



THE WORLD BANK



Session VII

Sampling and Power

Christel Vermeersch

January 2008

Introduction

- ❑ Objective: evaluate programs and interventions
(¿Does it have an effect or not?)
- ❑ We need enough data to be able to detect changes that are due to the program
 - Part 1: discuss power
- ❑ We need representative data:
 - Part 2: discuss sampling

PART 1: Having “enough” data

- ❑ Evaluating a program is a significance test
- ❑ Example: y is a variable denoting test results. We try to estimate the effect of doubling the school’s budget, on the students’ test results.

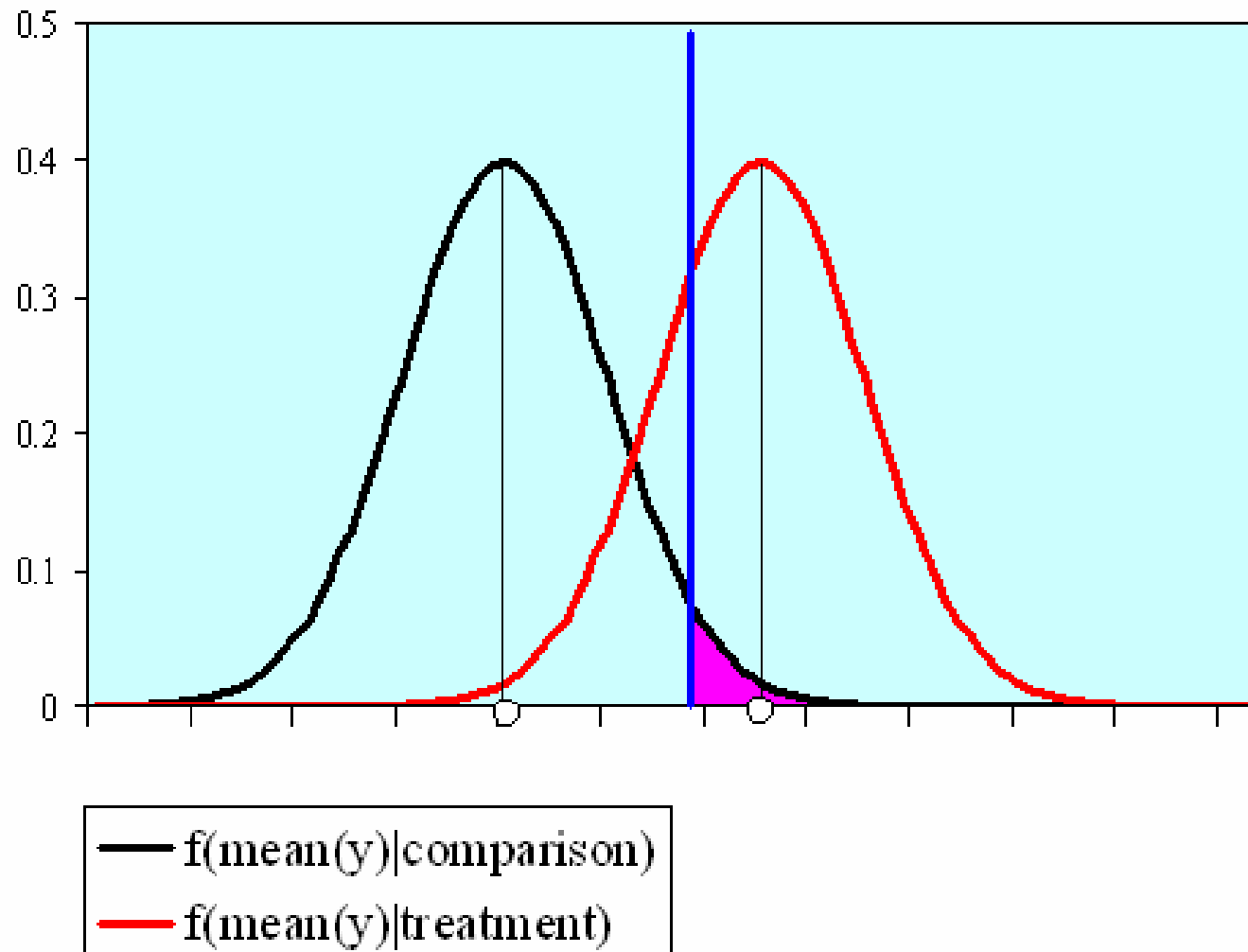
$H_0: E(y|\text{Treatment}) = E(y|\text{Control})$ ie the program has no effect

