Chapter 3: Experiments

Pre-requisites

- Chapter 1: Intro to STATA
- Chapter 2: Review of Regression

Contents

1. Introduction

In the previous two chapters, we were interested in introducing STATA along with a basic refresh of the linear regression model and hypothesis testing. The goal was pretty general, so we think the material previously covered could be helpful for any researcher interested in empirical economics. This chapter is the first one specifically related to impact evaluation.

In this chapter, we are going to show you how to use STATA to implement the design and analysis of a randomized controlled trial. As we have discussed in lecture, randomization is considered the "gold standard" of impact evaluation. This method has a set of nice properties that allow evaluators to construct a good counterfactual since it guarantees statistical balance between treatment and control units. In this scenario, one researcher can use those not assigned to treatment status to represent what would have happened to those assigned to treatment in case they were not treated, a feature usually absent in non-experimental research in which selection bias is prevalent.

Consider the following equation:

(1) $y_i = \alpha + \beta D_i + \mu_i$

Randomized assignment to treatment solves the selection bias problem present in observational studies. In terms of equation (1), randomization guarantees that all unobservable factors that can affect outcomes are balanced between treated (individuals with D=1) and control (those with D=0) such that the only factor that can impact outcomes is the treatment itself. In this scenario, we can uncover the causal relationship between treatment and the outcomes of interest via a simple comparison of means

because all potential factors that can cause bias in our estimates have been balanced out due to randomization.

During this chapter, you will be trained in the use of STATA to perform several tasks related to the design and analysis of randomized experiments. In terms of the evaluation design, we will teach you how to randomize treatment for the most common designs used in practice. Using a set of covariates, we will also show how to test pre-treatment balance between program participants and no participants. After covering these implementation issues, we will discuss how to analyze experimental data to evaluate the impact of the intervention.

Since this is one of the first chapters, we keep the discussion at a basic level. Obviously, this is not the manner researchers usually evaluate programs in practice. We are going to discuss in the last section of this chapter what other issues researchers explore during the analysis of an impact evaluation just as a matter of illustration. Over the next chapters, we will provide you more details about these topics.

We also cover a more detailed discussion about causal inference. The goal here is to offer a complementary discussion about the issue beyond the introductory discussion you may find in basic textbooks. We discuss causality and introduce the potential outcomes framework. We derive the parameters of interest in impact evaluation and discuss the issue of selection bias. We then relate that to standard regression analysis. We discuss then how experiments solve the selection bias problem. These sections are optional in the sense that most of these details are no needed to follow the STATA section, but we recommend it for those interested in a detailed explanation.

Although we routinely will use regressions to evaluate the impact of an intervention in an experimental design, some famous statisticians like the late David Freedman (professor at Berkeley until 2008 when he passed away) have criticized the use of regression to analyze experimental data under the argument that randomization is not enough to justify regression assumptions, particularly in the case of multivariate regression. We study alternative ways to test the impact of an intervention using t-tests for difference in means and non-parametric test of differences in distributions. We take advantage of this opportunity to explore the details behind t-tests for difference in means. So far, we have paid attention to t-tests for population parameters and the discussion in this chapter would be a good background for topics to be covered in the future such as power analysis.

We complement our discussion with a more depth treatment on related topics in the appendixes to this chapter. We use simulations to illustrate selection bias under the most common forms of endogeneity including omitted bias, reverse causality and measurement error in Appendix 3.1. We build on our Argentinean example from the previous chapter and use the actual estimates of a simple regression of income on education to see the magnitude of bias under different forms of endogeneity. We write some very simple programs and simulate the outcomes to measuring the extent and direction of selection bias. We also cover a more advanced randomization technique, pair-wise matching, in Appendix 3.2. This approach is becoming increasingly popular among researchers, so its inclusion in our review is well deserved.

In the following pages, we will use the same dataset based on the ENCEL 2007 survey; the survey used to evaluate the impact of Oportunidades, the famous Mexican cash transfer program. Readers

interested in more details about the way we constructed this dataset should review the previous chapter. We will use the ENCEL 2007 survey just for pedagogical purposes to illustrate the design of an evaluation, specifically alternative ways to randomize treatment and check for treatment balance. To discuss how to analyze experimental data to evaluate the impact of an intervention, then we will use the actual experimental data from PROGRESA. We will use a collection of cross-sectional datasets from several rounds of the original ENCEL survey for years 1997, 1998 and 1999.

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

- Understand the role of selection bias in estimating the causal effect of an intervention and to get familiar with the potential outcomes framework, the most influential theory of causality in the treatment effects literature.
- Perform random assignment to treatment using the most common approaches of the modern literature.
- Check balance between treatment and control groups.
- Test the impact of an intervention.

2. Causal Inference and Impact Evaluation

2.1. From the CEF to causal estimates

In the previous chapter, we introduced regression as a tool to estimate conditional expectation functions (CEF). We showed that regression provides a linear estimate to the CEF under very weak conditions (Godgerber 1991 and Angrist and Pischke 2008). This very powerful result is valid without any reference to causality. Therefore, researchers interested in the association between variables can always run linear regressions and, provided they correctly interpret regression coefficients, learn about the behavior of conditional means.

However, in impact evaluation we are interested in knowing whether a program has an effect in an outcome. Think in the case of a cash transfer program and assume for a moment that you don't know how the program was assigned to treatment and control groups. We want to make sure that the effect of a cash transfer on enrollment (the outcome of interest) is due exclusively to the cash and rather than a product of other factors. For instance, it is possible that mother's education play a critical role in explaining children's enrollment. Then, failing to control for this factor, would bias our estimate of the impact of the program since those children with more educated mothers would be performing better in terms of enrollment, independently of the impact of the program itself. The coefficient associated to the treatment variable in a linear regression would recover the true causal estimate plus a bias due to differences in terms of mother's education between treatment and control groups.

In this scenario, we want to establish conditions under which regression coefficients can be interpreted in a causal manner. In our previous example, this would imply that mother's education can explain both access to the program and enrollment. For instance, more educated mothers are more aware of availability of social programs, therefore they are more likely to enroll in a program of this kind. They are also more likely to send their children to the school. Therefore, a regression of access to a

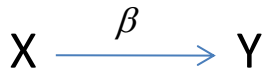
cash transfer and enrollment is likely to be cofounded by mother's education. The estimated impact of the program would be biased.

In this context, a regression will still estimate a CEF and establish a level of association between the treatment and enrollment, but arguably this is not enough if the goal is to decide whether to extend the coverage of a program or decide about the use of government budget among alternative welfare-improving policies. Therefore, it is critical to understand what a regression might be estimating and the potential biases that could be present in practice when a researcher tries to estimate the impact of an intervention.

2.2. Correlation does not imply causation

In impact evaluation we are interested in causal estimates of an intervention. To illustrate the point, we discuss now the basic problem of estimating the effect of changing the value of an explanatory variable X on a dependent variable Y. In our previous example, X is a cash transfer program and Y is enrollment. In particular, we pay attention to the conditions under which a causal interpretation for β (the coefficient of a regression of treatment status on enrollment) is provided.

Figure 1



In Figure 1, β is the causal effect of X on Y? We will argue later that this will be true in the case of a randomized experiment under which X is experimentally manipulated by the researcher. Without such an experiment, it is highly likely that a situation like Figure 2 can better describe what actually happens in reality. Assuming that X, Y, W and Z are observed variables (where W and Z are other factors that may affect X and Y) and that u and v are unobserved variables, this figure summarizes pretty much all the potential problems that an analyst will face when doing empirical impact evaluation.

Figure 2

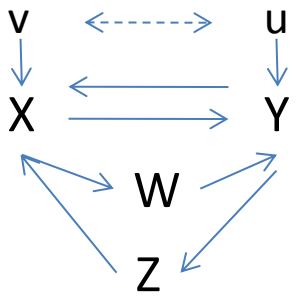


Figure 2 illustrates a key idea about the interpretation of causal relationships: cofounding. Cofounding can be explained by a third observable dimension like W or Z. Following our example above, mothers' education is observed and therefore it can be controlled for in a regression. Figure 2 also shows the most common problem in empirical impact evaluation: **omitted variable bias**. In this case, X

and Y are correlated by a third unobserved variable u. Following our previous example, it can be argued that the observed relationship between a cash transfer program (X) and enrollment (Y) may reflect differences in mother's ability (u) across treated and non-treated households rather than a real causal effect of a cash transfer on enrollment. More talented mothers are more likely to enroll the program and to send their children for basic education. Since mothers' ability is unobserved by the evaluator and correlated with take up of a cash transfer program, then there is correlation between the error term and our treatment variable leading to a problem of spurious correlation[1].

Other common problem in impact evaluation is **reverse causality**. In this case, Y affects X. For instance, it can be program officials target households with children with low levels of enrollment. When X affects Y and Y affects X at the same time we have a **reverse causality** or **simultaneity** problem. Other typical example in economics is the feedback between education and income. It should be obvious that education have a positive effect on income but it is also true that individuals with higher income can invest more on education. This simultaneity makes hard to attribute causality to a regression between income and education.

A third important problem in impact evaluation –no well captured by Graph 2- is **measurement error**. It is hard to explain intuitively how this leads a bias in the econometric analysis of an impact evaluation, but it can be shown formally that this problem causes an underestimation of the true effect of the program. Measurement error is not uncommon in impact evaluation since this type of studies usually implies collecting data in the field. If some households are incorrectly classified as treated or the information about the actual treatment status of a household is poorly recorded, then measurement error may create problems to evaluate the impact of the intervention.

In Appendix 1.1, we show the implications of these problems in terms of the estimation of the true effect of a program using some basic derivation and simple simulations in STATA.

3. Approaches to Causal Inference in Impact Evaluation

The estimation of a causal estimate is always based on a theory of causality. In impact evaluation, there are two competing approaches: a) the traditional **structural approach** (Havelmo 1943, Heckman 2000) and, b) the modern **potential outcomes framework** or experimental approach, also known as the Neyman-Rubin-Holland potential outcome model (Rubin 1974, Holland 1986 and Neyman 1923). Holmes (2010) offers a nice description of both approaches:

"The structural approach in empirical economics starts with a fully-specified economic model, well grounded in theory. The goal of the approach is the estimation of the underlying "deep model parameters" of preferences and technology. Once the model parameters are obtained, we have an artificial economy in hand that we can put to work, simulating the impact of various policy alternatives. Importantly, we can even use the estimated model to study the impacts of polices that have never yet been implemented. The structural approach can be contrasted with the experimentalist approach that originated in the labor economics literature. In this approach, the identification of the causal impact of a policy treatment is the goal. But rather than look through the lens of economic theory, it relies on finding

[1] It is important to note that no all omitted variable in the analysis lead to a bias. There exists a bias only when the omitted variable is correlated with X.

"natural experiments" or clever instruments to tease out the impacts of policies that have already taken place. Finding these experiments enable the researcher to treat behaviors as exogenous that would otherwise have to be treated as endogenous."

We will pay attention here to the potential outcomes framework, the dominant theory of causality in the modern treatment effects literature. A short explanation of the structural approach is discussed in Box 1. The next section discusses the potential outcomes framework.

Box 1: Structural modeling in impact evaluation

In this box we briefly discuss the traditional approach in impact evaluation, largely due to the work of econometricians associated with the Cowles Commission. This section is largely based on Heckman (2000).

In the structural approach, causality is defined within a well-specified economic model about the impact of a policy or intervention. Particularly, this approach formalizes the idea developed originally by Alfred Marshall that a "**ceteris paribus**" change is closely related to a notion of causality in which causal effects are approached by mean of comparative statics exercises. Other key contribution of this group of econometricians is the formalization of the idea that many models are consistent with the same data implying the some additional restrictions must be placed on models in order to recover causal parameters.

