AGRODEP

Household survey data course

**Dakar, 8-10 October 2012**

# Sampling for Impact Evaluation




# Introduction

- This is a basic introduction to sampling for impact evaluation.
- Focuses mainly on sample size calculations for randomized cluster samples but basic ideas are transferrable to more complex randomized designs and non-random sample designs.
- Main Question: how do we construct a sample to credibly detect a given effect size within our evaluation budget constraints?

# Impact Evaluation

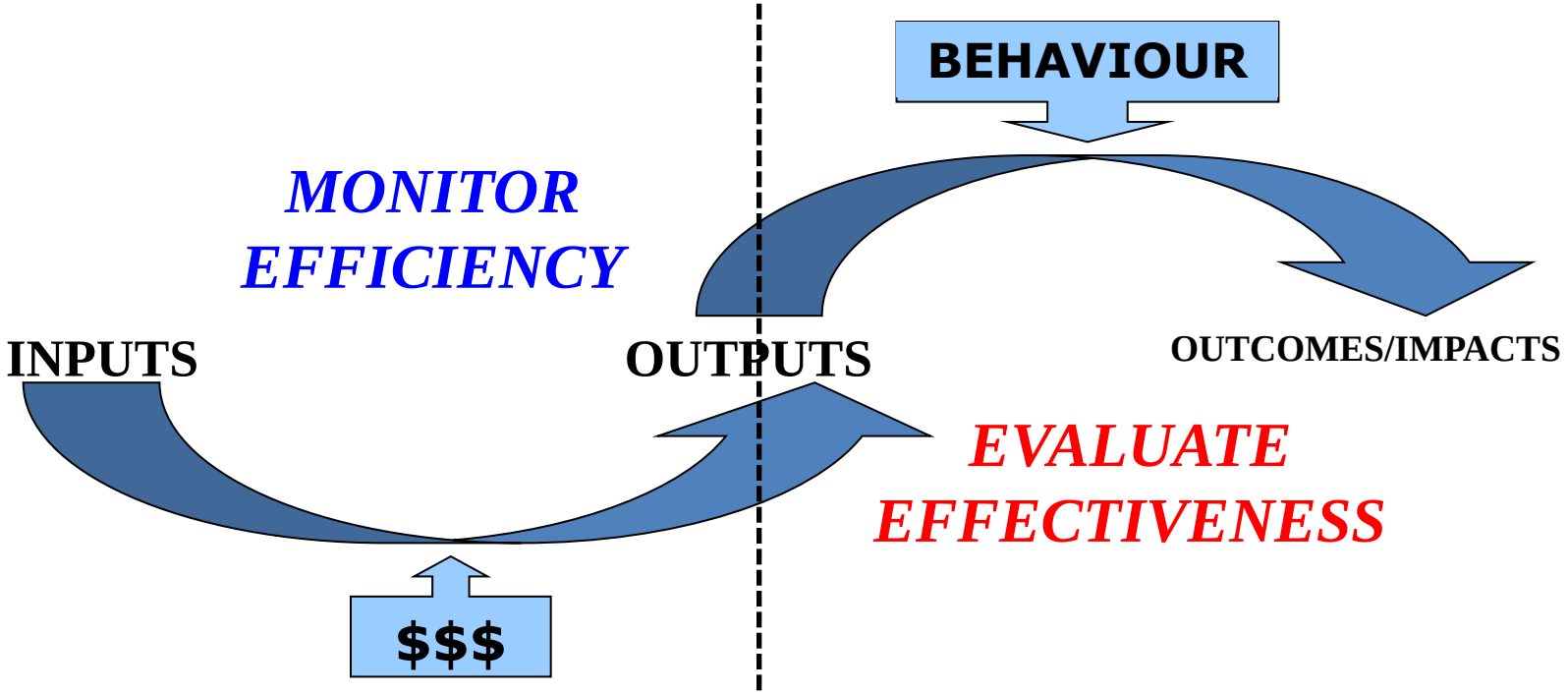
- Determine the causal effect of the project on outcomes (not only on outputs):
  - Farmers' wellbeing?
  - Land productivity?
  - Input supply, labor productivity, environment, women's conditions, health and nutrition,…?
  - …all of the above plus-> for whom? For which development domain? For which type of households? For which livelihood?
- What would be the impact with a different intervention?

# Monitoring

# Evaluation

'Traditional' M&E: monitoring to track implementation efficiency (input - output)

Impact Evaluation: estimate causal effectiveness on outcomes (output - outcome)

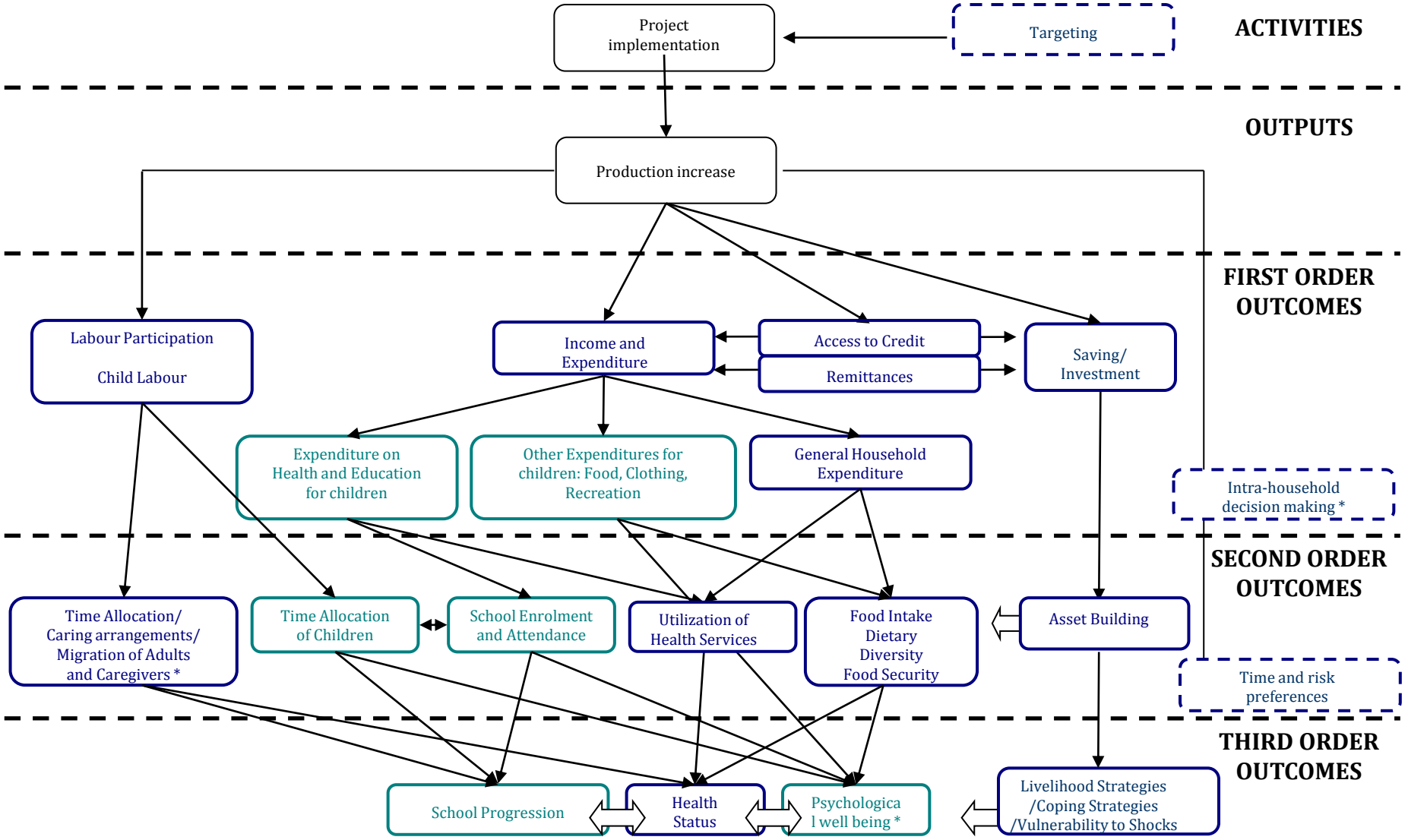



Note: Diagram from WorldBank training material produced by Arianna Legovini, Lead Economist - AIEI

# Theory of Change

- Impact evaluation must be based on a set of hypotheses on the change that can be achieved as a consequence of the intervention

- How would you think the project can affect the life of the beneficiaries?