$H_1: E(y|\text{Treatment}) > E(y|\text{Control})$ ie the program has a positive effect

		Intervention has an effect (in reality)	
		No	Yes
Statistical test	We reject H_0	Type I error	OK
	We don't reject H_0	OK	Type II error

Type I error

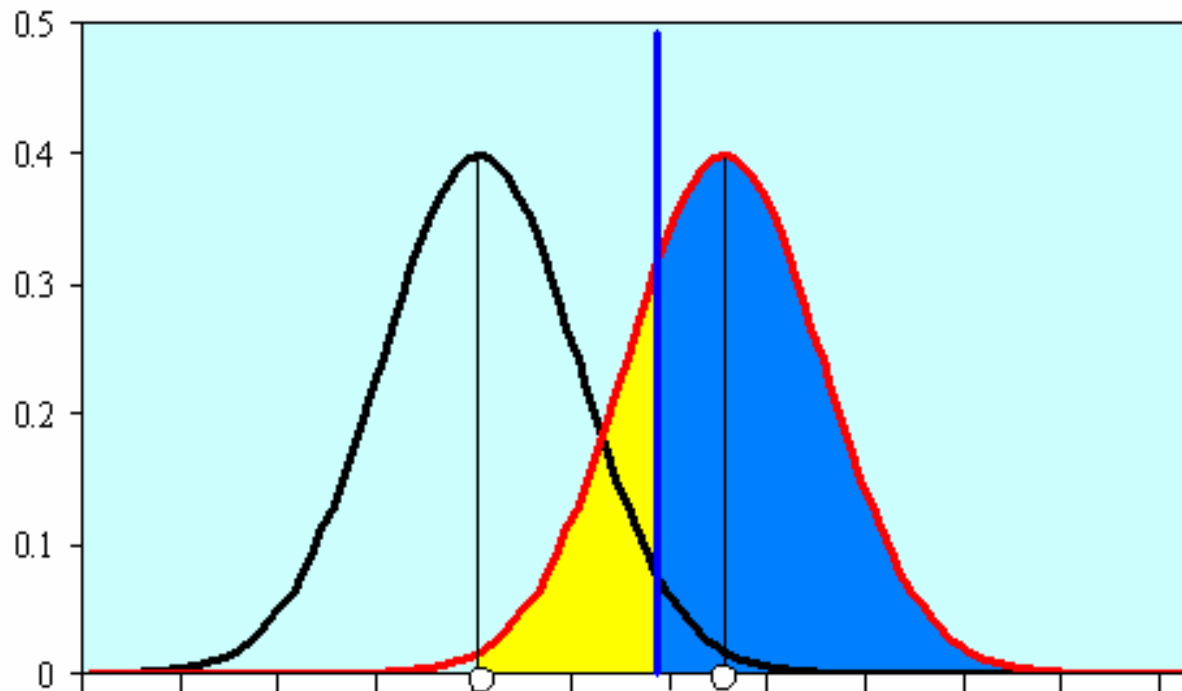
Type I error



Type II error and power

Type II error

Power



— $f(\text{mean}(y)|\text{comparison})$
 — $f(\text{mean}(y)|\text{treatment})$

Type I and Type II errors

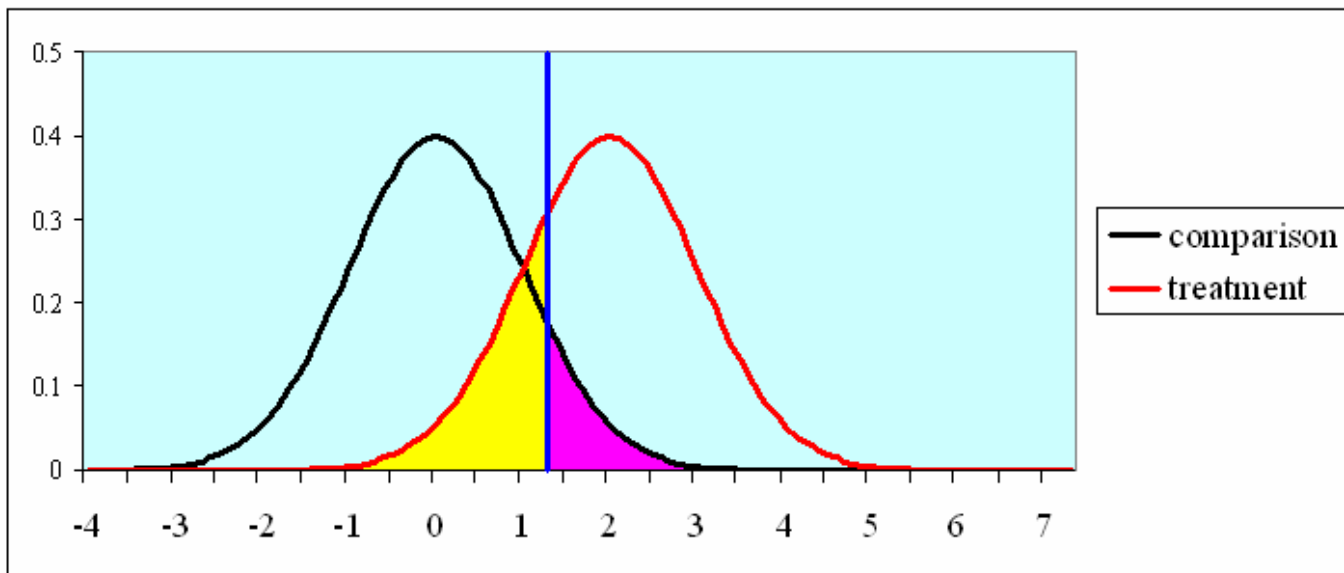
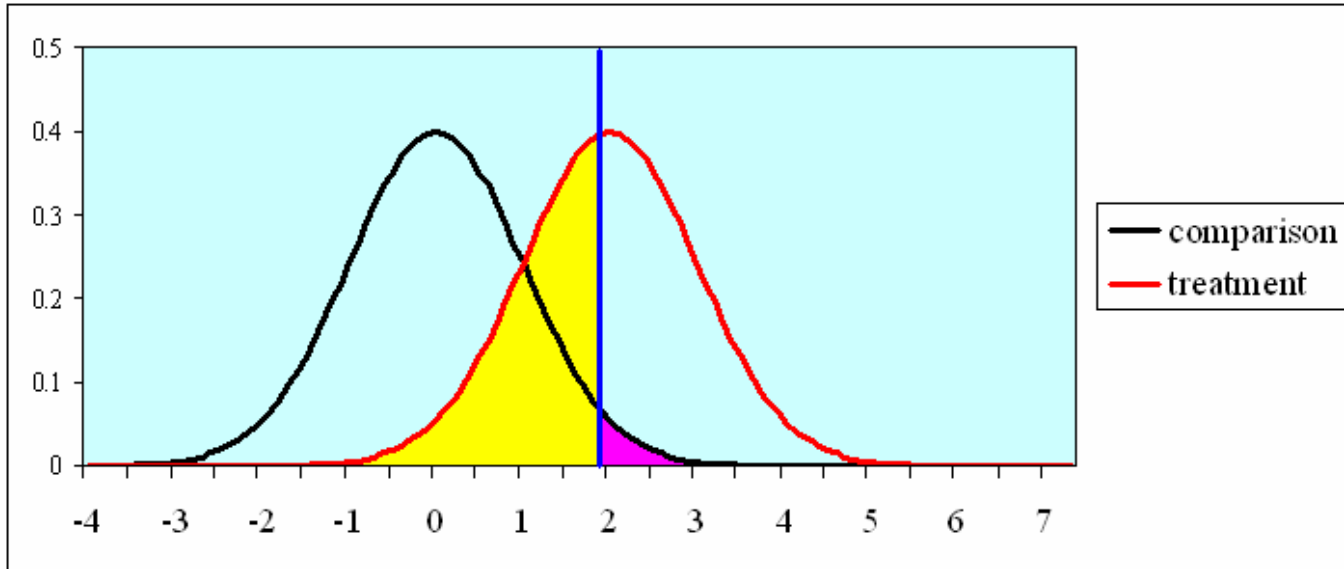
- Type I error
 - = rejecting the null hypothesis when it is true
 - **Significance level α = probability that we will conclude that the intervention has an effect, when in reality it has no effect**
 - Typical values for α : 0.01, 0.05, 0.1
- Type II error
 - = failing to reject the null when the null is false
- Power= $1 - \beta$
 - Probability that we will reject the null hypothesis when the null hypothesis is false
 - = probability that we will reject the null hypothesis when the alternative hypothesis is true
 - **=probability that we will conclude that the intervention has an effect, when it does really have an effect**
- Goal: maximizing power, for a given level of α

What affects the power of a test?