To illustrate the idea, let's assume that you are hired to evaluate the impact of a cash transfer program on learning outcomes measured by test scores. Under a structural approach, you need a theoretical model to link educational inputs (books, classrooms, teachers, etc.) and other resources (including the cash transfer and other inputs provided by parents) to educational outcomes. Researchers in education usually define an **education production function** to model this relationship. Let's consider the following simple education production function:

$$Y = F(X_1, ...X_N);$$

where F is well-behaved production function in which each X_i represents an input. In this context, each X_i represents a ***cause***. If each X_i can be varied independently (clearly, a strong assumption), a causal effect can be represented in the following way (assuming differentiability):

$$\frac{\delta Y}{\delta X_j} = F_j(X_1, ..., X_N) \Big/ X = x$$

Notice that the definition of a causal relationship in this model does not depend on what we can actually observe from the data. What it is required is a hypothetical model and the assumption that X_j can be varied independently. Under this model these parameters are known as **structural parameters** and represent cause-effect relationships.

More generally, a structural model consists of the following elements (Cameron and Trivedi 2005):

- A set of variables W, composed by endogenous (Y) and exogenous (Z) variables.

- A joint probability distribution of W.
- A priori ordering of W according to hypothetical cause-effect relationships and a set of restrictions in the hypothesized model.
- A specification of the functional form.

This may seem too theoretical for our purposes but it is important to have it in mind. What really matters is to understand that –in this framework- causality is theoretically grounded. This contrasts with the potential outcome framework in which the role of theory is less important.

Using theoretical models along experimental data is a frontier area in impact evaluation, implying that these two models of causality are converging. For instance, Todd and Wolpin (2006) use experimental data from PROGRESA to estimate a dynamic model of child schooling and fertility. The reason of this movement is that researchers are more interested not only on knowing whether a program works but also what factors explain why this program works. To explore these factors, a model is usually needed.

4. Potential Outcomes Framework

This section is largely based on Holland (1986), Angrist et al (2009) and Morgan et al (2007). The model was proposed originally by Neyman (1923) and further developed by Rubin (1974). We introduce here the basic terminology.

Assume that i is an index for individuals in a given population. We define a variable D_i as the treatment or potential cause of which we want to estimate the effect. For simplicity, we assume that the treatment variable is binary, so $D_i = 1$ if individual i has been exposed to treatment and $D_i = 0$ if individual i has not been exposed to treatment. We also define $Y_i(D_i)$ is the outcome or the effect we want to attribute to the treatment. Since the treatment is binary, we define two potential outcomes; $Y_i(1)$ is the outcome in case of treatment and $Y_i(0)$ is the outcome in case of no treatment.

Note that the outcome for each individual can be written as follows:

(2) $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$

Or simply, $Y_i = Y_i(1)$ if $D_i = 1$ and $Y_i = Y_i(0)$ if $D_i = 0$. This means that we only observe the outcome for each individual in one specific state (with treatment or with no treatment).

4.1. The fundamental problem of causal inference

So far, we have defined for each individual two potential outcomes: one in the case of treatment and one in the case of no treatment. As a consequence, if it were possible to observe an individual in both treatment statuses in the same moment of time, it would be easy to compute the causal effect of an intervention by simply comparing her outcomes in both states. The definition of a causal effect would be the following:

Box 2: Definition of causal effect

For every individual i, the causal effect of $D_i = 1$ is $\Delta_i = Y_i(1) - Y_i(0)$.

The problem with this definition is that we don't observe the same individual being exposed to treatment and not treatment at the same time. This is the so-called "fundamental problem of causal inference". This is stated in the following way:

Box 3: Fundamental problem of causal inference (Holland 1986)

It is not possible to observe for the same individual i the values $D_i = 1$ and $D_i = 0$ as well as the values $Y_i(1)$ and $Y_i(0)$. Therefore, it is not possible to estimate the effect of D_i on Y_i for each individual i.

The problem of estimating a causal effect is essentially a problem of missing data since we don't observe the treated units in the situation in which they were no treated as well as we don't observe the untreated units in the situation in which they were treated. On other words, we don't observe the counterfactuals for treated and untreated units. The following table summarizes this idea:

Table 1

The Fundamental Problem of Causal Inference

Group	Y(1)	Y(0)
Treatment (D=1)	Observable as Y	Counterfactual
Control (D=0)	Counterfactual	Observable as Y

Source: Morgan and Winship (2007).

We are required to think in terms of "counterfactuals"; i.e what would have happened with a treated individual if he or she would not have received the treatment and viceversa. For the treatment group, we can observe the potential outcome Y(1) but the potential outcome in case of no treatment Y(0) is not observed for this group. For the case of the control group, we observe the potential outcome in the case of no treatment Y(0), but the potential outcome in case of treatment Y(1) is not observed. The key idea in this literature is to establish conditions under which what we observe for the case of the control group can be used to approach the missing counterfactual for the case of the treatment group.

This is a key idea to have in mind during the rest of the course since all the estimators are based on the idea of finding a control group that can represent the missing counterfactual or what would have happened with treated individuals if they were not exposed to treatment.

4.2. Solutions to the fundamental problem of causal inference

Holland (1986) suggests two types of solutions: a) the scientific solution and b) the statistical solution. We are going to focus on the later. You are required to study the first one on your own.

The statistical solution is based on estimating the average effect of the treatment instead of doing so at an individual level. We want to estimate such average effects. The first one in known as **average treatment effect (ATE)**, which is defined as follows:

(3) $ATE = E[\Delta_i] = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)]$

This average effect is still not estimable without further assumptions on the relationship between the potential outcomes $Y_i(1)$ and $Y_i(0)$ with the treatment D_i. From an economic point of view, this effect is not really interesting because is defined for a random person. Notice that we obtain a difference in means after the second equality based on a basic statistical property that states that the mean of a difference is the same as the difference in means.

More interesting for practitioners in impact evaluation is the **average treatment effect on the treated (ATT)**, which is defined as follows:

(4) $ATT = E[\Delta_i / D_i = 1] = E[Y_i(1) - Y_i(0) / D_i = 1] = E[Y_i(1) / D_i = 1] - E[Y_i(0) / D_i = 1]$

As in the previous case, we cannot estimate this parameter without further assumptions. In some particular cases ATE will be equal to ATT, and we will discuss later that this is true when we have a *randomized experiment* or in general in situations that resembles such an experiment as when we have a "natural" experiment.

Notice that we can define also a parameter called the **average treatment effect for the untreated (ATU)**, defined as:

(5) $ATU = E[\Delta_i / D_i = 0] = E[Y_i(1) - Y_i(0) / D_i = 0] = E[Y_i(1) / D_i = 0] - E[Y_i(0) / D_i = 0]$

There are other parameters of interest in the literature. For instance, we will study later the **local average treatment effect (LATE)** in the context of experiments with imperfect compliance. James Heckman has been advocating for the use of the **marginal treatment effect (MTE)**; see for instance Heckman and Vytlacil (2007), which we are not going to cover in this course. The interested student should read Heckman (2005) as a starting point to this new literature.

Notice that we can extend these parameters by conditioning on a set of covariates X. By doing so, we can –for instance- explore how the effect of the treatment varies according to education, income, and other individual characteristics.

4.3. Understanding selection bias

So far we have defined what a causal effect is and discuss the statistical solution by defining average parameters. These parameters are still not estimable since they depend on $E[Y_i(0)]$, an average that is not directly observed by the evaluator.

Let's focus what we actually estimate every time we compare means of the outcome variable between treated and non-treated individuals. We call this the **mean difference in outcomes** (MDO) or naïve estimator:

(6)
$$\begin{aligned} MDO &= E[Y_i / D_i = 1] - E[Y_i / D_i = 0] \\ &= E[Y_i(1) / D_i = 1] - E[Y_i(0) / D_i = 0]. \end{aligned}$$

We compare the outcomes for both groups. Since we know that the observed outcome for the treated group (those with D=1) is equivalent to the potential outcome in case of treatment Y(1), we just replace it in the second line of the equation. Using the same logic, we proceed to replace the potential outcome in the case of no treatment for those untreated (D=0).

It can be shown that the simple difference of outcomes between treated and control units provides a biased estimate of the impact of the program. To see that, consider the following derivation:

$$
\begin{aligned}
MDO &= E\left[Y_i/D_i=1\right]-E\left[Y_i/D_i=0\right] \\
&= E\left[Y_i(1)/D_i=1\right]-E\left[Y_i(0)/D_i=0\right] \\
&= E\left[Y_i(1)/D_i=1\right]-E\left[Y_i(0)/D_i=1\right] \\
&\quad +E\left[Y_i(0)/D_i=1\right]-E\left[Y_i(0)/D_i=0\right] \\
&= ATT+\underbrace{\left\{E\left[Y_i(0)/D_i=1\right]-E\left[Y_i(0)/D_i=0\right]\right\}}_{\text{selection bias}}
\end{aligned}
\tag{7}
$$

The first two lines are exactly the same as in equation (6). In lines 3 and 4, we add the expressions in red that represent the counterfactual for the control group (what would have happened to those untreated if they were exposed to treatment). This does not affect the whole expression since both terms cancel each other because they were entered in the equation with opposing signs but they are useful for our purposes since the two expressions in the third line form the ATT whereas the expressions of the fourth line form what we call **selection bias**. Therefore, each time we compare the outcomes between treatment and control units, we estimate the parameter of interest (in this case ATT) but also an additional term that recovers selection bias. This bias is just another expression to the problem of endogeneity discussed above and illustrated using simulation in the appendices to this document.

We want to find conditions to eliminate this bias, so the simple comparisons of means can recover ATT. We will see in the next section that random assignment to treatment solves the selection bias problem.

4.4. Experiments as a solution of the selection bias problem

There is consensus in the treatment effect literature that random assignment to treatment is the most credible and influential research design because it solves the selection bias problem in a simple way (Angrist and Pischke 2009).

To see this, recall from Table 1 that the observed outcome for treated (control) units is the same as the potential outcome in case of treatment (no treatment):

$$
\begin{aligned}
&E\left[Y_i/D_i=1\right]=E\left[Y_i(1)/D_i=1\right]. \\
&E\left[Y_i/D_i=0\right]=E\left[Y_i(0)/D_i=0\right].
\end{aligned}
\tag{8}
$$

The key question is whether we can approach the counterfactual for the treatment group using the observed outcome for control units or:

(9) $E[Y_i(0)/D_i=0]=E[Y_i(0)/D_i=1]$.

Alternatively, we may want to estimate ATU, in which case we would like to know under which conditions whether the observed outcome for the treatment group can be used to represent the missing potential outcome for those untreated. In other words,

(10) $E[Y_i(1)/D_i=1]=E[Y_i(1)/D_i=0]$

Generally, none of these conditions hold with observational data due to the existence of selection. People select themselves into treatment or control groups due to some observable or unobservable characteristics. However, as we mentioned above, there is an important case in which these conditions are met. That is the case of a **randomized experiment**.

In an experimental design, the treatment D_i is randomly assigned. Because of that, the treatment D_i is independent (or orthogonal) of the potential outcomes $Y_i(1)$ and $Y_i(0)$. In such a situation, the conditional distribution of the potential outcomes conditional on the treatment is equivalent to the unconditional distribution. Therefore,

(11) $E[Y_i(0)/D_i=0]=E[Y_i(0)/D_i=1]=E[Y_i(0)]$

(12) $E[Y_i(1)/D_i=1]=E[Y_i(1)/D_i=0]=E[Y_i(1)]$

Then, we can compute ATE by simply computing;

(13)
$$\begin{aligned} ATE &= E[\Delta_i]=E[Y_i(1)-Y_i(0)]=E[Y_i(1)]-E[Y_i(0)] \\ &= E[Y_i(1)/D_i=1]-E[Y_i(0)/D_i=0]=E[Y_i/D_i=1]-E[Y_i/D_i=0]. \end{aligned}$$

Notice that, since the conditional distribution and the unconditional distribution are the same under a randomized experiment, the following condition is also true:

(14) $ATE=ATT=ATU$.

Therefore, we only need to compute the difference between the average value of Y_i in the treatment group with the average value of Y_i in the control group. This is the reason why randomized experiments are considered the "gold standard" in terms of research design in many areas of social and natural sciences.