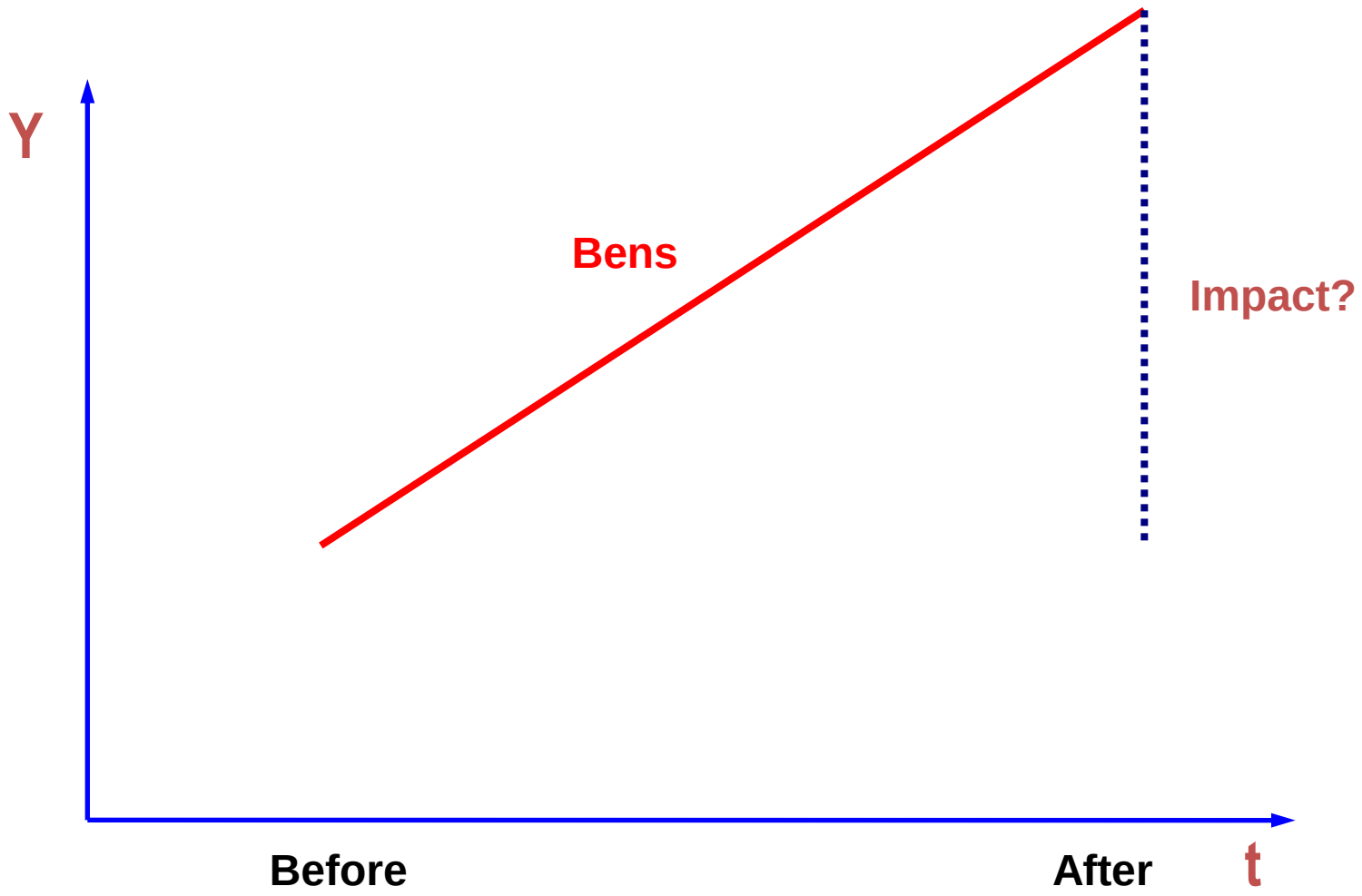

Project implementation
Targeting
ACTIVITIES
OUTPUTS
Production increase
FIRST ORDER OUTCOMES
Labour Participation
Child Labour
Income and Expenditure
Access to Credit
Remittances
Saving/ Investment
Expenditure on Health and Education for children
Other Expenditures for children: Food, Clothing, Recreation
General Household Expenditure
Intra-household decision making *
SECOND ORDER OUTCOMES
Time Allocation/ Caring arrangements/ Migration of Adults and Caregivers *
Time Allocation of Children
School Enrolment and Attendance
Utilization of Health Services
Food Intake Dietary Diversity Food Security
Asset Building
Time and risk preferences
THIRD ORDER OUTCOMES
School Progression
Health Status
Psychological well being *
Livelihood Strategies /Coping Strategies /Vulnerability to Shocks

# Impact Evaluation

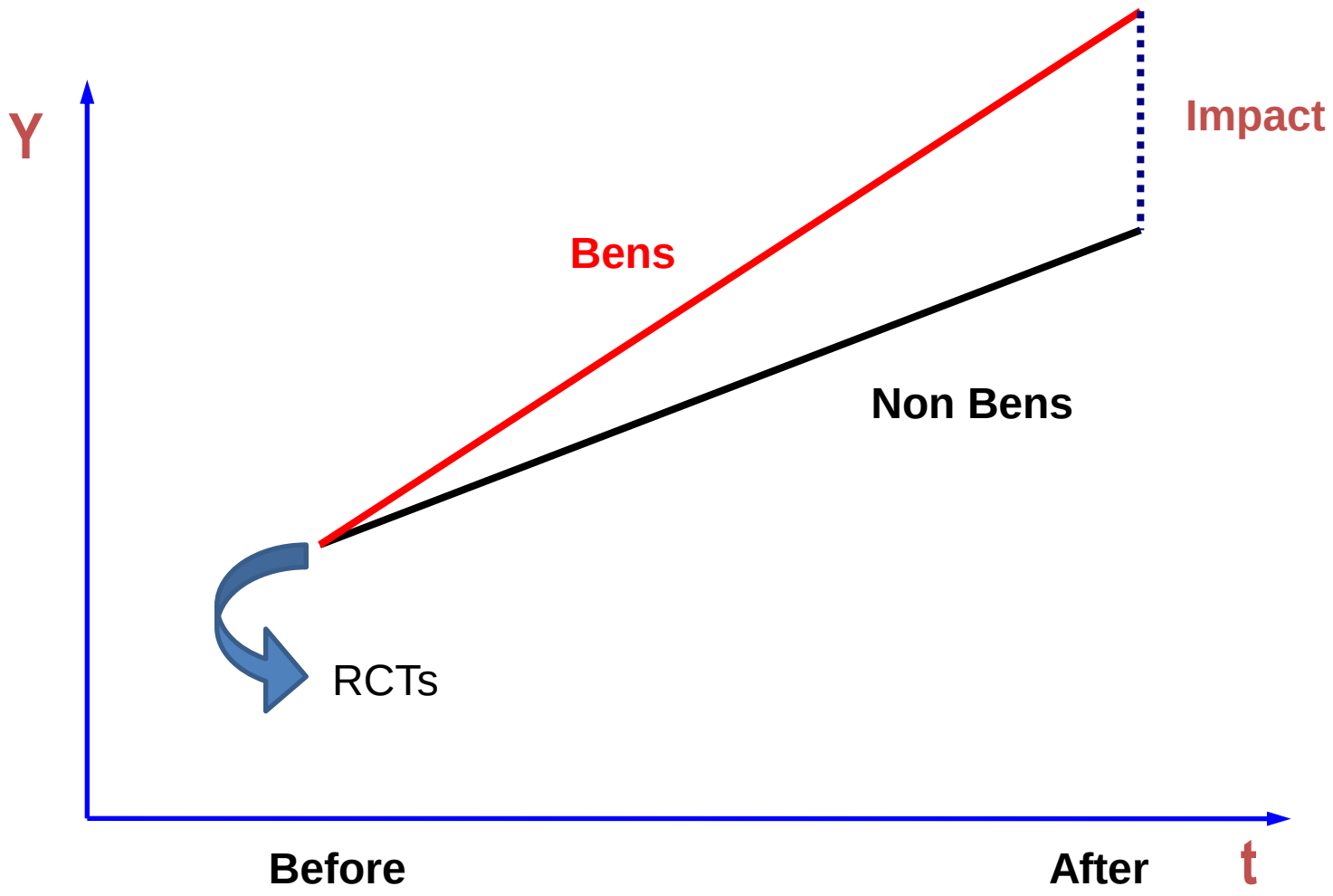
- How would you go about measuring the causal impact of AR on …

  -productivity?