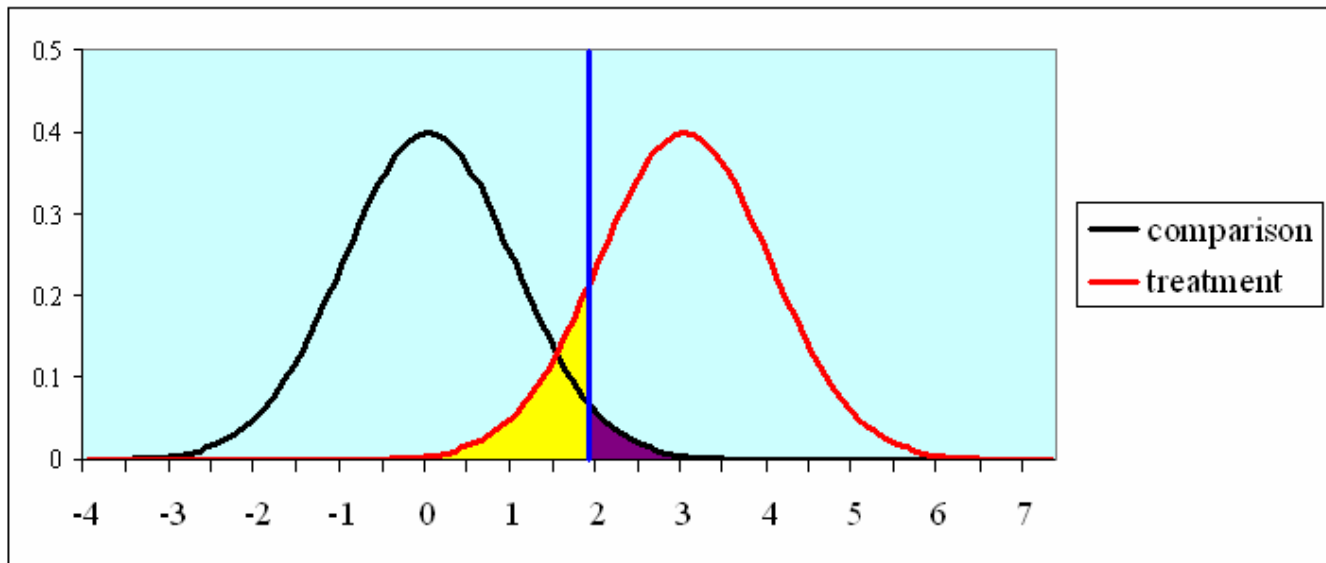
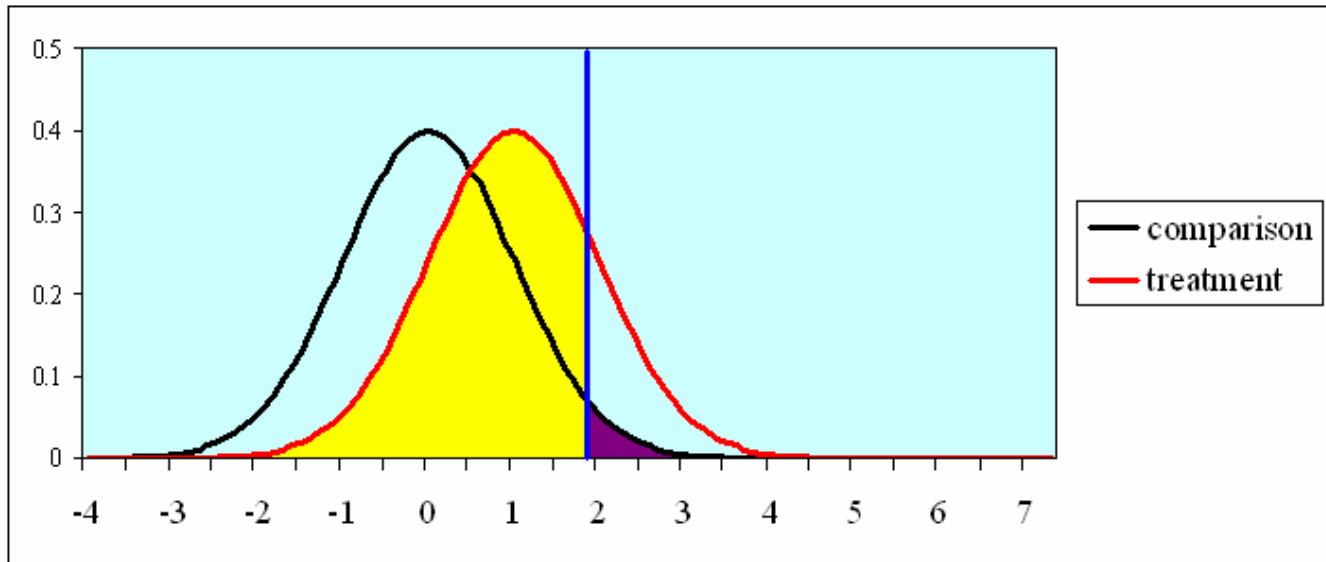
The power of a test increases when...