4.5. Potential outcomes and regression

One final aspect to cover is the relationship between potential outcomes and regression, the most common tool to analyze the impact of an intervention. To see that, recall equation (2). After a simple re-arrangement we can obtain an expression similar to a regression:

$$(15)\quad \begin{aligned} Y_i &= D_i Y_i(1) + (1-D_i) Y_i(0) \\ &= Y_i(0) + \{Y_i(1) - Y_i(0)\} D_i \end{aligned}$$

This is similar to a simple regression like $Y_i = \alpha + \beta D_i + \varepsilon_i$. Therefore, we can re-write (15) in the following way:

$$(16)\quad \begin{aligned} Y_i &= Y_i(0) + \{Y_i(1) - Y_i(0)\} D_i \\ &= \underbrace{E\{Y_i(0)\}}_{\alpha} + \underbrace{\{Y_i(1) - Y_i(0)\}}_{\beta} D_i + \underbrace{Y_i(0) - E\{Y_i(0)\}}_{\varepsilon_i} \end{aligned}$$

As showed in the equation, the coefficient associated to the treatment variable recovers the difference between potential outcomes, or –in other words- the impact of the program. When we run a regression of the outcome variable on the treatment, we are basically comparing mean outcomes as in the case of a simple t-test of difference in means. To see that, recall the MDO is the following difference: $E[Y_i/D_i = 1] - E[Y_i/D_i = 0]$. In the context of regression, these expectations can be written in the following way:

$$(17)\quad \begin{aligned} E(Y_i/D_i = 1) &= \alpha + \beta + E(\varepsilon_i/D_i = 1) \\ E(Y_i/D_i = 0) &= \alpha + E(\varepsilon_i/D_i = 0) \end{aligned}$$

Therefore, the MDO would be estimated as follows:

$$(18)\quad \begin{aligned} E(Y_i/D_i = 1) - E(Y_i/D_i = 0) = & \underbrace{\beta}_{\text{Treatment effect}} \\ & + \underbrace{E(\varepsilon_i/D_i = 1) - E(\varepsilon_i/D_i = 0)}_{\text{Selection bias}} \end{aligned}$$

This illustrates again the idea discussed above that, in absence of randomization, a regression estimates the causal effect of interest plus a selection bias. When treatment is randomly assigned, this selection bias term becomes zero, and the simple difference in means recovers the true impact of the program. Therefore,

$$(19)\quad E(Y_i/D_i = 1) - E(Y_i/D_i = 0) = \underbrace{\beta}_{\text{Treatment effect}}.$$

5. Hypothesis Testing for Difference in Means

In impact evaluation we usually test the impact of an intervention or the balance between treatment and control units using difference in means. We compare the mean outcome of the treatment group against the mean for the control units and then evaluate whether the difference in means is significant in a statistical sense. This last step implies the use of some form of hypothesis testing for comparing means, something slightly different to the approach we have been using so far.

Although it can be shown that, under certain conditions, a test for comparing means approach is essentially the same as one regarding a population parameter, we believe that discussing this particular case offers some useful insights that would be useful later when we discuss power analysis.

To motivate this discussion, let's return to our Argentinean example. Imagine for a moment that the treatment of interest in higher education and the outcome a measure of income. We want to test whether the average income of those with higher education is different than the average income for those without higher education.

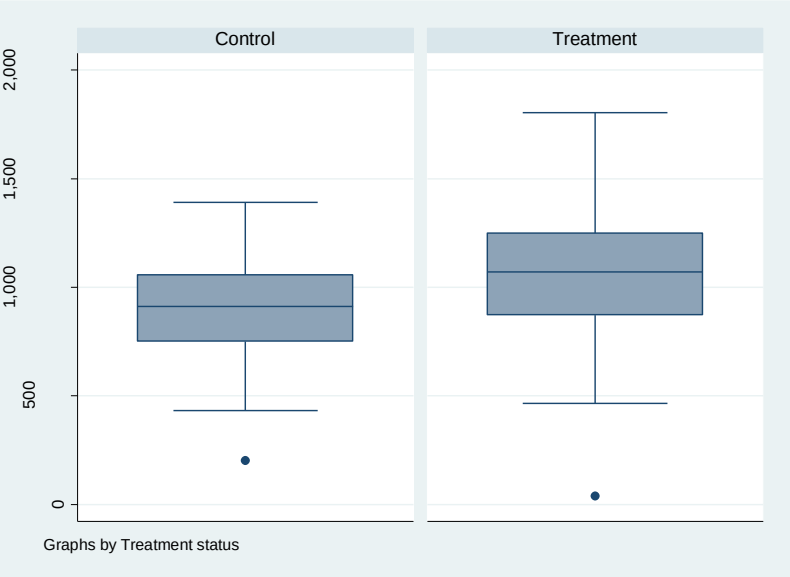
5.1. Logic of t-test for differences in means

We can use a simple example to understand the logic behind a t-test for difference in means. Imagine the following scenario. You ask you research assistant to compute the means and standard deviations for labor income between those with and without higher education. She reports to you that the mean of income for treatment group is 1,050 pesos (standard deviation of 325) and the mean for the control group is 915 pesos (standard deviation of 250). You want to evaluate what can be learnt about the impact of higher education by looking at the distribution of incomes between these two groups using a simple box-plot. Using this information, you write a simple code in STATA to answer that question. The code is below:

```
* SCENARIO 1

clear all
set obs 100

gen group = 1 in 1/50
replace group = 2 in 51/100

label var group "Treatment status"
label define group 1 "Control" 2 "Treatment"
label values group group

gen x=rnormal(915, 250) in 1/50
replace x=rnormal(1050, 325) in 51/100

label var x "Labor income"

graph box x, by(group) graphregion(fcolor(white)) ylabel (0(500)2000)
```

The resulting graph is presented below:

Figure 3: Boxplot of labor income by treatment status (scenario 1)



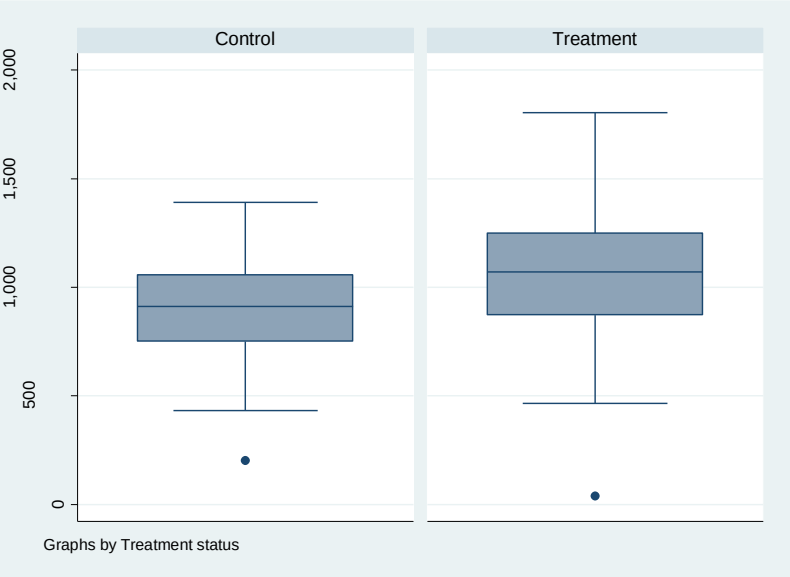
The graph shows that there is an important overlap between both distributions. Although the mean for the treatment group is higher, its distribution is also more disperse. Therefore, it is unclear whether it can be concluded that both means are truly different.

Consider an alternative scenario. Suppose that your research assistant tells you she made a mistake in her first estimates of mean and standard deviations for the treatment and control group. Now she reports that the true mean and standard deviation for the treatment group are 1,260 and 420 respectively. For the control group, she reports 915 and 250. Using this new information, you replicate the previous code. The code is below:

```
* SCENARIO 2

clear all
set obs 100

gen group = 1 in 1/50
replace group = 2 in 51/100

label var group "Treatment status"
label define group 1 "Control" 2 "Treatment"
label values group group

gen x=rnormal(915, 250) in 1/50
replace x=rnormal(1260, 420) in 51/100

label var x "Labor income"

graph box x, by(group) graphregion(fcolor(white)) ylabel (0(500)2000)
```

The resulting graph is the following:

Figure 4: Boxplot of labor income by treatment status



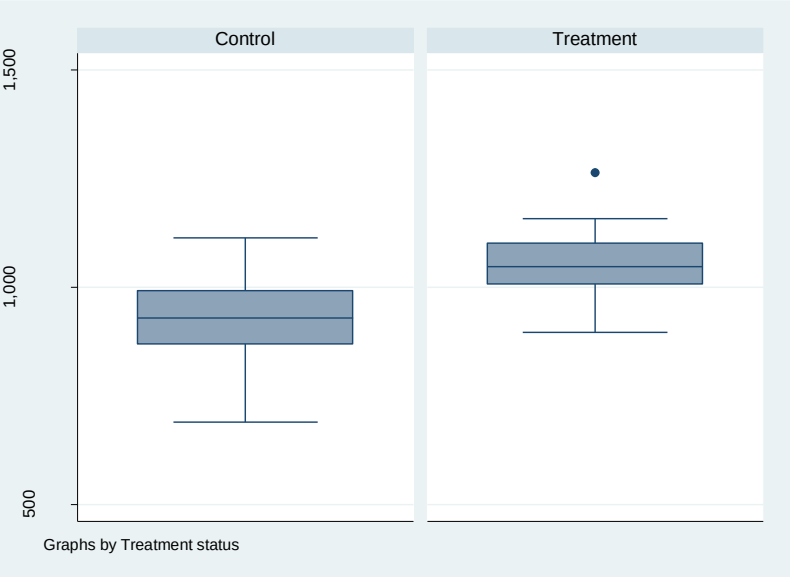
Now, it is a little clearer but still we cannot be sure regarding the difference between treatment and control groups. Therefore, although the difference between means is higher now, the level of dispersion makes still unclear whether those more educated have higher incomes.

Now consider a third scenario. Your research assistant tells you that the means she estimated in the first place were fine, but the problem was with the standard deviations. The standard deviation for the treatment group is 75 whereas the one for the control group is 100. The new STATA code is the following:

```
* SCENARIO 3

clear all
set obs 100

gen group = 1 in 1/50
replace group = 2 in 51/100

label var group "Treatment status"
label define group 1 "Control" 2 "Treatment"
label values group group

gen x=rnormal(915, 100) in 1/50
replace x=rnormal(1050, 75) in 51/100

label var x "Labor income"

graph box x, by(group) graphregion(fcolor(white)) ylabel (500(500)1500)
```

The resulting graph is presented below:

Figure 5: Boxplot of labor income by treatment status



In this case we are confident that there is a difference between treatment and control group is terms of incomes. The reason is that, with lower standard deviations, the dispersion around the mean is less acute. In this scenario, although the means are closer each other compared to the second case, the lower standard deviation allows for a precise estimation of the difference between the means for treatment and control groups.

This simple example illustrates the logic of significance test. Each time we run a simple t-test for difference in means we always take into account differences in terms of means but also difference in standard deviations. The difference between means is usually known as **signal** whereas the difference in terms of standard deviations is known as **noise**. A test of difference in means is just a ratio between signal and noise. The notion is captured in the following formula:

$$(20)\ t = \frac{\overline{Y_T} - \overline{Y_C}}{\frac{s_T}{\sqrt{n_T}} + \frac{s_C}{\sqrt{n_C}}};$$

where $\overline{Y_T}$ and $\overline{Y_C}$ are respectively the means of the outcome (labor income) for treatment (those with higher education) and control group (those without higher education). s represents the standard deviation and n the sample size for a group.

There is a third element in the previous formula that we did not discuss in detail yet: sample size. It played a role during our previous discussion but in an implicit manner. We will explicit this role in the next section.