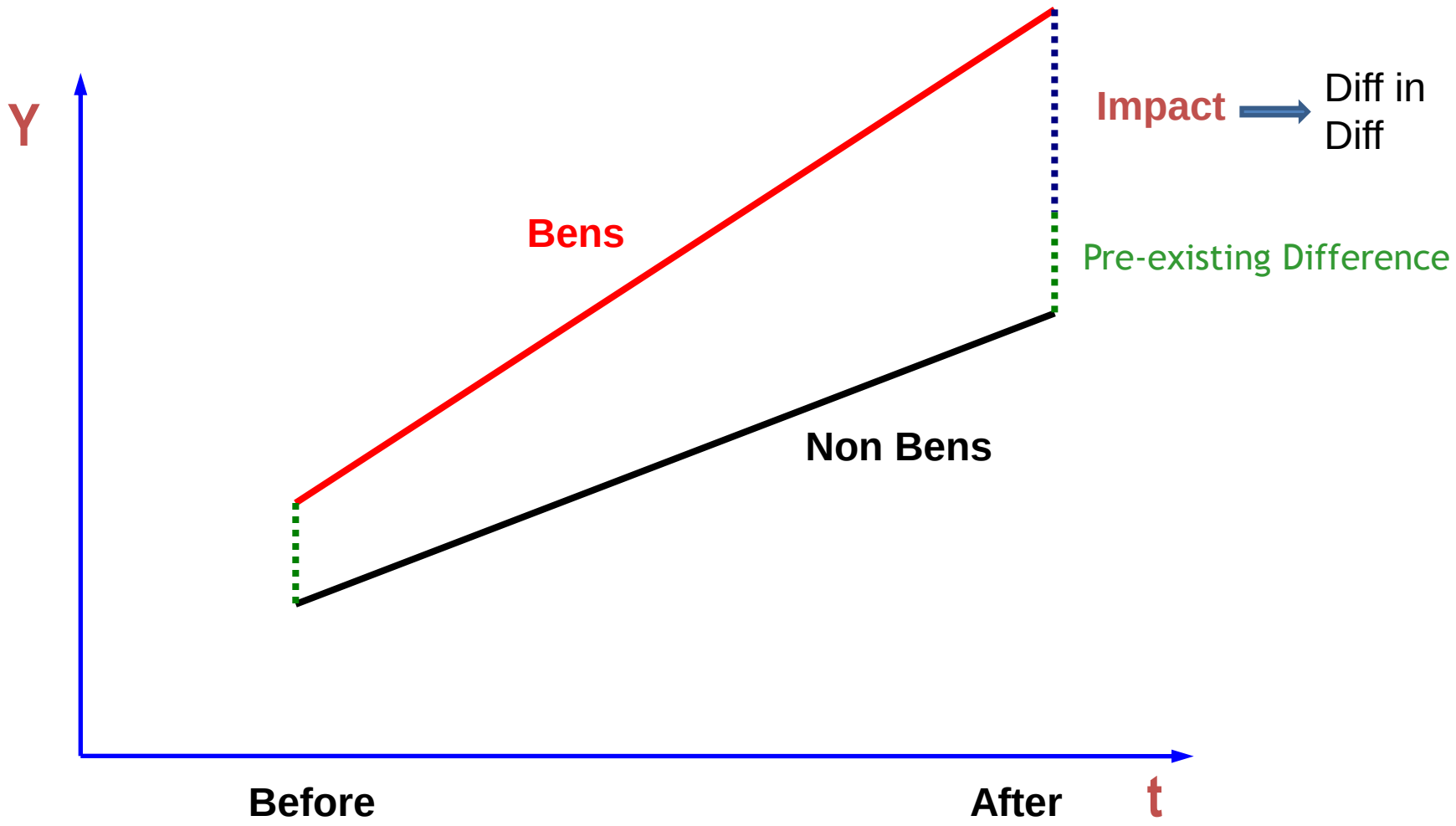
# Impact Evaluation - Method



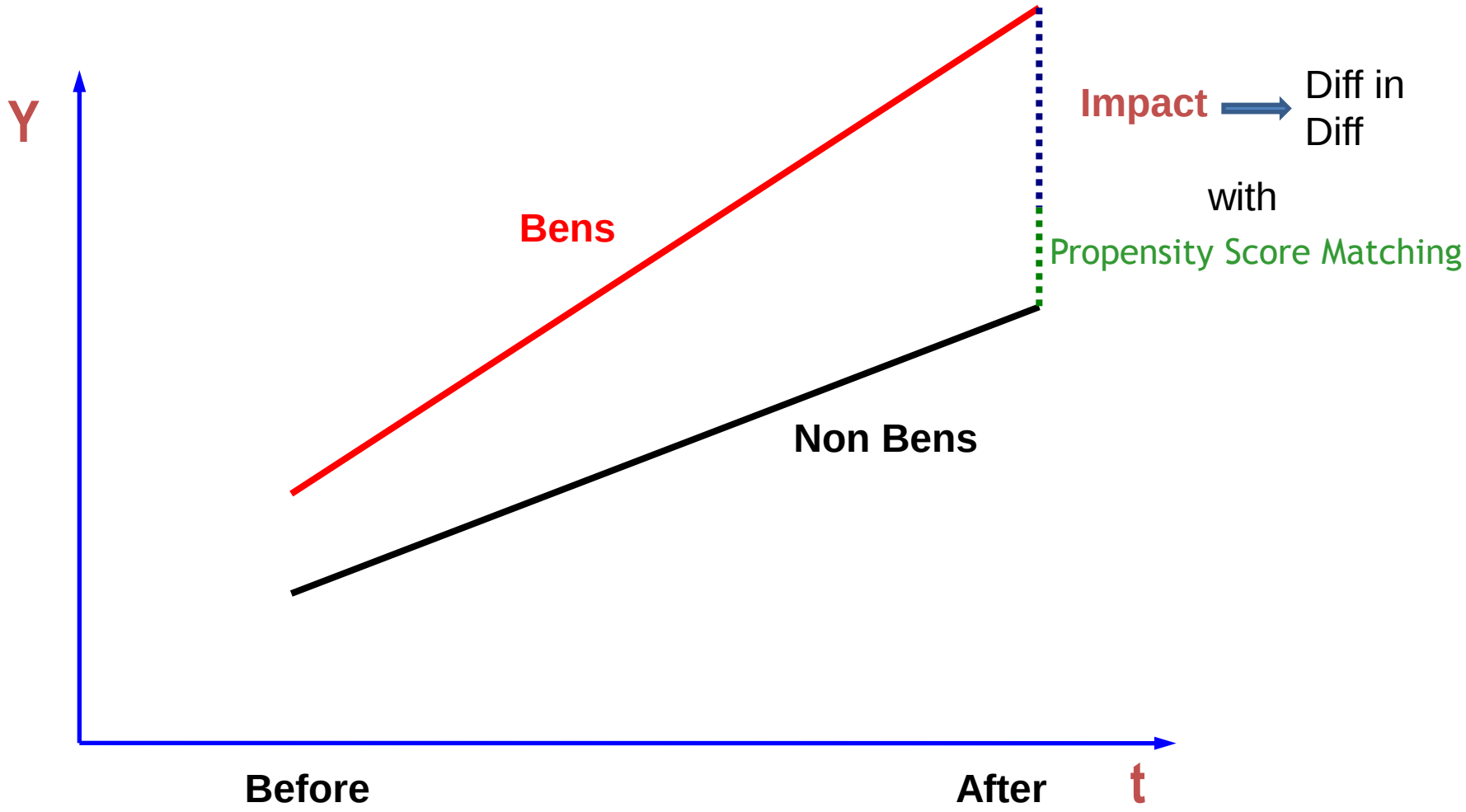
# Impact Evaluation - Method



# Impact Evaluation - Method



# Impact Evaluation - Method

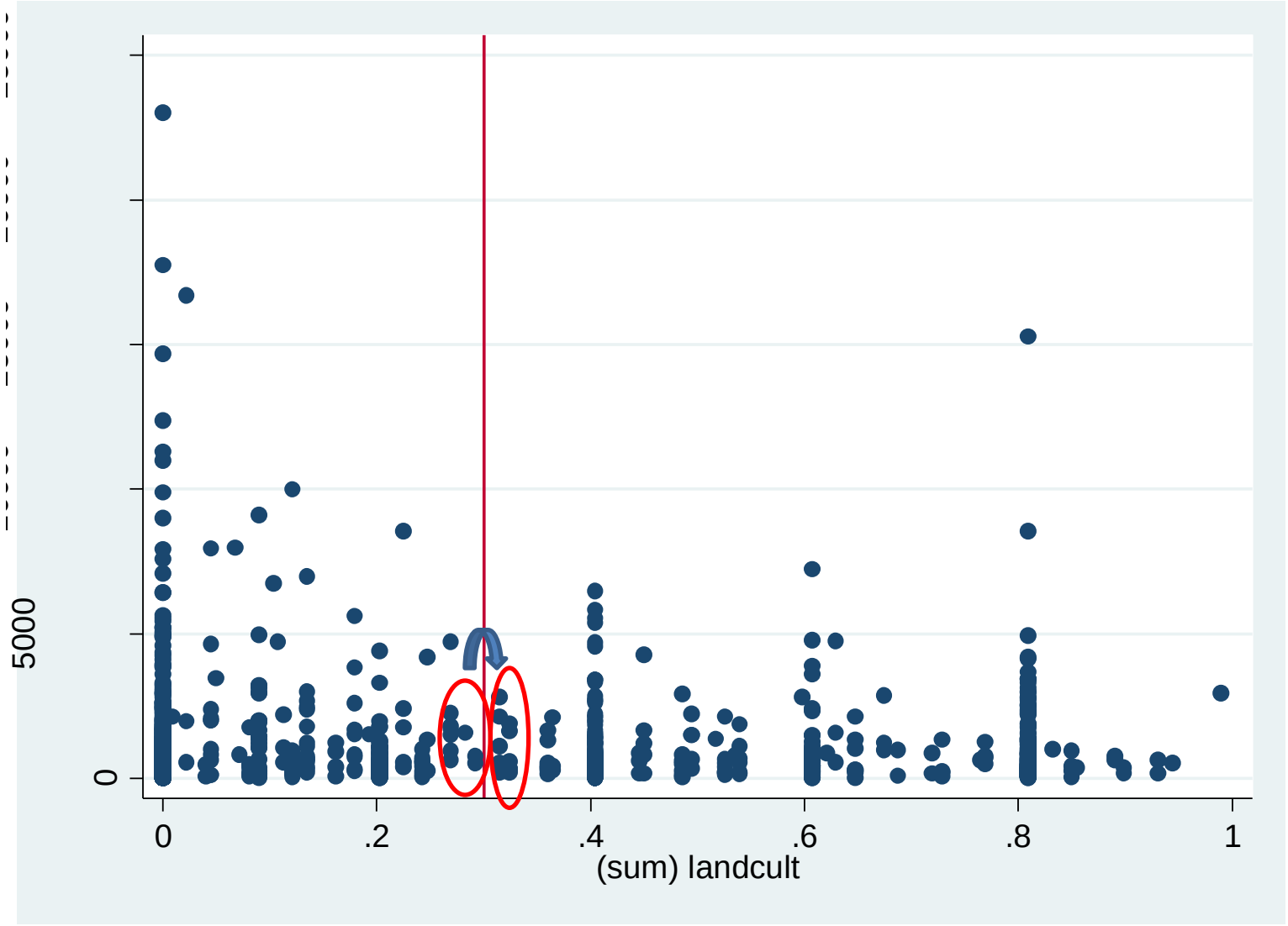


# Impact Evaluation

- What about if we have a sharp eligibility cut-off point?