- ❑ α increases (we are compromising on the type I error)
 - ❑ When the expected effect is larger
 - ❑ When the variance of the indicator increases... how does this happen???
 - Remember that the variance of an average is proportional to $1/n$, where n is the sample size
 - So: increasing the sample size will decrease the variance of an average (many indicators are averages!)
- ⇒ To increase the power of a sample, when the size of the effect is fixed and the type I error is fixed, we'll have to increase the sample size.
- ⇒ In practice, we calculate the sample size that is necessary to detect a change X in the indicator, with a Type I error of 5% and a power of 80 or 90 %.

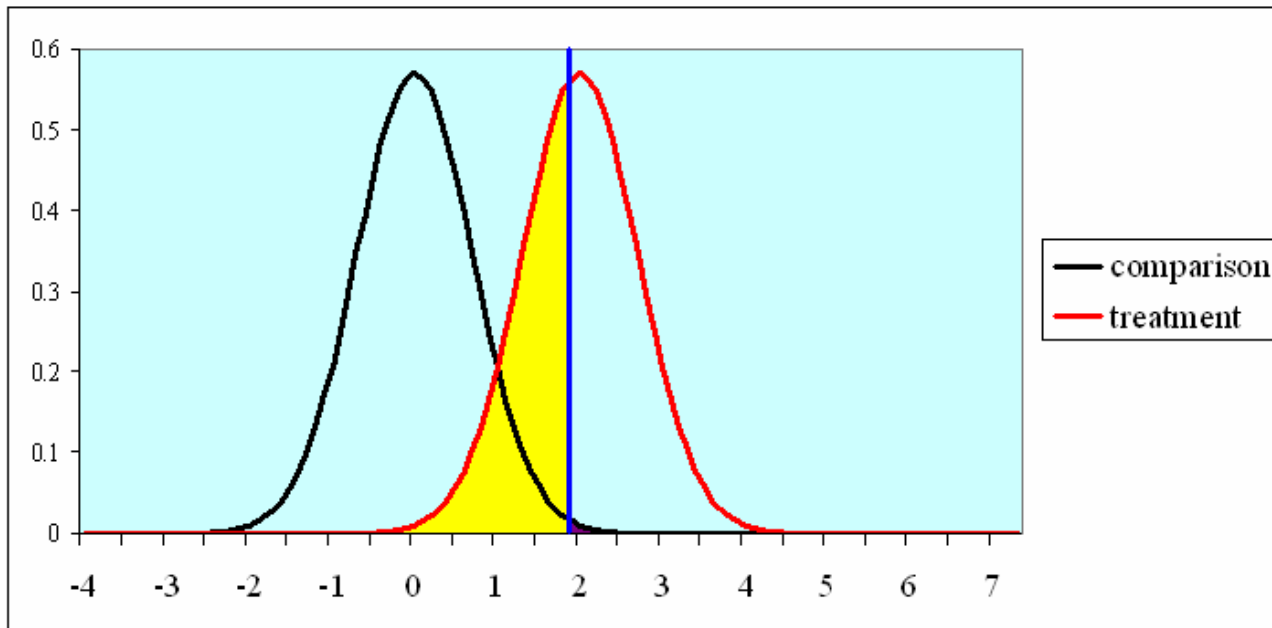
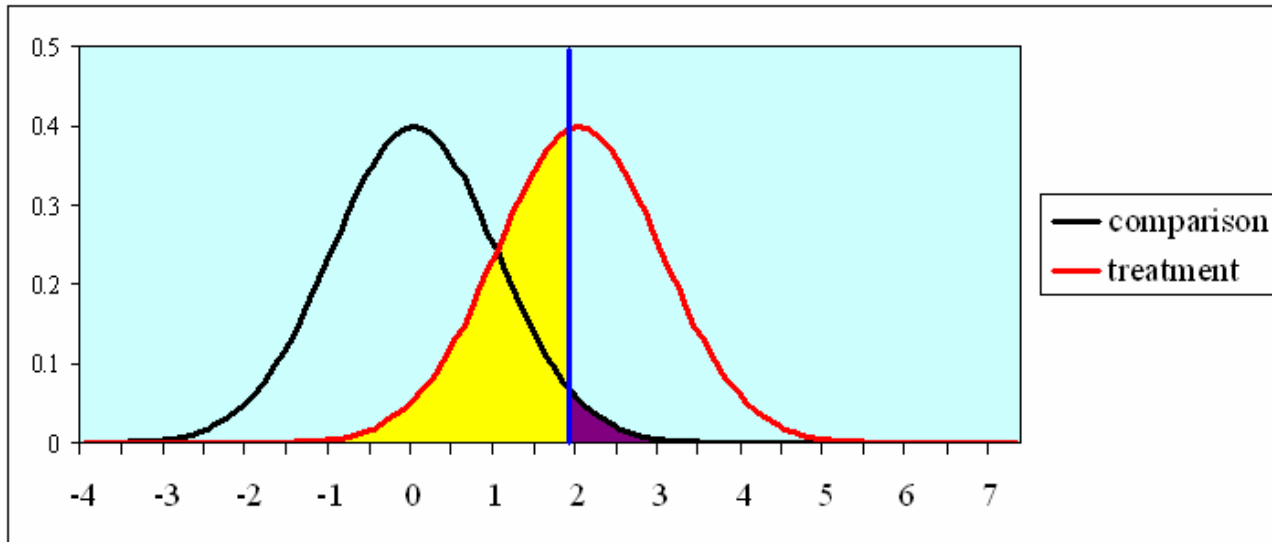
What happens when we increase the Type I error?