5.2. Role of sample size

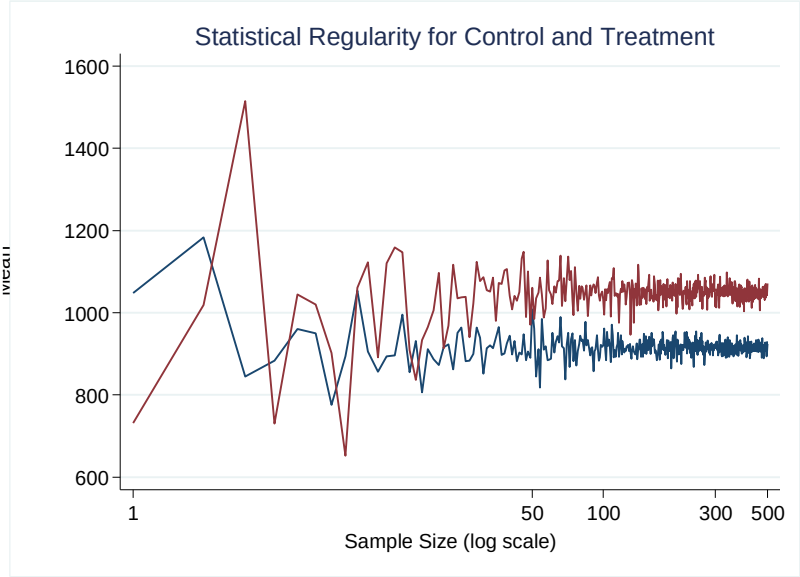
Our intuition dictates that larger sample sizes are related with more precise estimates. We showed in the previous chapter that this was the case for simple simulations of the law of large numbers. We

basically replicate that logic here but now in the context of two different groups. We want to make sure that enough observations are available to estimate a difference between two groups with precision. Using the means and standard deviation for our original Argentinean example, we run a simple simulation in STATA to shed light on this issue. The code is the following:

```
clear
set obs 500
gen nvar1=.
gen pvar1=.
gen nvar2=.
gen pvar2=.
gen var1=.
gen var2=.
set seed 999
forvalues x = 1/500 {
  quietly replace var1=rnormal(915, 250) in 1/`x'
  quietly replace var2=rnormal(1050, 325) in 1/`x'

  sum var1 , meanonly
  quietly replace pvar1=r(mean) in `x'
  quietly replace nvar1=`x' in `x'

  sum var2 , meanonly
  quietly replace pvar2=r(mean) in `x'
  quietly replace nvar2=`x' in `x'
  }
*
#delimit ;
twoway (line pvar1 nvar1) (line pvar2 nvar2)
 , ylabel(600(200)1600,angle(horizontal)) xlabel(1 50 100(200)500)
   ytitle(Mean)
   xtitle("Sample Size (log scale)",height(5))
   title("Statistical Regularity for Control and Treatment" , size(*0.9))
   xscale(log) graphregion(fcolor(white)) legend(off)
 ;
#delimit cr
```

The first part of the code defines the number of observations and creates a set of auxiliary variables. The second part includes a loop that creates each time two random variables with the means and standard deviations for labor income of treatment and control groups and saves the mean of for a set of replications (2 for the two first replications, 3 for the three firsts and so on). The last part just produce a graph with the results. The graph is shown below:

Figure 6: Statistical regularity for treatment and control groups



The results of the simulation suggest that, for small sample sizes, it is more difficult to distinguish any difference between the means of treatment and control group. As sample size increases, the distinction between both groups becomes more evident. This aspect is captured in the t-test formula by the parameter n related to sample size. Therefore, each time we run a t-test for difference in means, whether a difference in means is statistically significant depends on the signal (the distance between means), the noise (the difference in standard deviations) and the sample size of each group.

5.3. Sampling distribution of t-test under the null hypothesis

Each time we run a t-test a critical assumption is that the null hypothesis is true. In a t-test of difference in means the null hypothesis is that the difference is zero, or, in the context of impact evaluation, that there is no effect of the intervention. Under the null hypothesis, we construct a t-distribution consistent with no treatment effect. In the case of our example, this implies that the mean of incomes between those with and without higher education is the same. In this t-distribution consistent with no treatment effect, we impose a convention regarding the level of error for making an inference regarding the impact of an intervention we are willing to assume. This convention (usually 5% or 10%) implies to assume that extreme values of the t-statistic are more unlikely in the case of no effect, therefore inconsistent with the null hypothesis of no difference.

We can use STATA to illustrate this logic using simulations. We construct a sample of 100 observations and then create a variable x that represents income. We split the sample in two groups of equal size and create a normal variable with the same mean and variance. We assume this since if the null hypothesis is true, then mean and variance of treatment and control group are alike. We then compute a t-test of difference in income means and save the computed t-statistic. We replicate the experiment 10,000 times, saving the t-statistic after each replication. We know in advance that the average of the estimated t-statistics would be concentrated around 0. The STATA code for this experiment is given below:

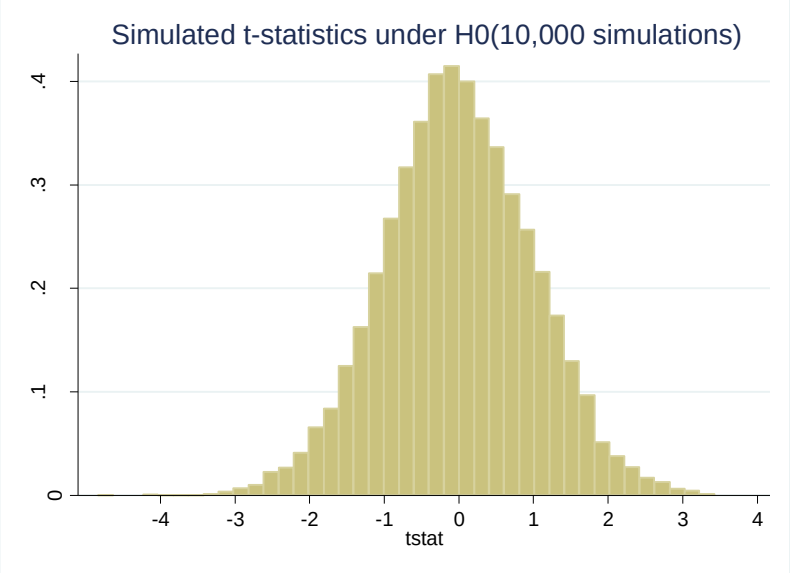
```
clear
set obs 10000
gen x=.
gen tstat=.
gen treatment = 1 in 1/50
replace treatment = 2 in 51/100
label define treatment 1 "treatment" 2 "control"
label values treatment treatment
set seed 123

forvalues i = 1/10000 {
  quietly replace x=rnormal(915, 250) in 1/50
  quietly replace x=rnormal(915, 250) in 51/100
  quietly ttest x, by(treatment) unequal
  quietly replace tstat=r(t) in `i'
  quietly list tstat in 1/`i'
}

sum tstat
```

As expected, the average of the t-statistic after 10,000 simulations is almost zero. We construct a histogram with the estimated t-test from this simulation. The STATA code is the following:

```
histogram tstat , bin(41) graphregion(fcolor(white)) xlabel(-4(1)4)
title("Simulated t-statistics under H0(10,000 simulations)")
```

The resulting graph is the following:

Figure 7: Simulation of t-test under the null hypothesis



We have derived empirically the t-distribution under the null hypothesis. This illustrates the logic of large sample approximations for statistical inference. Under the null hypothesis, we observe that most of estimated t-statistics are close to the mean but there are also some extreme values. The area below

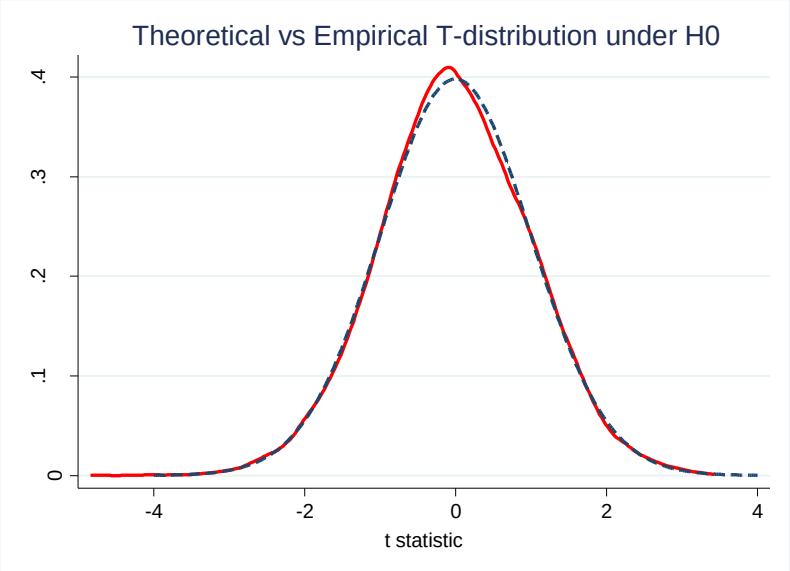
the curve is the probability mass. We know that the middle 95% of these t-values are located between -1.96 and 1.96. Therefore, the probability of obtaining a t-statistic outside the range (-1.96, 1.96) is really small (only 5%), so the convention is to consider this possibility as inconsistent with the null hypothesis. Although obtaining a t-statistic higher than the absolute value of 1.96 is still a possibility under the null hypothesis, we assume that this extreme value is unlikely to belong from the null distribution. Of course, we can be wrong and the observed t-statistics still belong to the null. For that reason, we say we can commit a type I error; that is, assuming that an effect exists when in fact it does not.

We can compare the results of our simulation with the theoretical t-distribution. The STATA code is below:

```
#delimit ;
twoway (kdensity tstat, lcolor(red) lwidth(*2))
       (function    y=tden(98,x),    range(-4    4)    lcolor(navy)    lwidth(*2)
lpattern(dash))
 , legend(off) ytitle("Density (Proportion)")
   xtitle("t          statistic",height(5))          graphregion(fcolor(white))
title("Theoretical vs Empirical T-distribution under H0")
 ;
#delimit cr
```

The resulting graph is the following:

Figure 8: Contrasting theoretical versus empirical t-distribution under the null hypothesis



As expected, the theoretical and the empirical distribution are very close each other.

5.4. Sampling distribution of t-test under the alternative hypothesis

We can also run a similar simulation for the case in which the alternative hypothesis is true. The steps are the same as below. The single change is to create different distributions of the outcome for treatment and control groups. We assume that the income distribution for treatment group has a mean

of 1,050 (standard deviation of 325) whereas the mean for the control group is 915 (250). The STATA code is below:

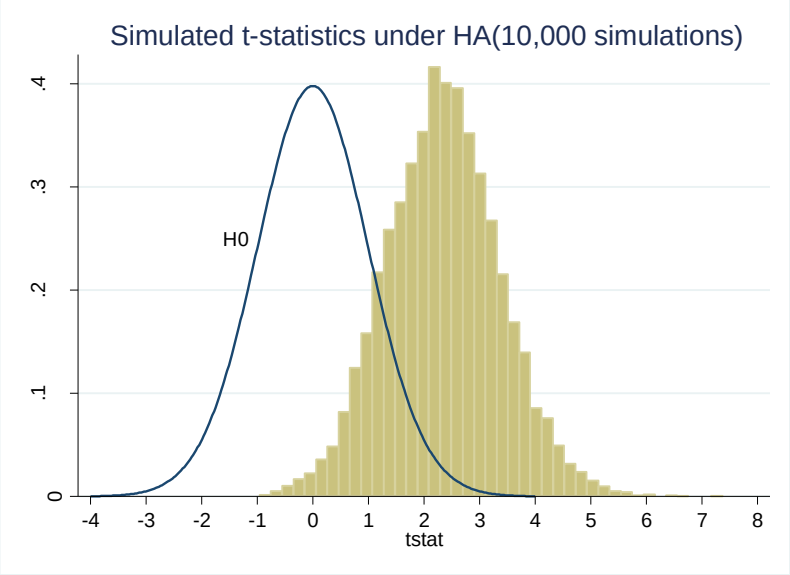
```
* SIMULATION

clear
set obs 10000
gen x=.
gen tstat=.
gen treatment = 1 in 1/50
replace treatment = 2 in 51/100
label define treatment 1 "treatment" 2 "control"
label values treatment treatment
set seed 123

forvalues i = 1/10000 {
  quietly replace x=rnormal(915, 250) in 51/100
  quietly replace x=rnormal(1050, 325) in 1/50
  quietly ttest x, by(treatment) unequal
  quietly replace tstat=r(t) in `i'
  quietly list tstat in 1/`i'
}

sum tstat
```

The average t-statistic is. We can also create a histogram with the estimated t-statistics. The routine is the following:

```
   histogram tstat , bin(41) graphregion(fcolor(white)) xlabel(-4(1)8)
title("Simulated t-statistics under HA(10,000 simulations)") ///

      addplot(function y=tden(98,x), range(-4 4) lcolor(navy) lwidth(*1.5))
legend(off) text(0.25 -1.4 "H0")
```

The resulting figure is the following:

Figure 9: Simulated t-test under the alternative hypothesis



We have added the theoretical distribution of the t-test under the null hypothesis for comparison. As we see, there is an area in which both distribution overlap. This implies that, for a group of t-values, is unclear whether a computed t-statistic belongs to the distribution of no effect (null hypothesis) or to the distribution consistent with a positive treatment effect. We say that, for a set of values, we have no enough power to detect a difference between treatment and control groups in terms of the impact of the intervention.