*-assume the project targets only farmers with <.3 ha*

# RDD

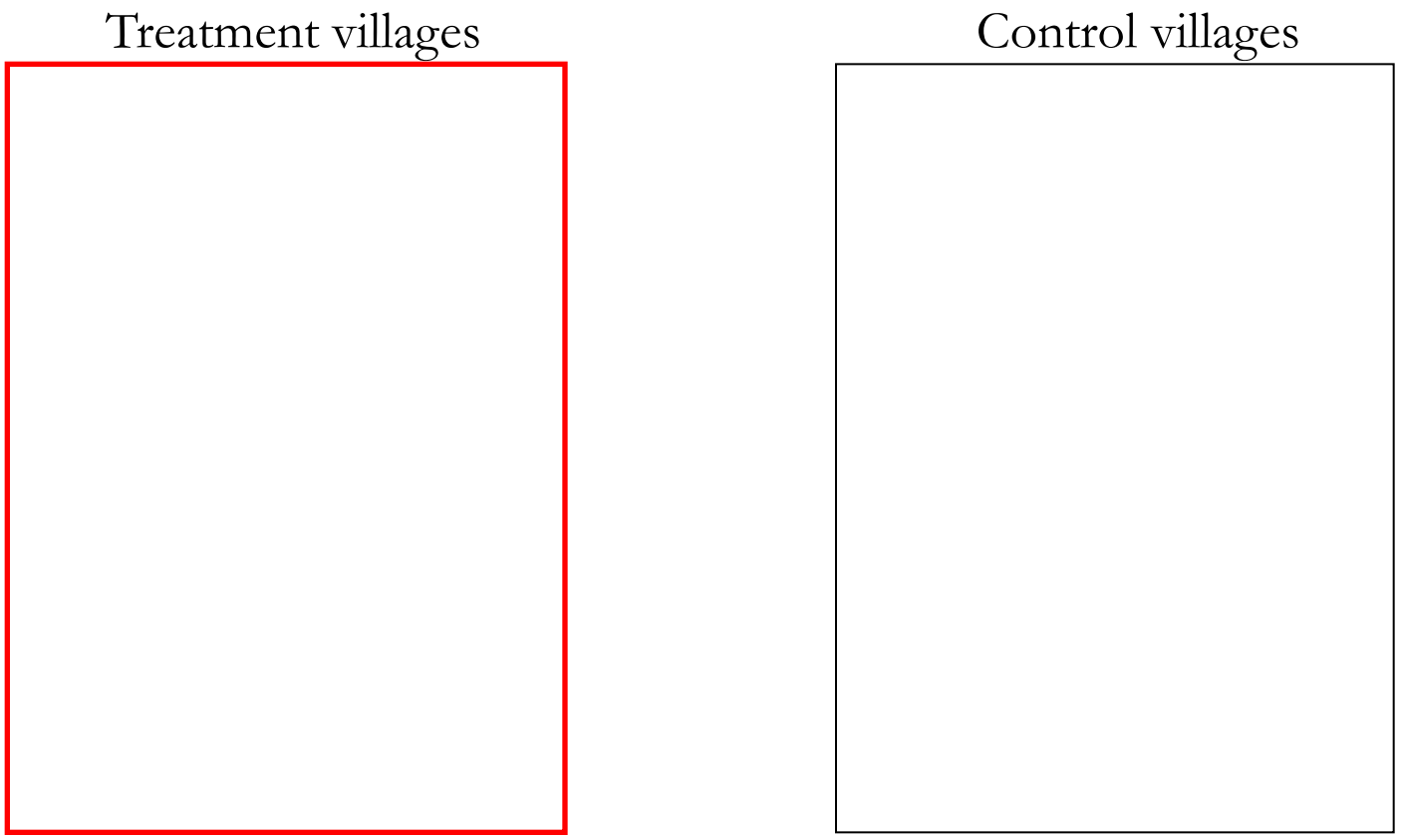


# Impact Evaluation - Method

- Causal effect: change that is due to AR and not to other actors or factors (confounders)
  - … taking into account any other factors also changing during the program period
  - … taking into account any systematic differences between beneficiaries and non-beneficiaries of AR intervention

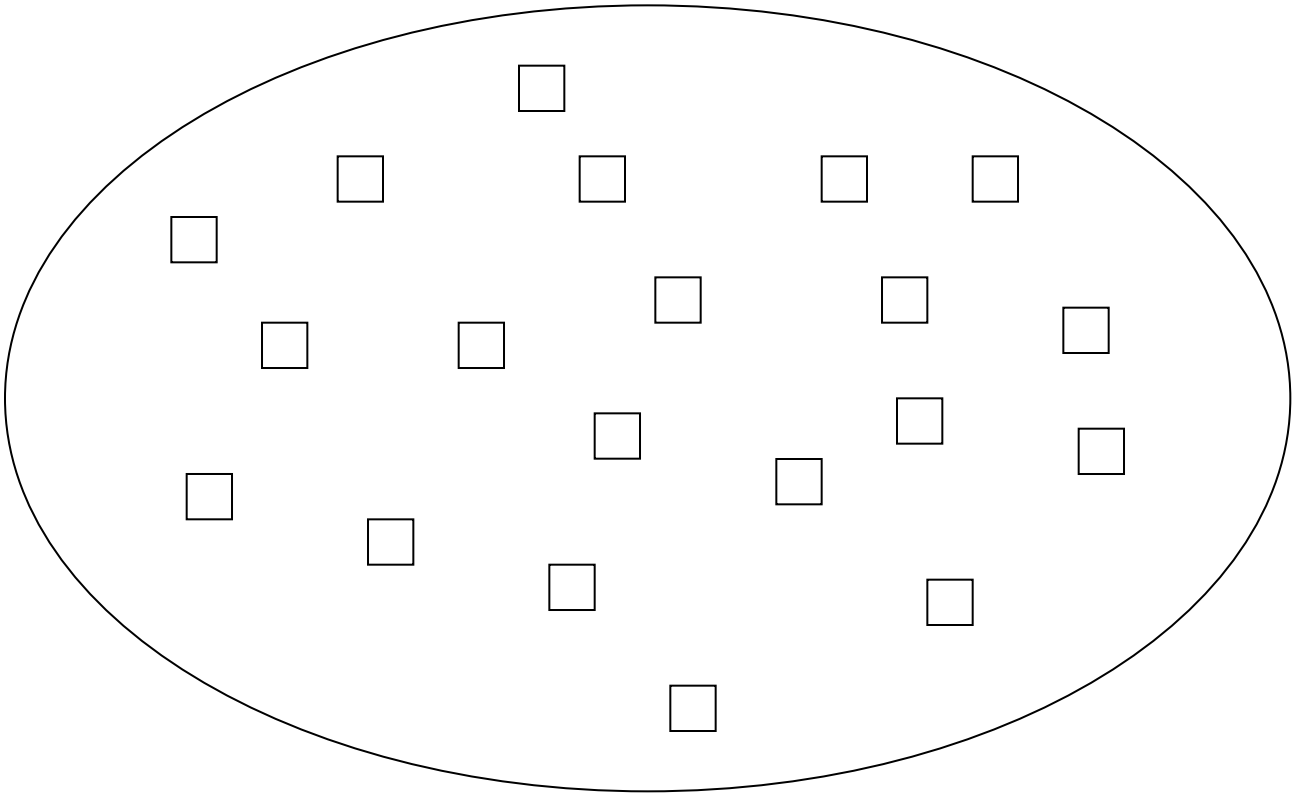
It is very important that the "control group" is comparable to the "treatment group"

# Evaluation Design



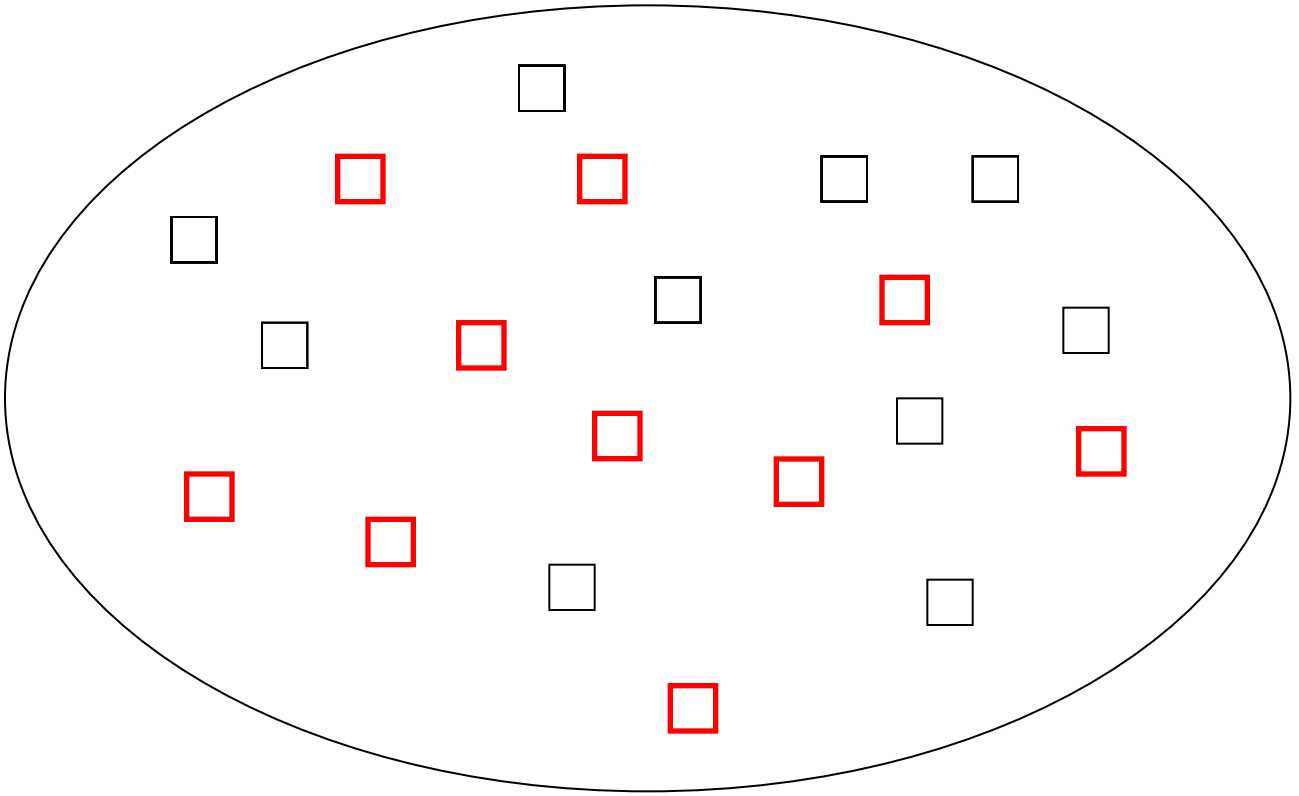

- How can we ensure that treatment and control **villages** are comparable?