What happens when we increase the expected size of the effect?



What happens when we increase sample size?



Home exercise: some simulations in Stata

***computing a sample size;**

```
sampsi 130 135, alpha (0.05) power(0.8) sd1(15)
sd2(18) onesided;
```

***Let's increase the expected size of the effect==> a smaller sample will be enough!;**

```
sampsi 130 145, alpha(0.05) power(0.8) sd1(15)
sd2(18) onesided;
```

***Let's increase the required power ==> we need a larger sample;**

```
sampsi 130 135, alpha (0.05) power(0.9) sd1(15)
sd2(18) onesided;
```

Home exercise: More simulations in Stata...

***let's compute the power of a test;**

```
sampsi 130 135, alpha (0.05) sd1(15) sd2(18)
      n1(100) n2(100) onesided;
```

***Let's increase the expected size of the effect
=> power goes up;**

```
sampsi 130 145, alpha (0.05) sd1(15) sd2(18)
      n1(100) n2(100) onesided;
```

***Let's increase the sample size => poder
aumenta;**

```
sampsi 130 135, alpha (0.05) sd1(15) sd2(18)
      n1(200) n2(200) onesided;
```

Parte 2: Sampling (→ Representative data)

□ Representative surveys

- Goal: learning about an entire population
 - Ex. LSMS/ national household survey
- Sample: representative of the national population

□ Impact evaluation

- Goal: measuring changes in key indicators for the target population of an intervention
- In practice: measuring the difference in indicators between treatment and control groups
- We sample strategically in order to have a representative sample in the treatment and control groups
- Which is not necessarily the same as a representative sample of the national population

Can we use a general household survey for impact evaluation?