5.5. Testing difference in means

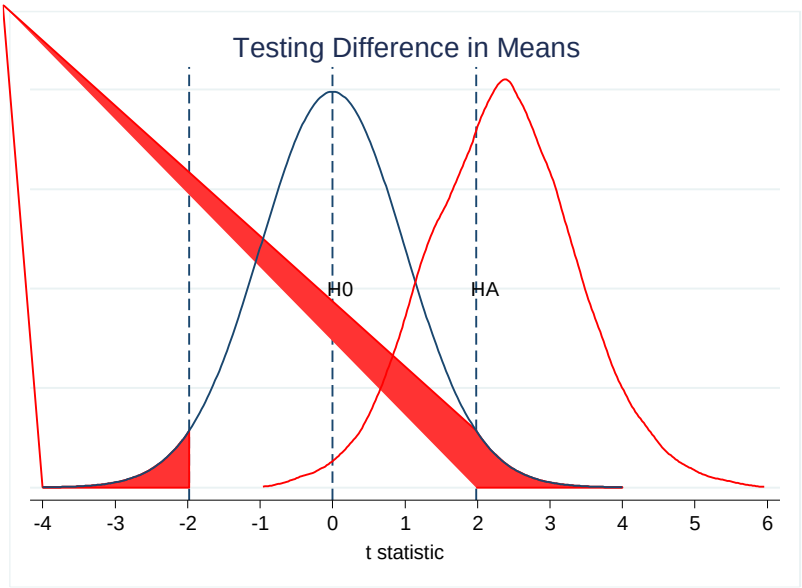
We are now in the position of understanding what a test of difference in means does. A graph comparing both distributions would be helpful. The STATA code is the following:

```
   #delimit ;
   graph twoway (function y=tden(98,x), range(-4 -1.98) bcolor(red)
                 recast(area) plotregion(style(none)))
              (function y=tden(98,x), range(1.98 4) bcolor(red)
                 recast(area) plotregion(style(none)))
                    (function y=tden(98,x), range(-4 4) clstyle(line)
                 clcolor(navy) clwidth(*1.5)
                 plotregion(style(none)))
                    (kdensity tstat if tstat<6, clstyle(line) clcolor(red)
clwidth(*1.5))
    ,yscale(off) legend(off)
     xlabel( -1.98 "t = -1.98"  0 "t = 0" 1.98 "t = 1.98")
     xtitle("t statistic", height(5))
     title(Testing Difference in Means)
     text(.2 2.1 "HA")
     text(.2 0.1 "H0")
     graphregion(fcolor(white))
     xline(-1.98 0 1.98, lcolor(navy) lpattern(dash))
```

```
    xlabel(-4(1)6)
  ;
#delimit cr
```

The resulting graph is the following:

Figure 10: Using t-test for difference in means



We basically have replicated the previous graph adding some extra features like the red shadow areas at the extreme areas of the null distribution. The red areas represent inconsistent values with the null distribution. So, if the estimated t-statistic is above 1.96, we say that the null hypothesis of no treatment effect is rejected at 5% confidence level (2.5% for each tail). Notice that an important area (about 40%) of the alternative distribution is below 1.96. That means that the level of power is low in this case, since for many values of the t-statistic it is not possible to distinguish to which distribution they belong. We will discuss the implications of this issue for a research design when we cover power analysis.

6. Randomization in Practice

Typically, allocating treatment between units requires performing some previous steps. Firstly, you need to compute power and sample size, a topic we are going to cover in a specific chapter later. Once the sample size is defined, you would need to decide how to choose this sample from your sampling frame, the universe of potential beneficiaries of your program. Under ideal conditions, you will be collecting baseline data you can use to evaluate whether your randomized assignment worked, although you can also rely on administrative data to check balance or data previously collected.

To show you how to randomize treatment, we use the ENCEL 2007 survey as if were the sample of our study. We will use this survey to check balance based on socio-economic characteristics and to evaluate which of the available randomization techniques minimizes variance and improves efficiency.

STATA does not have a command that randomly assign observations to treatment and control groups; therefore it is necessary to write a simple routine to perform such operation. Despite this, most of the available options are based on the command `uniform()` that generates random numbers in the bracket [0,1].

6.1. Simple Randomization

Consider the case of a simple randomization. We assign households to treatment and control status using the following steps:

- Step 1: Create a variable with random numbers from 0 to 1 and assign each household a number.
- Step 2: Sort households from low to high according to the value that was randomly assigned.
- Step 3: Create a variable that split the sample in treated and control units according to the value of the random variable created in the first step.

The following code considers the steps previously discussed:

```
* Create a random number
gen random = uniform()

* Sort observations according this random number
sort random

* Create a treatment variable (1=treatment/0=control)
gen treatment_simple = 0
replace treatment_simple = 1 if _n <= _N/2
```

The previous procedure ensures balance in pre-treatment characteristics between treatment and control households because the assignment process is completely independent of household's characteristics. Although this procedure is attractive for its simplicity, is little used in practice when the target population is heterogeneous.

To check pre-treatment balance, we can use a set of variables and compute descriptive statistics. The code is the following:

```
* Computing descriptive statistics by treatment status
tabstat    IncomeLab    famsize    Lanhead    sexhead    agehead    pov_HH,
by(treatment_simple) s(mean sd)
```

The results are presented below:

```
treatment_simple |  Income~b   famsize   Lanhead   sexhead   agehead    pov_HH
-----------------+------------------------------------------------------------
               0 |  2722.066  5.886892   .225416  .8585922  45.51167  .7937172
                 |  2968.581  2.898503  .4178746  .3484573  14.47656  .4046539
-----------------+------------------------------------------------------------
               1 |  2701.724  5.895809  .2224397  .8605748  45.25568  .7876146
                 |  2805.306  2.960603  .4159035  .3464051  14.21027  .4090144
```

```
-----------------+------------------------------------------------------------
           Total |  2711.894   5.891351   .2239277   .8595836    45.3837   .7906656
                 |  2888.044   2.929659   .4168835   .3474263   14.34432   .4068426
-------------------------------------------------------------------------------
```

We observe that all these variables seem to be similar between treatment and control units. For instance, the monthly labor income for the group assigned to treatment status is equal to 2,701 pesos whereas it is 2,722 for the case of control group. We can test formally whether these means in incomes are similar using a t-test, but this is something we will discuss later.

6.2. Randomized Block Designs

The simple randomized design described above is not the most efficient way to assign the treatment across participants. Whereas it is true that under this design treatment and control groups are, on average, alike, there is the chance that unbalance can be still present in some dimensions. One way to minimize such problem is using stratification (also known as blocking). In this method, units are assigned to treatment and control status within a strata or block, which are defined by a number of baseline variables.

In a stratified design, researchers aim to ensure that all strata are properly represented in the evaluation sample. It requires detailed knowledge about the population of interest in order to construct the blocks. In particular, it is suggested to stratify the sample with respect to variables that are strongly correlated with the outcome variable (Imai, King and Stuart 2008) and variables for which a subgroup analysis desired.

In the case of Oportunidades, we want consider **poverty** and **native language** as the variables for creating the blocks since we believe these variables are strongly related to the outcome variable (labor income). Each variable is binary, thus we can construct 4 strata in this simple case. Note however that the number of strata can exponentially increase with the number and type of variables (binary or categorical) included in the analysis[2].

Randomizing treatment within strata is slighted more complicated than the previous case. The steps are detailed below:

- Step 1: Create a variable with random numbers from 0 to 1 and assign each household a number.
- Step 2: Sort households from low to high according to the value that was randomly assigned to them within their strata.
- Step 3: Within each strata count the number of households and list each one.
- Step 4: Assign treatment status within each strata. That is, within each strata create a variable that assign treatment to a fraction of the households based on the ranking created in the previous step. The rest would be assigned to the control group.

[2] In particular if the variables are binary we can calculate the number of stratums as $2^N=2^2=4$. Note that inclusion of variables for which stratify, exponentially increases the number of stratums. For example, if we add 2 more variables, the number of stratums will be $2^N=2^4=16$. If these 2 additional variables were categorical (from 3 categories) the final number of stratums would be 2x2x3x3=36.

The routine is presented below:

```
* Create a random number
gen random = uniform()

* Sort the random number according to poverty and language
sort pov_HH Lanhead random

* Defining the size of each stratum
by pov_HH Lanhead: gen strata_size = _N

* Assigning a value for each household according to the order the household within each stratum
by pov_HH Lanhead: gen strata_index = _n

* Create a treatment variable (1=treatment/0=control)
gen treatment_block = 0
replace treatment_block = 1 if strata_index <= (strata_size/2)
```

In this case, there exists balance between treatment and control units within strata along with balance in the whole sample. To see that, let's compute some basic descriptive statistics using the same variables as before for the whole sample:

```
* Computing descriptive statistics by treatment status

tabstat IncomeLab famsize Lanhead sexhead agehead pov_HH, by(treatment_block) s(mean sd)
```

The results are presented below:

```
treatment_block |  Income~b   famsize   Lanhead   sexhead   agehead    pov_HH
----------------+------------------------------------------------------------
              0 |  2694.598  5.866002  .2239523  .8563039  45.58075  .7906397
                |  2664.067  2.930988   .416909  .3507969  14.36782  .4068701
----------------+------------------------------------------------------------
              1 |  2729.194  5.916704  .2239032  .8628638  45.18649  .7906915
                |  3095.919  2.928239  .4168765  .3440064  14.31868  .4068331
----------------+------------------------------------------------------------
          Total |  2711.894  5.891351  .2239277  .8595836   45.3837  .7906656
                |  2888.044  2.929659  .4168835  .3474263  14.34432  .4068426
------------------------------------------------------------------------------
```

Although results are not conclusive, it can be shown that the standard deviations are slightly smaller and more similar between groups than in the simple randomization case. This suggests that some gains in terms of efficiency are possible in this design compared to the simple randomized design.

It is also possible to assess the balance for each stratum. We can run the same analysis as before but for each stratum using the following code:

```
sort strata_size
```

```
by strata_size: tabstat IncomeLab famsize Lanhead sexhead agehead pov_HH,
by(treatment_block) s(mean sd)
```

We focus only in the first stratum. The results are below:

```
-> strata_size = 320

Summary statistics: mean, sd
  by categories of: treatment_block

treatment_block |  Income~b   famsize   Lanhead   sexhead   agehead    pov_HH
----------------+------------------------------------------------------------
              0 |  3793.925       5.4         1    .84375  46.49375         0
                |  2739.861  2.824422         0  .3642322  10.95617         0
----------------+------------------------------------------------------------
              1 |  3400.212   5.91875         1     .8375    47.575         0
                |  2344.594  2.933159         0  .3700671  11.99714         0
----------------+------------------------------------------------------------
          Total |  3597.069  5.659375         1   .840625  47.03437         0
                |  2553.522  2.886501         0  .3665987  11.48321         0
------------------------------------------------------------------------------
```

As you may notice, there is balance in terms of income, family size, household's head and age, although it is less efficient than the average case. Notice that, since language's head and household poverty are the variables used to stratify the sample, there is no balance in these dimensions within the stratum.

6.3. Cluster Randomized Designs

Randomization by cluster consists on randomly selecting a group of people that share certain characteristics or form a unit. This unit is called **cluster**. In the case of Oportunidades, these clusters are geographical units called villages (*villas*). PROGRESA, the original program that was called later Oportunidades, had a cluster randomized design in which 320 villages were assigned to the treatment group and 186 to the control one. The number of units selected in each cluster may vary depending on power calculations. In some cases, researchers include all the units from a cluster, but it is more common to select just a sample of them. For simplicity, we assume here that all households in a cluster are selected into the sample.