# Evaluation Design



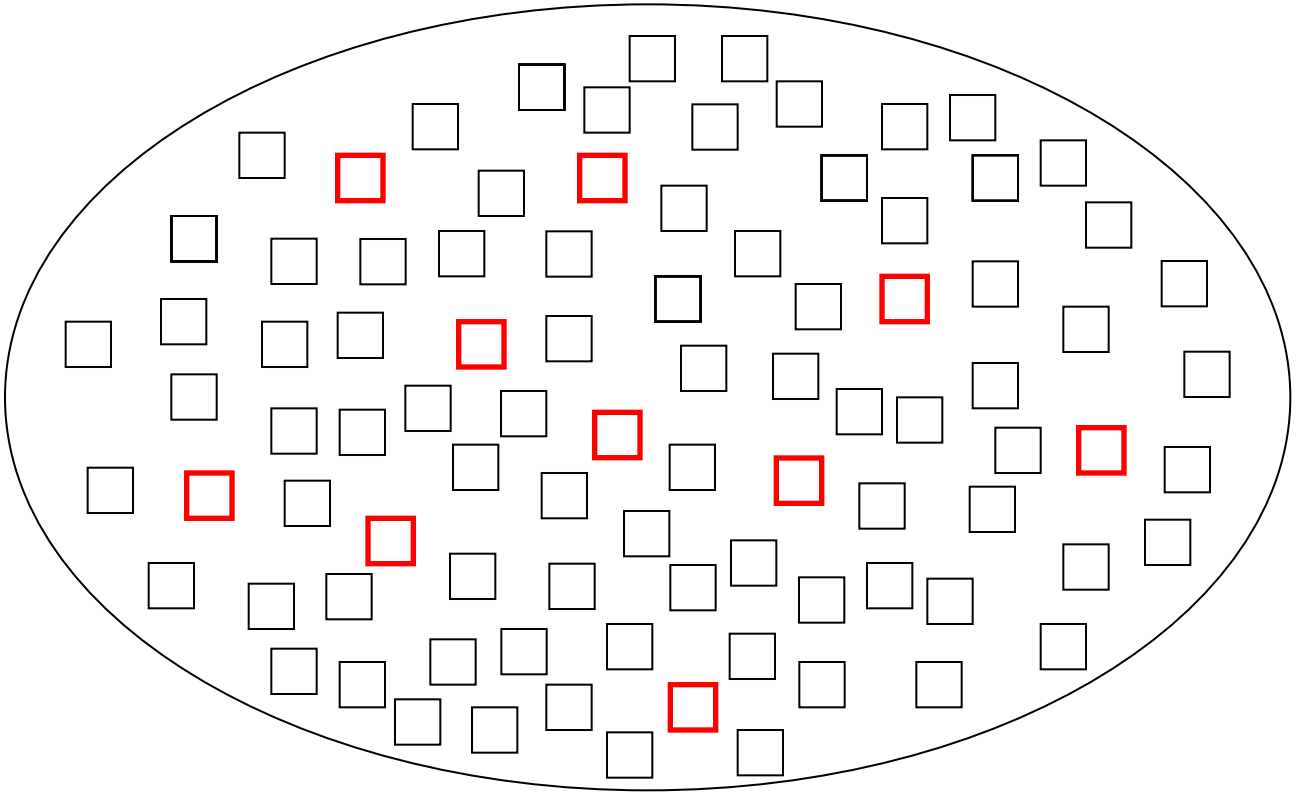
- How can we ensure that treatment and control **villages** are comparable?