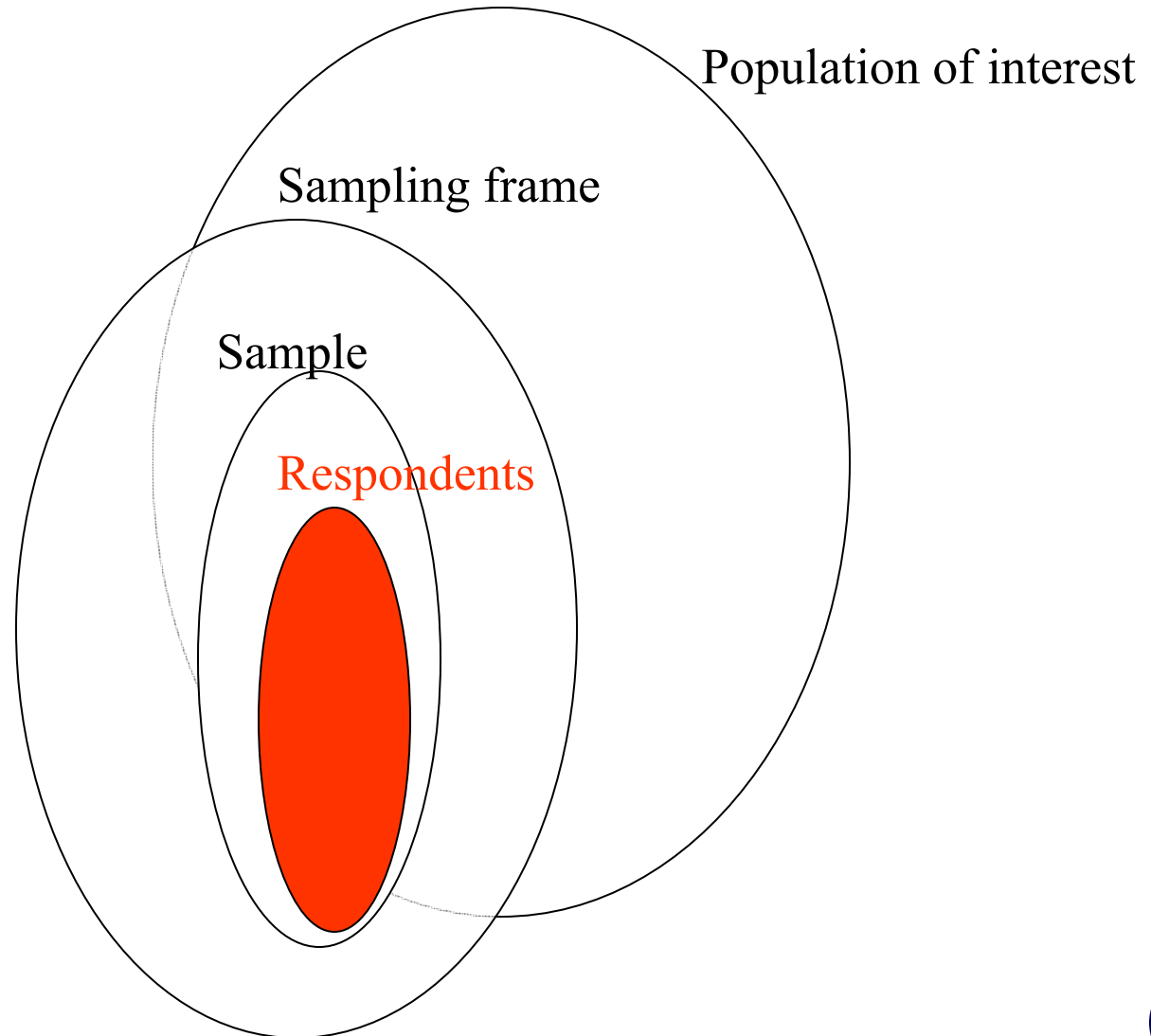
- Only if...
 - The general household survey is representative of the treatment and comparison groups
 - There are enough observations in both the treatment and control groups.

- Example: We evaluate the effect of an information campaign for mothers on child feeding under 1 year, 10 control districts, 10 treatment districts. Can we use the DHS as a baseline?
 - Many of the women in the DHS will not have a child under 1 year old.
 - The DHS is not representative at the district level.

Definitions

- ❑ **Unit of analysis:** The type of entity for which we want to gather information (persons, villages, schools,...)
- ❑ **Population:** The set of units of analysis for whom we want to know whether the intervention works or not
- ❑ **Sampling frame:** The physical list of units of analysis from which we will draw the sample (?What's the best case scenario?)
- ❑ **Sample:** The units of analysis that we draw from the sampling frame, and about which we want to collect information
- ❑ **Respondents:** The units of analysis for whom we obtain data (eg. Persons who answer the questions of the survey)
- ❑ **Response rate:** percentage of units of analysis in the sample, for whom we collect information (eg respond to the survey)

Graphically...