The procedure is a bit more complicated than previous ones, it is described below:

- Step 1: Create a variable that assigns a number for the place a household occupies in the list of households in the sampling frame.
- Step 2: Create a treatment variable that takes the value of 1 for a "representative household" in each cluster and zero otherwise.
- Step 3: Generate a variable with random numbers from 0 to 1 and assign a value to each household.
- Step 4: Sort the households by the treatment variable and the random allocation.
- Step 5: Keep the number of clusters.

- Step 6: From the list of "representative households", assign a fraction of them to the treatment and control status.
- Step 7: Assign the treatment status of the representative households to other households within the same cluster.
- Step 8: Sort households in the sampling frame according to Step 1.

The code is the following:

```
* Create an order variable
gen long order = _n

* Select one observation per cluster (village)
egen treatment_cluster = tag(villid)

* Create and sort random number
gen random = runiform()
sort treatment_cluster random

* Record the number of clusters
qui sum treatment_cluster
scalar nro=r(sum)

* Assign treatment status according to order of observation in each
cluster
replace treatment_cluster = _n > (_N - nro/2)

* Assign treatment to all households in a given village
bysort villid (treatment_cluster): replace treatment_cluster =
treatment_cluster[_N]

sort order
```

We compute some basic descriptive statistics as before. The results are presented below:

treatment_cluster	Income~b	famsize	Lanhead	sexhead	agehead	pov_HH
0	2744.205	5.830358	.2129366	.8570936	45.51994	.7910306
	2930.748	2.88018	.4094009	.3499925	14.32613	.40659
1	2677.463	5.956346	.2356401	.8622369	45.23857	.7902766
	2841.563	2.980239	.4244177	.344667	14.3629	.4071298
Total	2711.894	5.891351	.2239277	.8595836	45.3837	.7906656
	2888.044	2.929659	.4168835	.3474263	14.34432	.4068426

As you can see, the means between treatment and control units look alike, although an appropriate statistical test would be needed to evaluate this.

We consider a fourth alternative, the paired-matching design, in Appendix 3.1.

7. Evaluating Pre-treatment Balance

Once treatment status is assigned to households, regardless of the method used for the assignment, it is critical to assess balance in observable characteristics between both groups. The standard approach

is to compare averages between groups and determine whether they are statistically similar. We can compare the performance of the methods discussed above in terms of achieving balance. The basic command in STATA is `ttest`.

Recently, some scholars have suggested that testing difference in means between treatment and control groups could be misleading if differences in other moments besides the mean are present (Sekhon, 2012). In that scenario, it is better to study whether differences in terms of the distributions of the pre-treatment variables are present. The standard tool is the **Kolmogorov-Smirnov test for equality of distributions (KS-test)**. This test is implemented in STATA with the command `ksmirnov`.

The KS-test test computes a statistic that quantifies the distance of between the empirical distribution functions of the samples under comparison. The null hypothesis is that both sample distributions are drawn from the same distribution function.

7.1. Simple randomization

In the **simple randomization** case, we compute the difference in means of observable characteristics for all treated and control households in the sample. The code is the following:

```
*      Means
foreach covariates of varlist IncomeLab famsize Lanhead sexhead agehead
pov_HH {
      ttest `covariates', by(treatment_simple) unequal
      }
```

The results are below for the variable average family income:

```
Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |   11237    2722.066    28.00424    2968.581    2667.173     2776.96
       1 |   11239    2701.724    26.46162    2805.306    2649.855    2753.593
---------+--------------------------------------------------------------------
combined |   22476    2711.894    19.26391    2888.044    2674.136    2749.653
---------+--------------------------------------------------------------------
    diff |            20.34253    38.52862               -55.17627    95.86132
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   0.5280
Ho: diff = 0                     Satterthwaite's degrees of freedom =    22402

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.7012         Pr(|T| > |t|) = 0.5975          Pr(T > t) = 0.2988
```

The results show that the two means are very similar and we confidently reject the null hypothesis of equality of means.

One limitation of using the `ttest` command is that makes complicate comparisons when many control variables are available. In this scenario, it is better to have the possibility of present the results for many variables in a single table. We can do that using the user-written command `ttable2` (`findit ttable2`). The code is the following:

```
ttable2     IncomeLab     famsize     Lanhead     sexhead     agehead     pov_HH,
by(treatment_simple)
```

The results are presented in the table below:

```
------------------------------------------------------------------------------
Variables   G1(0)            Mean1     G2(1)          Mean2      MeanDiff
------------------------------------------------------------------------------
IncomeLab     11237        2722.066       11239     2701.724     20.343
famsize       11237           5.887       11239        5.896     -0.009
Lanhead       11237           0.225       11239        0.222      0.003
sexhead       11237           0.859       11239        0.861     -0.002
agehead       11230          45.512       11225       45.256      0.256
pov_HH        11237           0.794       11239        0.788      0.006
------------------------------------------------------------------------------
```

We found no evidence against imbalance between treatment and control units. The table reports the sample size for each group along with their means. The last column reports the difference in means. None of these differences in means are significant suggesting that the sample is balanced between treatment and control units.

We also perform the K-S test. The code is the following:

```
*     Distribution

foreach covariates of varlist IncomeLab famsize agehead {

      ksmirnov `covariates', by(treatment_simple)

      }
```

We consider only continuous variables since this test is appropriate only for this type of variables. Due to space constrains, we only report the result for the case of labor income below:

```
Two-sample Kolmogorov-Smirnov test for equality of distribution functions

 Smaller group       D       P-value  Corrected
 ------------------------------------------------
 0:                   0.0095    0.366
 1:                  -0.0062    0.648
 Combined K-S:        0.0095    0.696      0.690

Note: ties exist in combined dataset;
      there are 932 unique values out of 22476 observations.
```

We cannot reject the null hypothesis that the distribution of labor income for treatment and control groups come from the same distribution (p-value of 0.69). A similar result is found for the distribution of family size and household head age. This is strong evidence indicating that balance between treatment and control units was achieved by the randomization procedure.

7.2. Block-randomized designs

Consider the case of **block-randomization**. We can evaluate the balance between treatment and control units for the whole sample as well as balance for each stratum. Let's start by evaluating balance for the whole sample. The code is the following:

```
*      Means
foreach covariates of varlist IncomeLab famsize Lanhead sexhead agehead
pov_HH {
      ttest `covariates', by(treatment_block) unequal
      }
```

We report below the results for the case of labor income:

```
Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |   11239    2694.598    25.12935    2664.067     2645.34    2743.855
       1 |   11237    2729.194    29.20548    3095.919    2671.946    2786.442
---------+--------------------------------------------------------------------
combined |   22476    2711.894    19.26391    2888.044    2674.136    2749.653
---------+--------------------------------------------------------------------
    diff |            -34.5967    38.52849               -110.1153    40.92191
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -0.8980
Ho: diff = 0                     Satterthwaite's degrees of freedom =  21984.1

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.1846         Pr(|T| > |t|) = 0.3692          Pr(T > t) = 0.8154
```

We have no evidence to reject the null hypothesis of equal average labor incomes between treatment and control units for the whole sample. We also use the command `ttable2`. The code is the following:

```
ttable2 IncomeLab famsize Lanhead sexhead agehead pov_HH, by(treatment_block)
```

The results are shown below:

```
.    ttable2    IncomeLab    famsize    Lanhead    sexhead    agehead    pov_HH,
by(treatment_block)

------------------------------------------------------------------------------
Variables   G1(0)            Mean1     G2(1)          Mean2      MeanDiff
------------------------------------------------------------------------------
IncomeLab      11239        2694.598       11237     2729.194    -34.597
famsize        11239           5.866       11237        5.917     -0.051
Lanhead        11239           0.224       11237        0.224      0.000
sexhead        11239           0.856       11237        0.863     -0.007
agehead        11232          45.581       11223       45.186      0.394**
pov_HH         11239           0.791       11237        0.791     -0.000
------------------------------------------------------------------------------
```

We found no evidence against the null hypothesis in all variables but age of household head. Interestingly, the means for this variable for treatment and control groups are very close each other, still the difference is statistically significant at 5%.

We also compute the KS-test as above. The code is the following:

```
*     Distribution
foreach covariates of varlist IncomeLab famsize agehead {
      ksmirnov `covariates', by(treatment_block)
      }
```

The result for the case of labor income is reported below. We found no evidence against the null hypothesis of equal average incomes between treatment and control units:

```
Two-sample Kolmogorov-Smirnov test for equality of distribution functions

 Smaller group       D       P-value  Corrected
 ------------------------------------------------
 0:                   0.0103    0.306
 1:                  -0.0031    0.897
 Combined K-S:        0.0103    0.595      0.588

Note: ties exist in combined dataset;
      there are 932 unique values out of 22476 observations.
```

It is also possible to compare the balance on observables for the each stratum. The STATA code is presented below:

```
*BY STRATUM
foreach covariates of varlist IncomeLab famsize Lanhead sexhead agehead
pov_HH {
      foreach strata of numlist 320 4388 4713 13059 {
            ttest `covariates' if strata_size==`strata', by(treatment_block)
unequal
      }
}
```

We leave to the reader the analysis of this case but the key intuition is that, due to sample size issues, the difference between treatment and control units is estimated with less precision, therefore it is possible that in some cases lack of evidence against balance between treatment and control units can be due to power issues.

7.3. Cluster-randomized designs

A similar approach can be followed for the case of **cluster randomized** design. The code is the following:

```
*Means
foreach covariates of varlist IncomeLab famsize Lanhead sexhead agehead
pov_HH {
```

```
    ttest `covariates', by(treatment_cluster) unequal
    }
```

Due to space constraints, we focus again in the labor income case. The results are the following:

```
Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |   11119     2737.68    29.69555    3131.297    2679.471    2795.888
       1 |   11403    2689.122     24.6636    2633.698    2640.777    2737.467
---------+--------------------------------------------------------------------
combined |   22522    2713.095    19.25809    2890.125    2675.347    2750.842
---------+--------------------------------------------------------------------
    diff |             48.5575    38.60206               -27.10537    124.2204
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   1.2579
Ho: diff = 0                     Satterthwaite's degrees of freedom =  21685.3

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.8958         Pr(|T| > |t|) = 0.2084          Pr(T > t) = 0.1042
```

In this case, we cannot reject the null hypothesis of mean difference in average household income equal to zero. Using the command `ttable2`, we study whether this is true for the other variables. The code is the following:

```
ttable2 IncomeLab famsize Lanhead sexhead agehead pov_HH,
by(treatment_cluster)
```

The results are the following:

```
.  ttable2 IncomeLab famsize Lanhead sexhead agehead pov_HH,
by(treatment_cluster)

------------------------------------------------------------------------------
Variables   G1(0)            Mean1     G2(1)          Mean2      MeanDiff
------------------------------------------------------------------------------
IncomeLab     11119        2737.680       11403     2689.122     48.558
famsize       11119           5.870       11403        5.912     -0.041
Lanhead       11119           0.215       11403        0.233     -0.019***
sexhead       11119           0.866       11403        0.853      0.013***
agehead       11099          45.407       11380       45.356      0.051
pov_HH        11101           0.789       11375        0.792     -0.003
------------------------------------------------------------------------------
```

We find balance in means between treatment and control groups in all variables but language and sex of household head. We also test whether the distributions of these variables are alike between treatment and control units. The code is the following:

```
*Distribution

foreach covariates of varlist IncomeLab famsize agehead {
```

```
      ksmirnov `covariates', by(treatment_cluster)
      }
```

The results are presented below for the case of labor income:

```
Two-sample Kolmogorov-Smirnov test for equality of distribution functions

 Smaller group       D       P-value  Corrected
 ------------------------------------------------
 0:                   0.0086    0.431
 1:                  -0.0231    0.002
 Combined K-S:        0.0231    0.005      0.005

Note: ties exist in combined dataset;
      there are 932 unique values out of 22522 observations.
```

We reject the null hypothesis of equal distributions. This is also true for the case of family size. This is evidence that, at least in these two dimensions, we were not able to achieve balance between the distribution of treatment and control groups.

8. Evaluating the Impact of an Intervention using Experimental Data

We have shown how to assign treatment status and assess pre-treatment balance. Now we are in the position of evaluating the impact of the intervention. We will start with the basics in this section. In the next section we will discuss some common practices followed by applied researchers.