# Evaluation Design



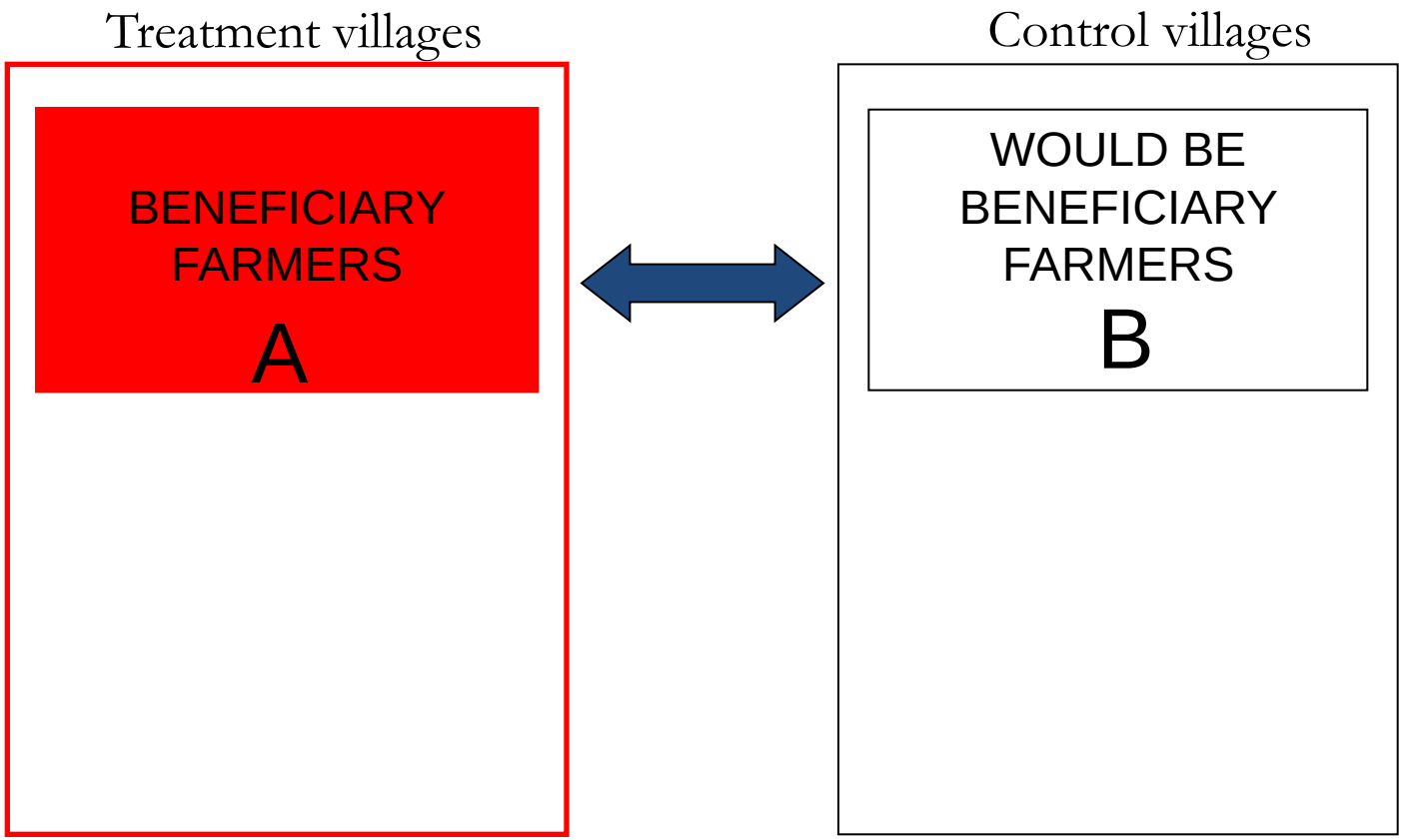
Random Treatment Assignment

# Evaluation Design

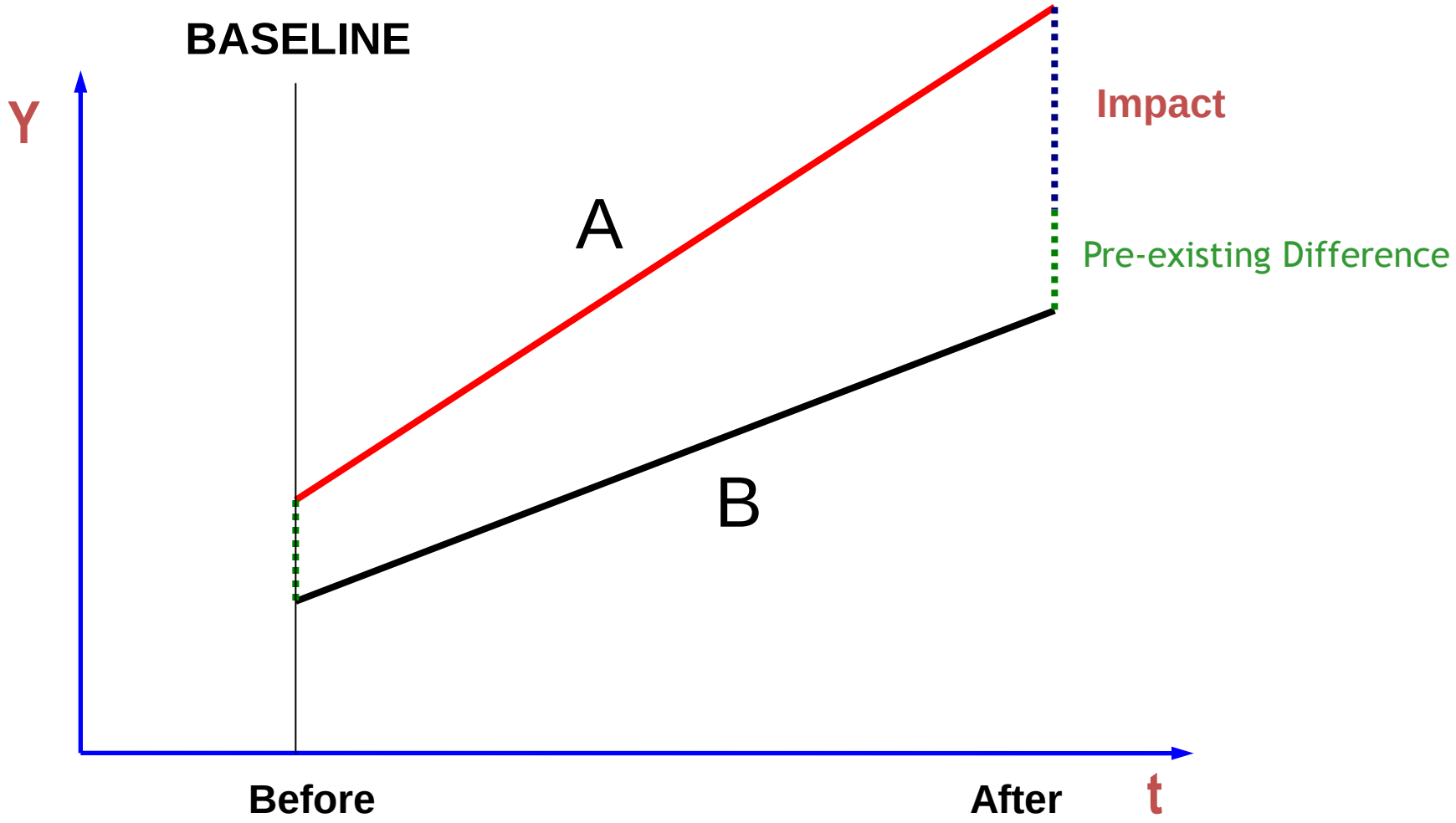


Random Treatment Assignment

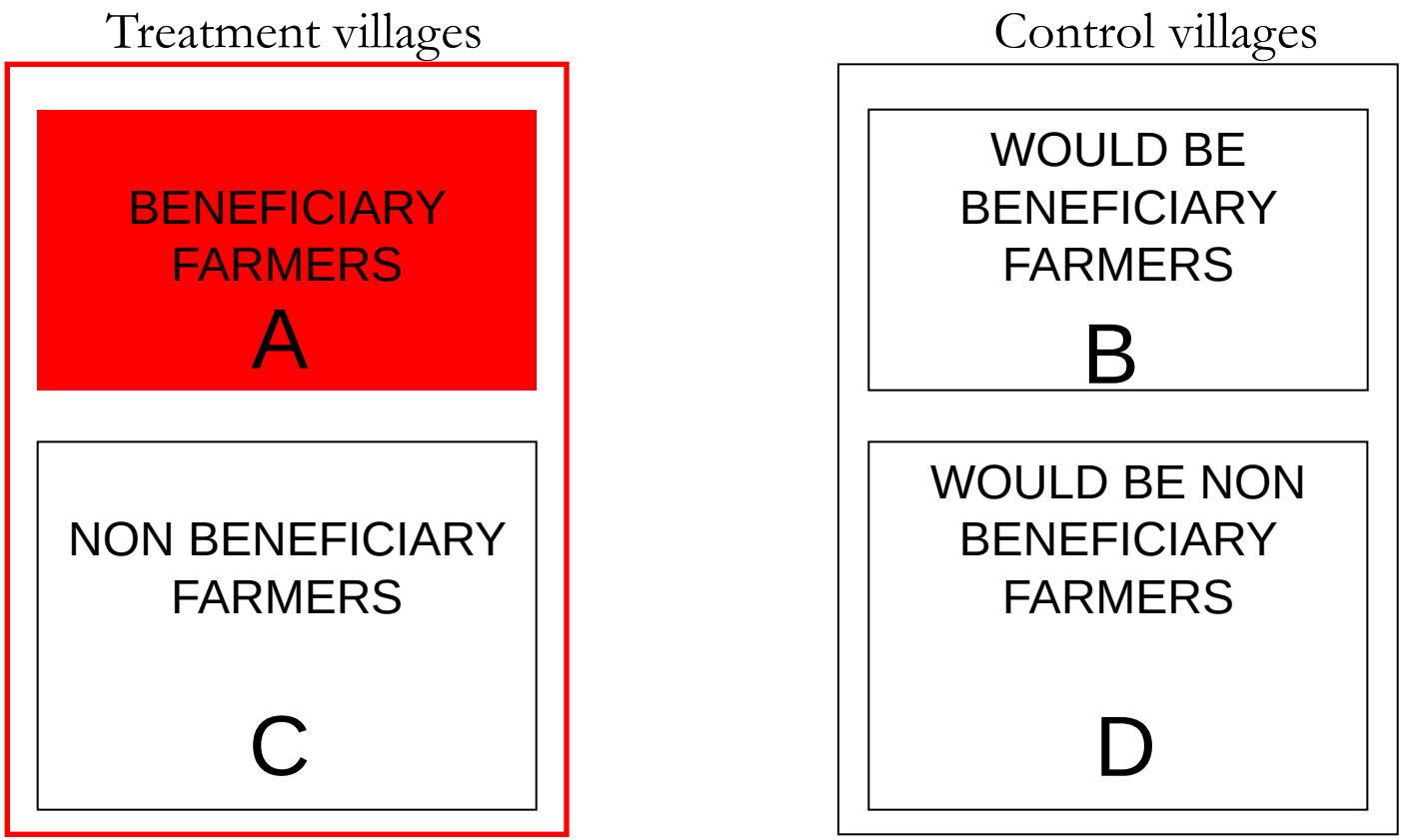
# Evaluation Design



# Where do we stand?



# Evaluation Design



# Randomization/1

- Experiments can be seen as lotteries that randomly assign subjects to separate groups, each of which is offered a different "treatment"
- Randomizing subjects to experimental groups eliminates all systematic pre-existing group differences, as only chance determines which subjects are assigned to which group
- After the experiment, we compare the outcome of interest in the treatment and the control group

  **Effect = Mean in treatment - Mean in control**

# Randomization/2

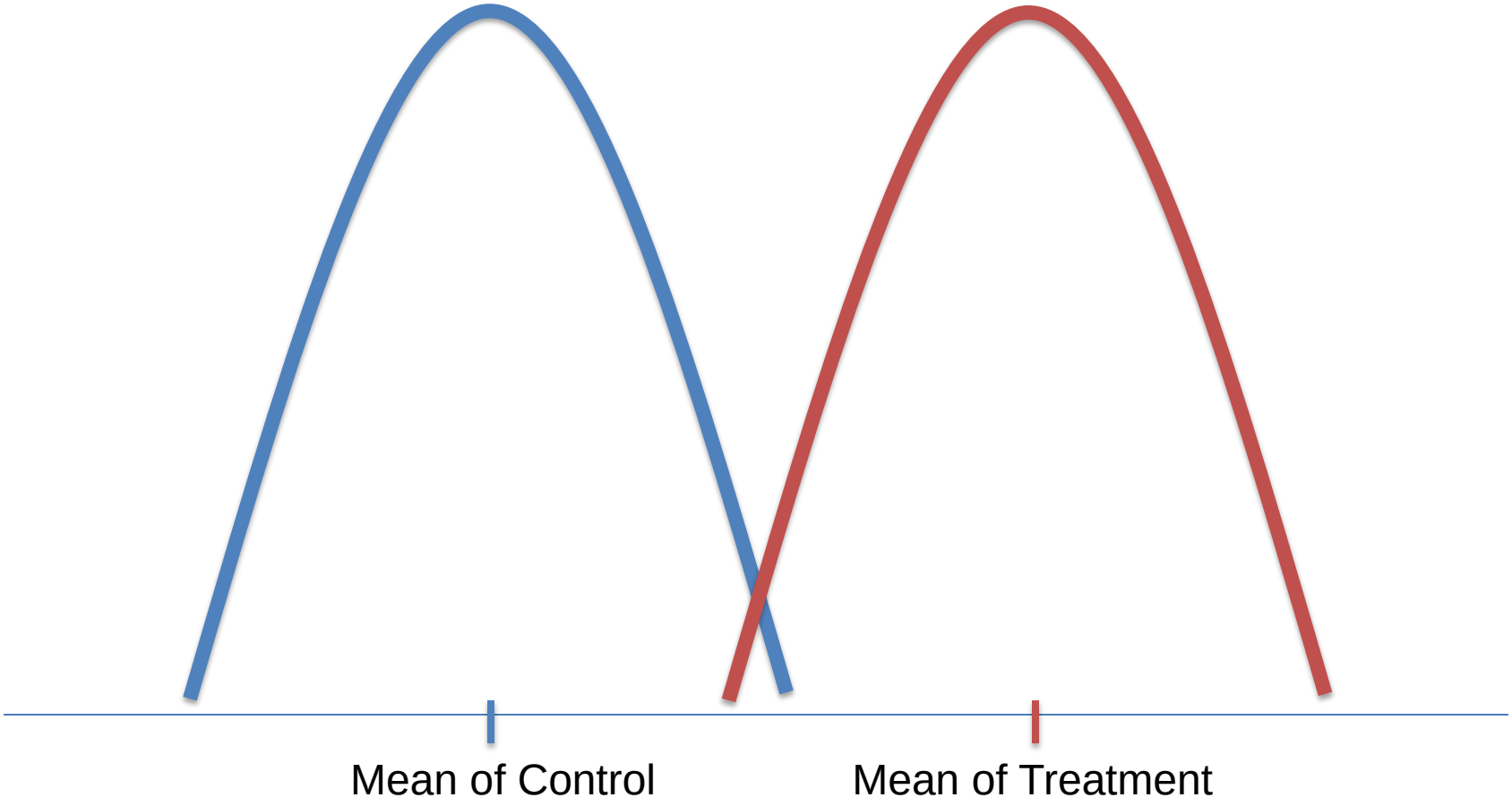
- The design of each experiment differs, but generally we use a two stage approach to selecting a random sample for an impact evaluation.
- First we select larger areas called primary sampling units [PSU] into treatment and control.
- Then we select units of analysis (such as households or farms or clinics) within the selected PSUs.
- These resulting groups have no systematic differences and will be what we compare in the analysis.