The response rate ...

- ❑ Best case scenario: it's 100 %
- ❑ When it is low, there is a risk of selection bias, for example when the control group does not want to answer the survey.
“non-response bias”
- ❑ Rules of thumb:
 - don't believe the results if the response rate is under 70 percent for either the control or the treatment group
 - There should be no significant difference in response rates between the treatment and control groups
- ❑ ?? All well and good.... But what do we do when the response rate is low ???

When the response rate is low ...

- ❑ How about: “ increasing the sample size in advance, because we know some people won’t respond”
 - ❑ It won’t work because any non-response bias will also be present in a larger sample
- ❑ Better solutions:
 - ❑ Additional efforts to collect information on the non-respondents (eg 3 home visits instead of one)
 - ❑ Sub-sampling the non-respondents to check whether they are different from the respondents.
 - ❑ In a follow-up survey: Checking the baseline characteristics of respondents and non-respondents to check whether they are similar.
- ❑ Always: check that the consultants report on the response rate in their evaluation report.

Sampling methods

□ Random sampling

- Simple: each unit in the sampling frame has the same probability of being selected into the sample
- Stratified: we first divide the sampling frame into strata (groups), and then we do a simple random sample in each strata
- clustered: we sample clusters of units: eg. Villages with all the persons that live there
- Multi-stage – a combination of two or more random sampling methods from the above, for example: stratified random sample of villages, then simple random sample of persons within the sampled villages

□ Systematic sampling

- Ex: Case studies: careful: we cannot generalize the results of these samples

The Design Effect with clustering/ multi-stage sampling

- ❑ When samples are clustered, this affects the power of the significance test.
- ❑ Example: let's estimate the height of 7 year old school children in Egypt
 - **Case 1**: survey 200 children, randomly chosen among all 7 year old Egyptian school children
 - **Case 2**: first randomly choose 20 schools among all Egyptian schools, then randomly choose 10 children in each of those 10 schools

Which sample gives us the most information??

Clustering changes the variance of the indicator

- Persons within a cluster tend to be more similar to each other than to persons in another cluster:
 - 20 children in the same school tend to come from more similar backgrounds than 20 children from different schools
- Therefore: the standard error on the mean height of 200 children in 20 schools (stratified random sampling) is larger than if we selected 200 children randomly (simple random sampling)

$$\text{Var}_{\text{Clustered}} = \text{Var}_{\text{SRS}} * \underbrace{(1 + \rho * (k - 1))}_{\text{design effect}}$$

- ρ is a measure of homogeneity between the units (children) belong to the same cluster (school)
- k =number of units (children) sampled in each cluster (school)

What does this mean for the sample size

- ❑ Larger variance \rightarrow lower power for a given sample size and type I error (α)!
- ❑ The loss of power due to the design effect can be compensate by increasing the overall sample size
- ❑ By how much do we need to increase the sample size when we have clustering?

Determining the sample when there is clustering

- Step 1: use standard methods to determine the sample assuming simple random sampling. (see first part of the presentation)
- Step 2: use the following formula to adjust the sample size for clustering

$$\begin{aligned}n_{Cluster} &= n_{SRS} * \text{design effect} \\ &= n_{SRS} * (1 + \rho * (k - 1))\end{aligned}$$

- You will need to estimate ρ from existing surveys...
- Use Stata!

When do we *really* worry sample size?

- When
 - We have very small samples *at unit of treatment!*
 - Suppose treatment in 20 schools and control in 20 schools
 - But there are 400 children in every school
 - This is *still* a small sample
- When
 - Our evaluation has a cross-over design
 - Sample size requirements increase exponentially
- When
 - We use Regression Discontinuity as our evaluation strategy

Conclusions

- ❑ Unfortunately, in many cases we can NOT use general surveys like LSMS and DHS as a baseline survey.
- ❑ The sampling frame and the sample selection method determine the representativity of the sample for the population
- ❑ When we want to measure very small differences in indicators between the treatment and control groups, we need a very large sample
- ❑ When there are clusters in the sample, we need a larger sample!