The ICC has the following structure:

$$\rho = \frac{\tau^2}{\tau^2+\sigma^2}.$$

The ICC is the percentage in the outcome variation due to cluster level characteristics.

It can be shown that the MDE for a cluster randomized design can be obtained by multiplying the MDE for the individual randomized case by the design effect. The equation is the following:

(10) $$MDE_{cluster} = MDE_{individual} * D = (t_{(1-k)} + t_{\alpha}) * \sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{mJ}} * \sqrt{1+(m-1)\rho}.$$

This is a very important result since we can use the estimation for the individual randomized case as a benchmark. We will only need to compute the ICC and the design effect to estimate the MDE for the cluster case using the previously estimated MDE for the individual randomized design.

7.2. Estimating the intra-cluster correlation (ICC)

Assume for this section that clusters are defined at village level and households are the unit of analysis. We start by estimating the ICC. There are several strategies to do that in STATA. The simplest way is estimating the ICC using the analysis of variance (ANOVA). We use the command `loneway` and add the variables that contain the information of household income and village identifier. The code is the following:

```
* A. ESTIMATING ICC

* Alternative 1: Computing rho from ANOVA

loneway IncomeLabHH1 villid
scalar rho = r(rho)
display rho
```

The results are the following:

```
. loneway IncomeLabHH1 villid

One-way Analysis of Variance for IncomeLabHH1: HH Primary Monthly Income:
Pesos

                                              Number of obs =     17593
                                                  R-squared =    0.2229
```

```
    Source                SS         df      MS            F     Prob > F
------------------------------------------------------------------------------
Between villid         7.712e+09    693     11128189      7.00     0.0000
Within villid          2.688e+10  16899    1590823.2
------------------------------------------------------------------------------
Total                  3.460e+10  17592    1966527.8

         Intraclass       Asy.
         correlation      S.E.       [95% Conf. Interval]
         -------------------------------------------------
            0.19156     0.01376       0.16460     0.21853

         Estimated SD of villid effect            613.961
         Estimated SD within villid              1261.278
         Est. reliability of a villid mean        0.85705
              (evaluated at n=25.30)

. scalar rho = r(rho)

. display rho
.19156093
```

This result implies that 19% of the variability of household income is explained by variability across villages and the rest for variability across households.

An alternative consists in using to estimate the ICC using maximum likelihood. We first use the command `xtmixed` and then the user-written program `iccvar`. The code is below:

```
* Alternative 2: Computing rho using Maximum Likelihood

findit iccvar

quiet xtmixed IncomeLabHH1 ||  villid: , var
iccvar
```

The results are reported below:

```
. quiet xtmixed IncomeLabHH1 ||  villid: , var

. iccvar

Intraclass Correlation Estimates

------------------------------------------------------------------
             |    ICC      Std. Err.      [95% Conf. Interval]
-------------+----------------------------------------------------
      villid |  0.18707     0.01091       0.16569      0.20845
------------------------------------------------------------------
```

The estimated ICC is pretty similar to the one previously calculated.

7.3. Computing the MDE

Using the estimated ICC, we can proceed with the calculation of the design effect using the formula previously provided. Since we have households and villages, the sample size is going to be defined by the number of villages we cover and the number of households from each of these villages. Therefore, computing the design effect requires defining the number of observations to be collected in each village. We set this value in 20, following some standards used regular household surveys. Using this information and the estimated ICC (rho), we estimate the design effect as follows:

```
* B. COMPUTING MDE

scalar m_sample=20

scalar d_effect=(1+(m_sample-1)*rho)^(1/2)

display d_effect
```

The output is below:

```
. display d_effect
2.1539865
```

This implies that, whatever the MDE is for the case of an individual randomized design, it will be multiplied by a factor of 2.15. This suggests that, compared to an individual randomized design, the MDE for a cluster randomized design is larger.

Recall that the MDE for the individual randomized design was estimated in 287 Mexican pesos. We can tell STATA to report that value again using the `display` command as below:

```
. display mde_abs
287.49358
```

We directly estimate the MDE for the cluster randomized design by simply multiplying the original MDE for the individual randomized design with the design effect. The code is below:

```
scalar mde_abs_cluster=t_alphaplusbeta*p_exp*st_error*d_effect

display mde_abs_cluster
```

The results are the following:

```
. display mde_abs_cluster
619.25728
```

We have estimated a MDE for a cluster randomized design of 619 Mexican pesos, assuming a level of power of 90% and a sample size of 1,000 observations. Since we are assuming that 20 households were taken from each village, 50 villages are included in the sample. Notice that, whereas in the

individual randomized design we were able to detect an impact of a least 287 Mexican pesos, we can only detect impacts of at least 619 Mexican pesos in cluster randomized design.

7.4. Changes in the ICC

As in the case of the individual randomized design, we can also evaluate how the MDE results are affect when changes in the basic parameters are allowed. We focus here in the ICC and number of units per cluster. Since all the other parameters behave in the same way as in the individual randomized case, we just focus in these two.

We assume now that the ICC is 0.4. Using that value, we re-estimate the design effect. The code is below:

```
* C. CHANGES IN THE ICC

scalar rho_04=0.4

scalar d_effect04=(1+(m_sample-1)*rho_04)^(1/2)

scalar mde_abs_cluster_04=t_alphaplusbeta*p_exp*st_error*d_effect04

display mde_abs_cluster_04
```

The result is the following:

```
. display mde_abs_cluster_04
843.09668
```

We estimated a larger MDE. Since a higher ICC implies that units within a given cluster are more correlated to each other, the intuition is that less information is obtained from each cluster compared to the benchmark case. As a consequence, a bigger MDE is associated with this design.

7.5. Changes in the number of observations per cluster

Since we have two different levels (units and clusters), the total sample size is going to depend on both levels. In our previous example, we considered 50 villages (clusters) and 20 households (units). Now, we consider a scenario in which we increase the number of households per cluster to be equal to 50 and reduce the number of villages to 20. This modifies the design effect previously estimated and, as a consequence, the MDE for cluster randomized design. The STATA code is below:

```
* D. CHANGES IN NUMBER OF OBSERVATIONS PER CLUSTER

scalar m_sample50=50

scalar d_effect50=(1+(m_sample50-1)*rho)^(1/2)

scalar mde_abs_cluster_50=t_alphaplusbeta*p_exp*st_error*d_effect50
```

```
display mde_abs_cluster_50
```

The output is reported below:

```
. display mde_abs_cluster_50
926.53636
```

We estimate a MDE larger than the baseline case. This shows one critical insight for the trade-off between the number of units per cluster and the number of clusters. It is better include more clusters than units per cluster in terms of power since there is more variability across clusters than within clusters. Of course, including more clusters is more expensive, so researchers need to balance these two dimensions.

8. Final remarks

In this chapter, we have covered the basics of power analysis. We paid attention to the details of the individual randomized designs and the basics of the cluster randomized case. There are several relevant aspects that we were not able to cover due to space restrictions, including the accommodation of baseline data, the use of covariates and adjustments for multiple comparisons.

9. Further readings

Bloom, Howard (1995), "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs," *Evaluation Review*, 19(5), 547-556, October.

Cohen, Jacob (1988). *Statistical Power for the Behavioral Sciences*. Second Edition. Lawrence Erlbaum Publishers.

Duflo, Esther; Rachel Glannester and Micheal Kremer (2008). "Using Randomization in Economic Development Research: A Toolkit," *Handbook of Development Economics*, Vol. 4, Elservier Science.