# Sample Size

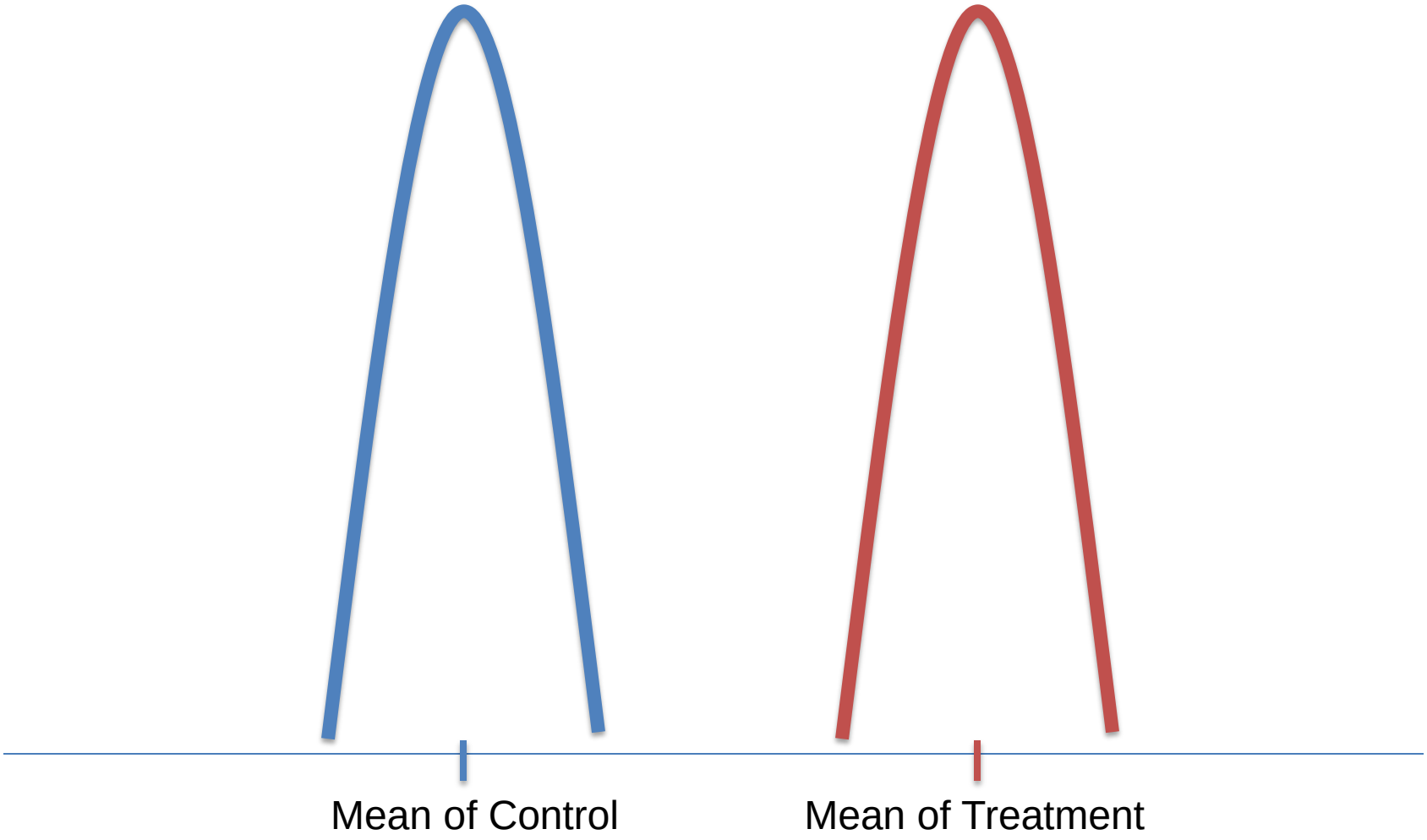
- Now, once you have your sampling frame, all you need to know is how many PSUs and units of analysis you need to credibly measure the impact of your project.

$$N = \left[ \frac{4\sigma^2 (z_{\alpha/2} + z_{\beta})^2}{D^2} \right] \left[ 1 + \rho(m-1) \right]$$