In this section, we use the original experimental sample of PROGRESA. We don't use the ENCEL 2007 survey here because its design is no longer experimental. In the previous section, we were interested in the evaluation design rather than the estimation of the impact. For this reason, it was fine to use the ENCEL 2007 as it was administrative data used as baseline.

We have assembled a new dataset, called `PanelPROGRESA_97_99.dta`. This dataset is a repeated cross-section of different waves of the ENCEL survey for March and October 1998, and November 1999. It also includes the baseline survey called ENCASEH 1997. This survey was used to compute the mean tested algorithm designed to define poor households in treated villages.

We first study the simple case in which a single cross-section after the implementation of the program is available to the researcher. We choose the wave of November 1999, more than a year after the start of distribution of cash to beneficiaries' households. We will discuss later how to take advantage of a baseline and several follow-ups to evaluate the impact of the program.

It is important to recall that PROGRESA has a two level assignment to treatment. In the first place, communities are selected into treatment and control status given a set of socio-economic characteristics. Then, within the treated communities, a fraction of the households are selected to receive the program. This implies that there are households in beneficiary villages that do not get access to the program. Consequently, we consider two treatment variables. The variable `D` is the treatment at village/community level and the variable `D_HH` defines whether a household is a PROGRESA's beneficiary.

The variable of interest is the family income per-capita (`Income_HH_per` in the sample). We restrict the sample to households. The code to keep only households in the sample is the following:

```
keep if HH==1            /*Sample of HH*/
```

We proceed in this way to simplify the analysis since the variables to be used in the examples are also constructed at household level.

8.1. Using hypothesis testing

The simplest way to evaluate the impact of the program is using a t-test between treatment and control units. We restrict the sample to observations from 1999. The code is the following:

```
ttest Income_HH_per if year==1999, by(D_HH) unequal
```

The results are presented below:

```
. ttest Income_HH_per if year==1999, by(D_HH) unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |   13608    444.2986    12.46543    1454.134    419.8647    468.7326
       1 |    7594     386.893     33.0812    2882.814    322.0447    451.7413
---------+--------------------------------------------------------------------
combined |   21202    423.7374    14.29779    2081.888    395.7127    451.7622
---------+--------------------------------------------------------------------
    diff |            57.40562    35.35184               -11.89129    126.7025
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   1.6238
Ho: diff = 0                     Satterthwaite's degrees of freedom =  9792.15

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9478         Pr(|T| > |t|) = 0.1044          Pr(T > t) = 0.0522
```

Although we find that average income is higher for the control group (444 over 387), the difference (57.4) is marginally insignificant (p-value of 0.104). We cannot reject the null hypothesis that both groups have the same incomes. To obtain further evidence in this regard, we perform a KS-test for equality of income distribution between treatment and control groups. The routine is the following:

```
ksmirnov Income_HH_per if year==1999, by(D_HH)
```

The results are presented below:

```
. ksmirnov Income_HH_per if year==1999, by(D_HH)

Two-sample Kolmogorov-Smirnov test for equality of distribution functions

 Smaller group       D       P-value  Corrected
 ------------------------------------------------
 0:                   0.0003    0.999
 1:                  -0.1469    0.000
```

```
 Combined K-S:       0.1469    0.000      0.000

Note: ties exist in combined dataset;
      there are 5226 unique values out of 21202 observations.
```

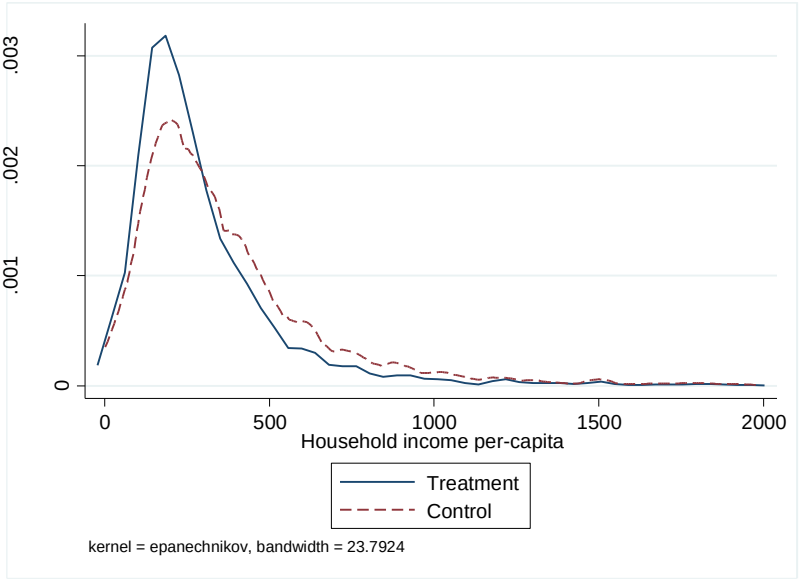
We reject the null hypothesis of equality of income distribution between treatment and control households. Therefore, although we cannot reject the hypothesis of the equality of these two means, we do reject the null hypothesis of equality of distributions. This result clearly speaks about the limitation of simple test about difference in means to evaluate the impact of an intervention.

We can illustrate graphically this point using kernel densities. The STATA code is the following:

```
kdensity Income_HH_per if D_HH==1 & year==1999 &
Income_HH_per<2000,addplot(kdensity Income_HH_per if D_HH==0 & year==1999 &
Income_HH_per<2000, lpattern(dash)) ///
      graphregion(fcolor(white)) legend(label(1 "Treatment") label (2
"Control")) xtitle("Household income per-capita") title("")
```

The resulting graph is the following:

Figure 11: Kernel density for distribution of income between treatment and control households



To facilitate the graphical analysis, we restrict the analysis to households with incomes lower than 2,000 pesos. Although both distributions follow a similar pattern, we do see differences around the mean.

8.2. Regression adjustment

The most common approach to analyze experimental data in impact evaluation is regression. In this context, regression estimates a CEF, but since the assignment to treatment is experimental, this estimated CEF has a causal interpretation. We run the following regression:

(21) $HH\_Income_{iv} = a + b(T\_HH_{iv}) + \varepsilon_{iv}$;

where is the HH_Income_{iv} per-capita household income for household i in village v and T_HH_{iv} is an indicator variable equal to one for treated households in village v. The STATA code is the following:

```
regress Income_HH_per D_HH if year==1999, vce(cluster villid)
```

We allow standard errors to be correlated among households from the same village by using the option `vce(cluster)`. The results are presented below:

```
. regress Income_HH_per D_HH if year==1999, vce(cluster villid)

Linear regression                                      Number of obs =   21202
                                                       F(  1,   499) =    2.34
                                                       Prob > F      =  0.1271
                                                       R-squared     =  0.0002
                                                       Root MSE      =  2081.8

                              (Std. Err. adjusted for 500 clusters in villid)
------------------------------------------------------------------------------
             |               Robust
Income_HH_~r |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        D_HH |  -57.40562   37.56596    -1.53   0.127    -131.2126    16.40133
       _cons |   444.2986   15.90243    27.94   0.000     413.0547    475.5426
------------------------------------------------------------------------------
```

We found that the coefficient associated to the treatment variable is -57.4, which implies that treated households have lower incomes (on average 57.4 pesos less than control households), although the lack of statistical significance suggest that PROGRESA has no impact of household incomes.

Recall the previous discussion about the connection between regression and potential outcomes. It was shown that the coefficient associated to the treatment variable recovers the difference between potential outcomes; i.e. the causal effect of interest. To see that, notice that this coefficient (57.4) is the same as the difference computed between treatment and control units in the t-test in section 8.1.

9. A More General Analysis of Experimental Designs

We have covered the basics of analyzing experimental data to evaluate the impact of an intervention. In this section we consider the most common extensions to such analysis.

9.1. Using control variables

The most basic extension to the simple analysis is the previous section is to include control variables in the basic regression (21). A first conceptual issue is to define what a **control variable** is and in which sense this variable differs from the **treatment variable**. Recall that we impose the following restriction for a treatment variable:

(22) $E(u_i / D_i) = 0$.

This assumption has several names such as **exogeneity** or **conditional mean-zero**. In a randomized experiment, the random assignment to treatment status guarantees such condition. The question is whether a similar requirement is needed for a control variable. Let's call W to such variable. We usually use control variables to avoid a biased estimate of the treatment effect due to a lack of perfect randomization or to reduce the variance of the causal estimates. Notice that we are not interested in the causal effect of the control variable W; we only use it with goal of improving the causal interpretation of D. In this scenario, the original assumption of conditional mean-zero for the causal interpretation of D is replaced with the following condition:

(23) $E(u_i/D_i, W_{1i}, ..., W_{ri}) = \gamma_0 + \gamma_1 W_{1i} + ... + \gamma_r W_{ri}.$

This assumption is known as the **conditional mean independence assumption** (Stock and Watson 2007, chapter 13). Notice that, under this assumption, the control variables are allowed to be correlated with the unobservable factors. We require that the conditional expectation of the unobservables –given D and the control variables W - cannot depend on D but it can depend on W. Consider the case of PROGRESA. Whether a household head speaks a native language might be correlated with access to economic opportunities like education or inadequate access to economic assets which has a direct effect on incomes and that might be correlated with the access to PROGRESA. Due to the fact that it is hard to measure some of these dimensions (for instance, access to economic assets is a proxy of wealth, a dimension that is hard to measure empirically), these factors enter the error term. Although randomization ensures balance in expectation, it is possible that balance cannot be achieved in all dimensions and we can fail to detect imbalance in some unobservable dimensions since we can only test balance with the covariates available in our sample. Therefore, using control variables help to minimize bias induced by imperfect randomization. It also helps to improve efficiency since the inclusion of control variables reduces the variance of the error term.

To illustrate the point, consider the following regression:

(24) $HH\_Income_{iv} = a + b(T\_HH_{iv}) + W_{iv}'\delta + \varepsilon_{iv};$

where W is a set of control variables. We consider as control variables the family size, the language, sex and age of household head. The STATA code is the following:

```
regress Income_HH_per D_HH famsize langhead sexhead agehead if year==1999,
vce(cluster villid)
```

The results are the following:

```
. regress Income_HH_per D_HH famsize langhead sexhead agehead if year==1999,
vce(cluster villid)

Linear regression                                   Number of obs =   19870
                                                    F(  5,   499) =   24.42
                                                    Prob > F      =  0.0000
                                                    R-squared     =  0.0029
                                                    Root MSE      =  2139.5

                              (Std. Err. adjusted for 500 clusters in villid)
```

```
------------------------------------------------------------------------------
             |               Robust
Income_HH_~r |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        D_HH |  -17.34478   36.93734    -0.47   0.639    -89.91666     55.2271
     famsize |  -26.55784   6.433828    -4.13   0.000    -39.19857   -13.91711
    langhead |  -131.2895   32.69871    -4.02   0.000    -195.5336   -67.04539
     sexhead |   -25.7107   51.27232    -0.50   0.616    -126.4469    75.02553
     agehead |   2.871198   .5577957     5.15   0.000      1.77528    3.967115
       _cons |   502.0264   71.06058     7.06   0.000     362.4116    641.6412
------------------------------------------------------------------------------
```

We observe a change in the magnitude of the estimated coefficient (from -57.4 to -17.34), although still statistically insignificant. As expected, the standard error associated to the treatment variable is also lower. We also observe that most of the coefficients of the control variables are significant, but we don't interpret these coefficients since they lack of causal interpretation as stated above. This approach differs from basic econometric books which implicitly assume that all variables are treatments since they require the conditional mean-zero assumption for all explanatory variables. In this context, we reserve the causal interpretation only for the treatment variable since this is the only one that has been randomly assigned.

9.2. Using fixed effects to control for heterogeneity

In some cases, it is important to control for relevant sources of heterogeneity using fixed effects. This is particular important for programs of national coverage in which the treatment may differ by region. One way to control for this heterogeneity is using region fixed effects. In this case, we will restrict the comparison between treatment and control households within the same region. Since households from the same region are exposed to the same type of shocks, controlling for region fixed effect helps to control for the heterogeneity of treatment responses.

In our basic example, we use an econometric specification that allowed the comparison between a treated household from the south Mexico, traditionally poorer than the average Mexican region, against a control household from the north of Mexico, usually less poor. Institutional and economic differences across the regions in which these households are located can explain differences in terms of treatment responses. For instance, if households from the north have access to better schools, then the impact of the transfer will be higher than a household from the south, in which schools are of lower quality. Therefore, it makes sense to restrict the comparison between treatment and control households within the same region.

The new econometric specification is the following:

(25) $HH\_Income_{ivj} = a + b(T\_HH_{ivj}) + W_{ivj}^{'}\delta + \gamma_j + \varepsilon_{ivj}$;

where γ_j is a set of dummies for each region (except one to avoid the dummy trap) in the sample. To add fixed effects to our design in STATA, there are two options. The first one is to create the dummies for each region and add them to the specification. A second option is to use the command `areg`. This command allows the use of a large set of dummies. In this case, we have only 7 dummies, so

we still can accommodate the dummies using the command `regress`. However, when the number of dummies is too large, it is better to use `areg`.

We use the second alternative. We need to create a variable with the regions considered in the sample. In the case of Mexico, these regions are called **entidad**. We can do that from the village id. We can also create similar variables for other levels of government. The code is below:

```
gen entidad=substr(villid,1,2)
gen munici =substr(villid,3,3)
gen locali =substr(villid,6,3)
destring entidad munici locali, replace
```

The STATA code for the regression is the following:

```
areg Income_HH_per D_HH famsize langhead sexhead agehead if year==1999, absorb(entidad) vce(cluster villid)
```

The results are presented below:

```
. areg Income_HH_per D_HH famsize langhead sexhead agehead if year==1999, absorb(entidad) vce(cluster villid)

Linear regression, absorbing indicators           Number of obs   =      19870
                                                  F(   5,    499) =      12.32
                                                  Prob > F        =     0.0000
                                                  R-squared       =     0.0044
                                                  Adj R-squared   =     0.0039
                                                  Root MSE        =  2138.1298

                              (Std. Err. adjusted for 500 clusters in villid)
------------------------------------------------------------------------------
             |               Robust
Income_HH_~r |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        D_HH |  -19.55199   34.78941    -0.56   0.574    -87.90377     48.7998
     famsize |  -26.79339   6.133213    -4.37   0.000     -38.8435   -14.74329
    langhead |  -57.51842    24.4117    -2.36   0.019    -105.4808   -9.556037
     sexhead |  -30.17082   50.61528    -0.60   0.551    -129.6162    69.27451
     agehead |   2.620776   .5469326     4.79   0.000     1.546202    3.695351
       _cons |   495.0883   71.44922     6.93   0.000     354.7099    635.4666
-------------+----------------------------------------------------------------
     entidad |   absorbed                                     (7 categories)
```

The estimated coefficient is now -19 pesos, slightly higher than the original -57.4 pesos estimated in the simple case from the previous section. This coefficient is still no significant, but the important drop in the estimated impact of the program on incomes is an indication of the existence of bias induced by the presence of heterogeneity in treatment response. Once we allow the comparison between treatment and control households to be restricted only to observations from the same region, we control for potential sources of heterogeneity. As a consequence, the estimated coefficients are closer to zero.

This highlights the importance of understanding what the counterfactual is in a given estimation. When we use region fixed effects, we are exploiting variation between treatment and control households in a given region. In this case, the counterfactual is estimated using control households in the same region. In the original case, the counterfactual was estimated using control households in all regions. Since control households in different regions may differ in many unobservable dimensions, this does seem to be a good idea.

9.3. Using baseline data

So far, we have been paying attention to cross-sectional estimates of the impact of an intervention. Many programs usually collect baseline data that can be exploited to provide more efficient estimates of the impact. One advantage of using baseline data is that it can be used for controlling for pre-program differences in the outcome variables.

Baseline data can be used in two ways. The covariates collected in the baseline survey can be used as control variables. These variables have the advantage of being pre-determined and therefore they cannot be affected by treatment itself, a problem that can be an issue with contemporary controls. The second way to exploit baseline data is exploiting its panel structure. This leads to a *difference in difference model* with experimental data.

We focus in the second case. The econometric specification for a single baseline and follow-up is the following:

(26) $HH\_Income_{ivt} = a + b(T\_HH_{iv}) + c(R_t) + d(T\_HH_{iv} * R_t) + W_{ivt}^{'}\delta + \varepsilon_{ivt};$

where R_t is an indicator variable of the round of the ENCEL survey and the variable $T\_HH_{iv} * R_t$ is an interaction between the indicator variable for the survey round and the treatment variable. The coefficient associated to the interaction recovers the causal effect of interest.

To implement this in STATA, we need to create some additional variables. In the first place, we need to use the original assignment to treatment rather than the actual treatment. The reason is simple: in 1997, there were no treated units since the program only started in 1998. Therefore, we need to create the assignment. The STATA code is the following:

```
gen aux=.
replace aux=0 if D_HH==0 & year==1998
replace aux=1 if D_HH==1 & year==1998
egen D_asigHH=max(aux),by(hogid)
replace D_asigHH=D_HH if year==1999
drop aux
label var D_asigHH "HH level assignment to treatment"
```

We basically assign the treatment status in 1998 to 1997. We keep the original treatment status from the survey for 1999 since represents the actual treatment status. Then, we have the assignment to treatment in 1997, a year in which no actual treatment was provided.

We also need to drop one year from the sample. We choose to drop 1998 since we want to evaluate the impact of the program after more of one year. We then create a dummy variable for the second

period. Finally, we create the interaction between the assignment to treatment status and the dummy equal to one for the second period. The STATA code is the following:

```
drop if year==1998

tab year, gen(dyear)

gen I_dhXdyear2=D_asigHH*dyear2
label var I_dhXdyear2 "Interaction: HH treatment and second period dummy"
```

Now, we are in the position of estimating the difference in difference model. The STATA code is the following:

```
regress Income_HH_per D_asigHH dyear2 I_dhXdyear2, vce(cluster villid)
```

The results are the following:

```
. regress Income_HH_per D_asigHH dyear2 I_dhXdyear2, vce(cluster villid)

Linear regression                                       Number of obs =   43038
                                                        F(  3,   505) =    3.61
                                                        Prob > F      =  0.0133
                                                        R-squared     =  0.0002
                                                        Root MSE      =  3722.3

                              (Std. Err. adjusted for 506 clusters in villid)
------------------------------------------------------------------------------
             |               Robust
Income_HH_~r |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    D_asigHH |  -64.31267   91.31702    -0.70   0.482    -243.7207    115.0954
      dyear2 |   81.59896   28.91951     2.82   0.005     24.78159    138.4163
 I_dhXdyear2 |    6.90705   97.92011     0.07   0.944    -185.4739     199.288
       _cons |   362.6997   26.07715    13.91   0.000     311.4666    413.9327
------------------------------------------------------------------------------
```

The coefficient of interest is 64.31 but it is not significant. The coefficient is now positive. Notice that we did not include covariates in this case since the ones we were using before are time invariant like household head sex or language, but they can be easily incorporated in the previous specification as long as they vary over time.

9.4. Heterogeneous effects

It is common to study how the impact of the program varies according to the characteristics of program participants. For instance, we might be interested in understanding whether the program has a different impact according to the gender of household head. Some theories suggest that gender matters in terms of intra-household allocation of resources.

To explore heterogeneous effects there are several alternatives. One approach consists in using interactions between treatment status and gender. This is the approach we will explore here. Another

alternative is to test for difference in sub-groups using techniques for adjusting for multiple comparisons.

To illustrate this, we return to the cross-sectional case for simplicity. The econometric specification is the following:

(27) $HH\_Income_{ivj} = a + b(T\_HH_{ivj}) + c(S_{ivj}) + d(T\_HH_{ivj} * S_{ivj}) + \gamma_j + \varepsilon_{ivj}$;

where S_{ivj} is a dummy variable equal to one for male household heads and $T\_HH_{ivj} * S_{ivj}$ is the interaction between treatment status and household head gender.

To implement this regression in STATA, we need to create the interaction between treatment status and the household gender. The code is the following:

```
gen I_dhXsexhead=D_HH*sexhead
label var I_dhXsexhead "Interaction: HH treatment and sex of household head"
```

The regression in STATA is the following:

```
areg Income_HH_per D_HH sexhead I_dhXsexhead if year==1999, absorb(entidad)
vce(cluster villid)
```

The results are reported below:

```
. areg Income_HH_per D_HH sexhead I_dhXsexhead if year==1999, absorb(entidad)
vce(cluster villid)

Linear regression, absorbing indicators           Number of obs   =      21202
                                                  F(   3,    499) =       1.59
                                                  Prob > F        =     0.1913
                                                  R-squared       =     0.0029
                                                  Adj R-squared   =     0.0024
                                                  Root MSE        =  2079.3520

                              (Std. Err. adjusted for 500 clusters in villid)
------------------------------------------------------------------------------
             |               Robust
Income_HH_~r |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        D_HH |  -15.16252   108.2794    -0.14   0.889    -227.9022    197.5772
     sexhead |  -78.70923   54.58844    -1.44   0.150    -185.9607    28.54229
I_dhXsexhead |  -37.89679   114.3473    -0.33   0.740    -262.5583    186.7647
       _cons |   510.5489   53.24143     9.59   0.000     405.9439    615.1539
-------------+----------------------------------------------------------------
     entidad |   absorbed                                     (7 categories)
```

The coefficient associated to the interaction recovers the differential impact of treatment for male household heads compared to females ones. We find that the impact of the program is lower for households with male heads, although the coefficient is not statistically significant.

10. Final Remarks

In this chapter we have introduced the basics of experimental designs for impact evaluation. In the first part of this chapter, we have offered an extended discussion about potential outcomes. This discussion was designed to be complementary material for those students interested in a more detailed exposition of such conceptual framework but it is no required.

The second part of this chapter covered issues of implementation. We started discussing hypothesis testing in the context of mean comparisons, a natural approach to evaluate pre-treatment balance and the impact of an intervention without imposing strong parametric assumptions like in the case of linear regression. As mentioned above, this approach is less sensitive to the Freedman critique.

We then show different ways to allocate treatment across units and study pre-treatment balance between treated and control groups using t-tests and Kolmogorov-Smirnov tests of equality of distributions. Using the experimental sample of PROGRESA, we then discuss how to use experimental data to estimate the impact of a program or intervention. We use hypothesis testing and linear regression showing that, in some cases, non-parametric distributional tests like the Kolmogorov-Smirnov are more informative than regular t-tests and regression to analyze experimental data. We also cover some basic extensions of regression adjustment of experimental data.

We covered the basic of experimental designs. In the next chapter, we would be paying attention to extensions for cases in which one or more assumptions are not meet.

11. Further Readings

Angrist, Joshua and Jorn-Steffen Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Baum, Cristopher (2006). *An Introduction to Modern Econometrics using STATA*. STATA Press.

Cameron, A. Colin and Pravin Trivedi (2009). *Microeconometrics using STATA*. STATA Press.

Chen, Bryant and Judea Pearl (2013). "Regression and causation: A critical examination of six econometric textbooks." Mimeo, UCLA.

Haavelmo, Trygve (1944). "The Probability Approach in Econometrics," *Econometrica*, 12, 1-115.

Holland, P. W. (1986). "Statistics and Causal Inference", *Journal of the American Statistical Association*, 81, 945-970.

Gerber, Alan S., and Donald P. Green (2012). *Field Experiments: Design, Analysis, and Interpretation*. W.W. Norton.

Goldberger, Arthur (1998). *Introductory Econometrics.* Harvard University Press.

Goldberger, Arthur (1991). *A Course in Econometrics.* Harvard University Press.

Pearl, Judea (2009). *Causality: Models, Reasoning, and Inference*. Second Edition. Cambridge University Press.

Rubin, Donald B. (1974). "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66(5), 688-701.

Stoddard, Gregory (2012). *Biostatistics and Epidemiology Using Stata: A Course Manual.* Mimeo, University of Utah School of Medicine.

Stock, James y Mark Watson (2007). *Introduction to Econometrics*. Pearson/Addison Wesley.

Wooldridge, Jeffrey (2009). *Introductory Econometrics: A Modern Approach*. Cengage Learning.