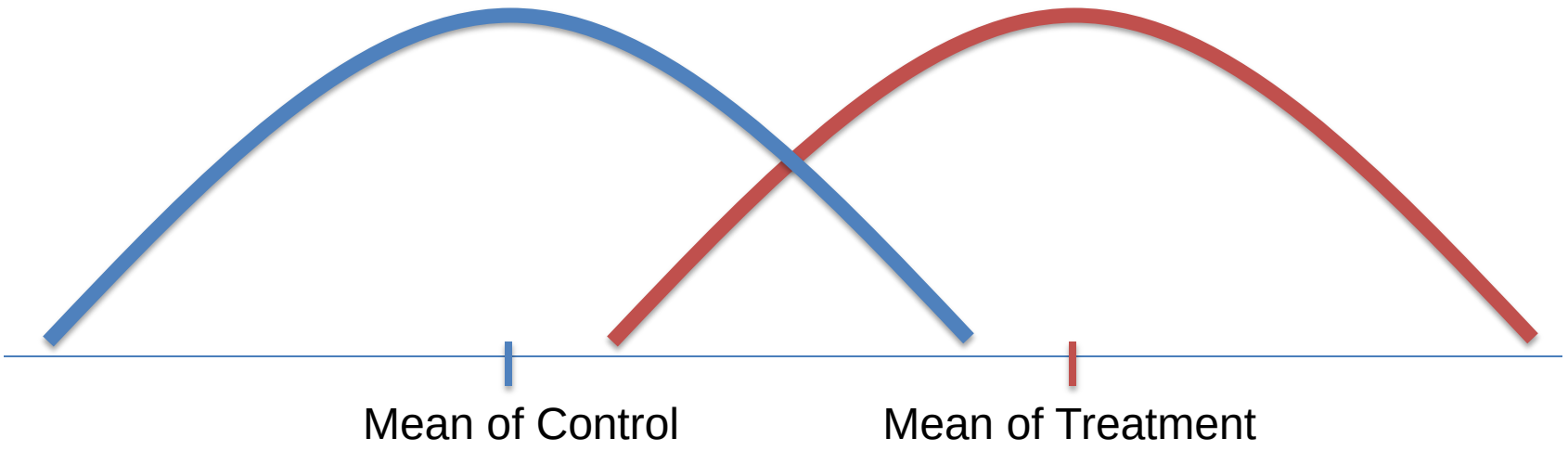
# Sample Size



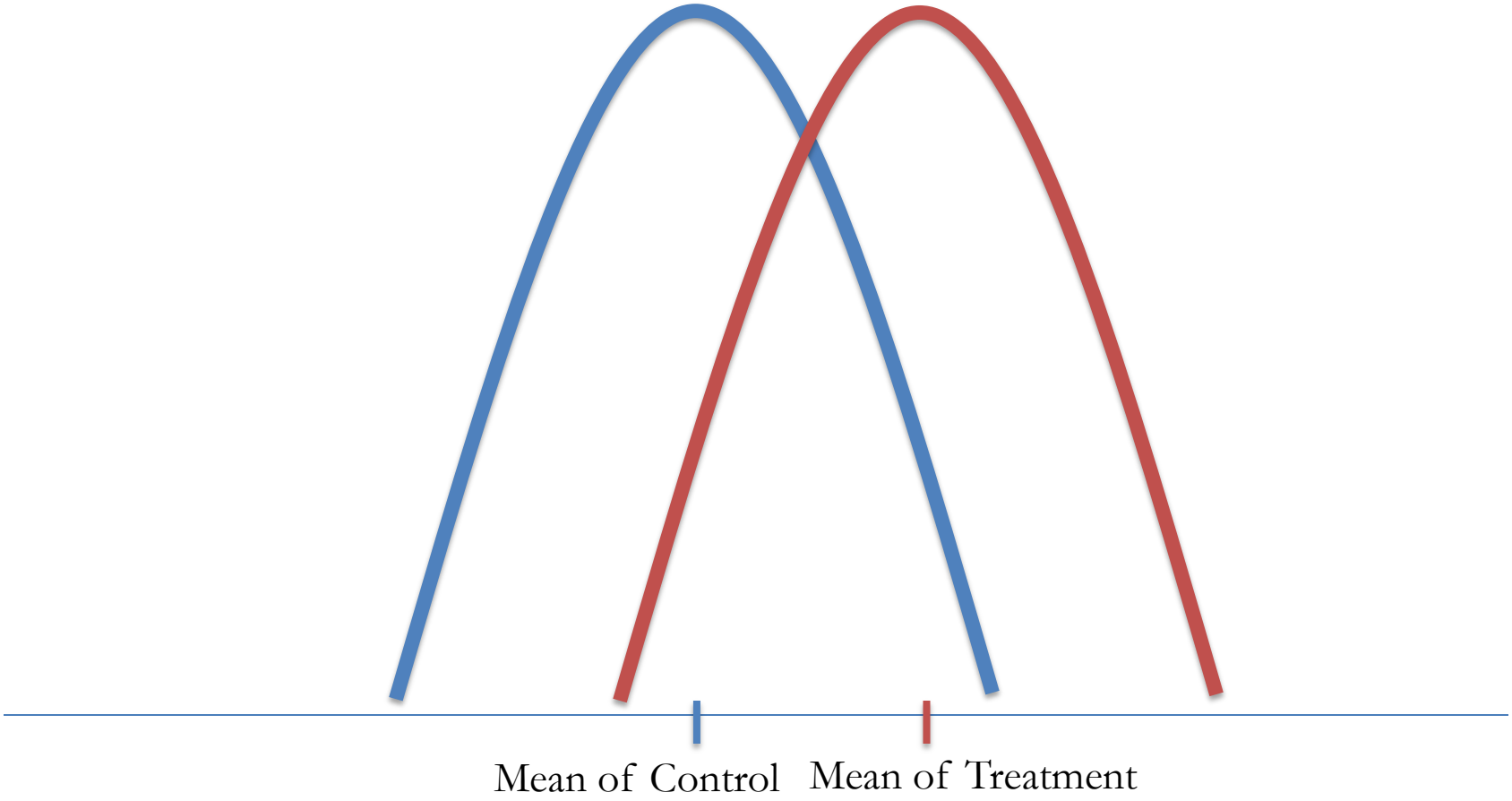
# Sample Size



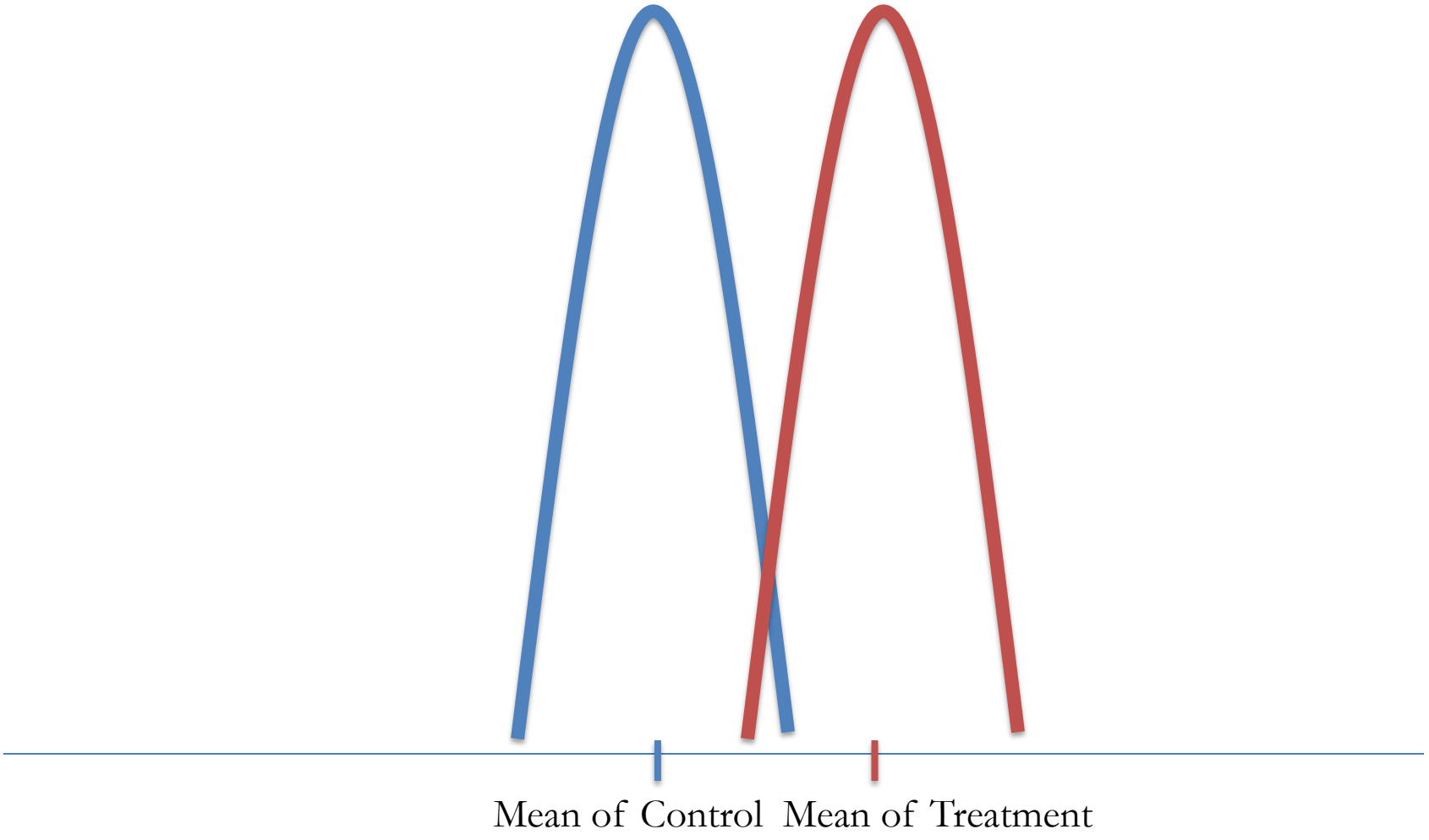
# Sample Size



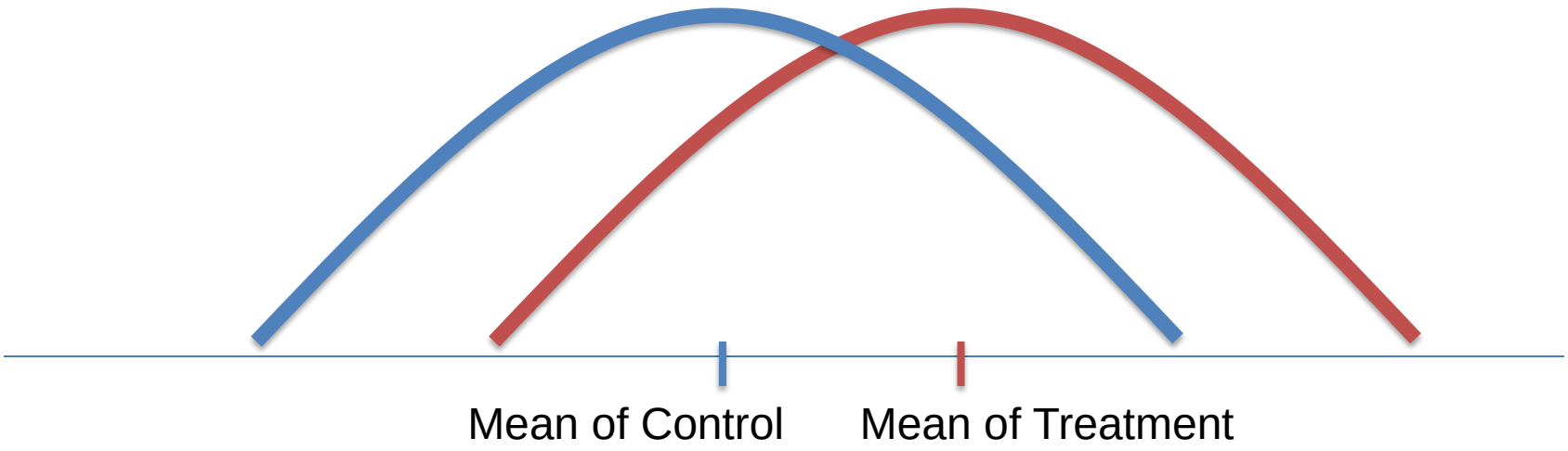
# Sample Size



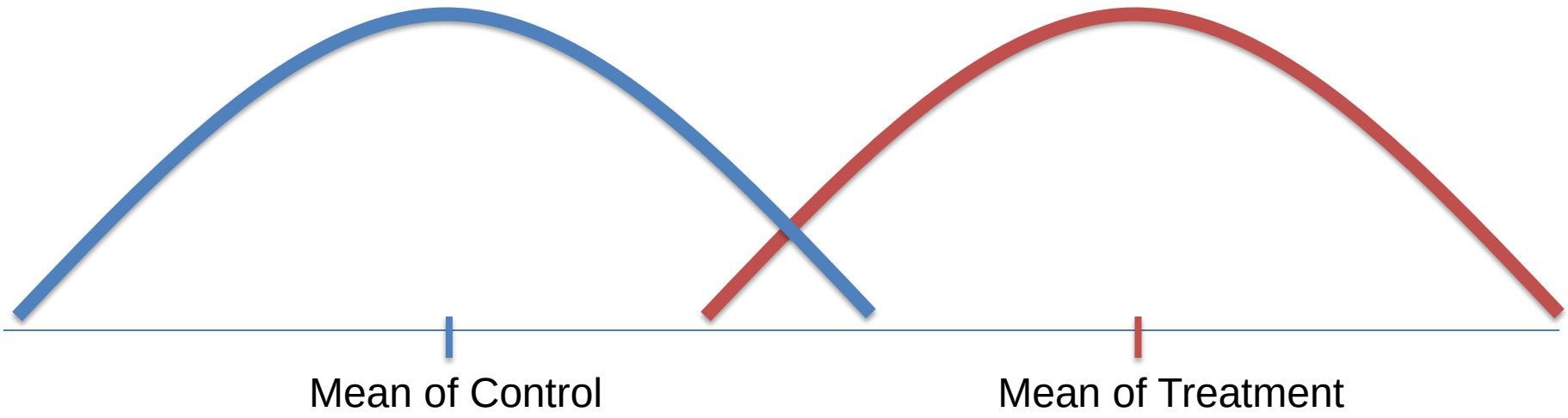
# Sample Size



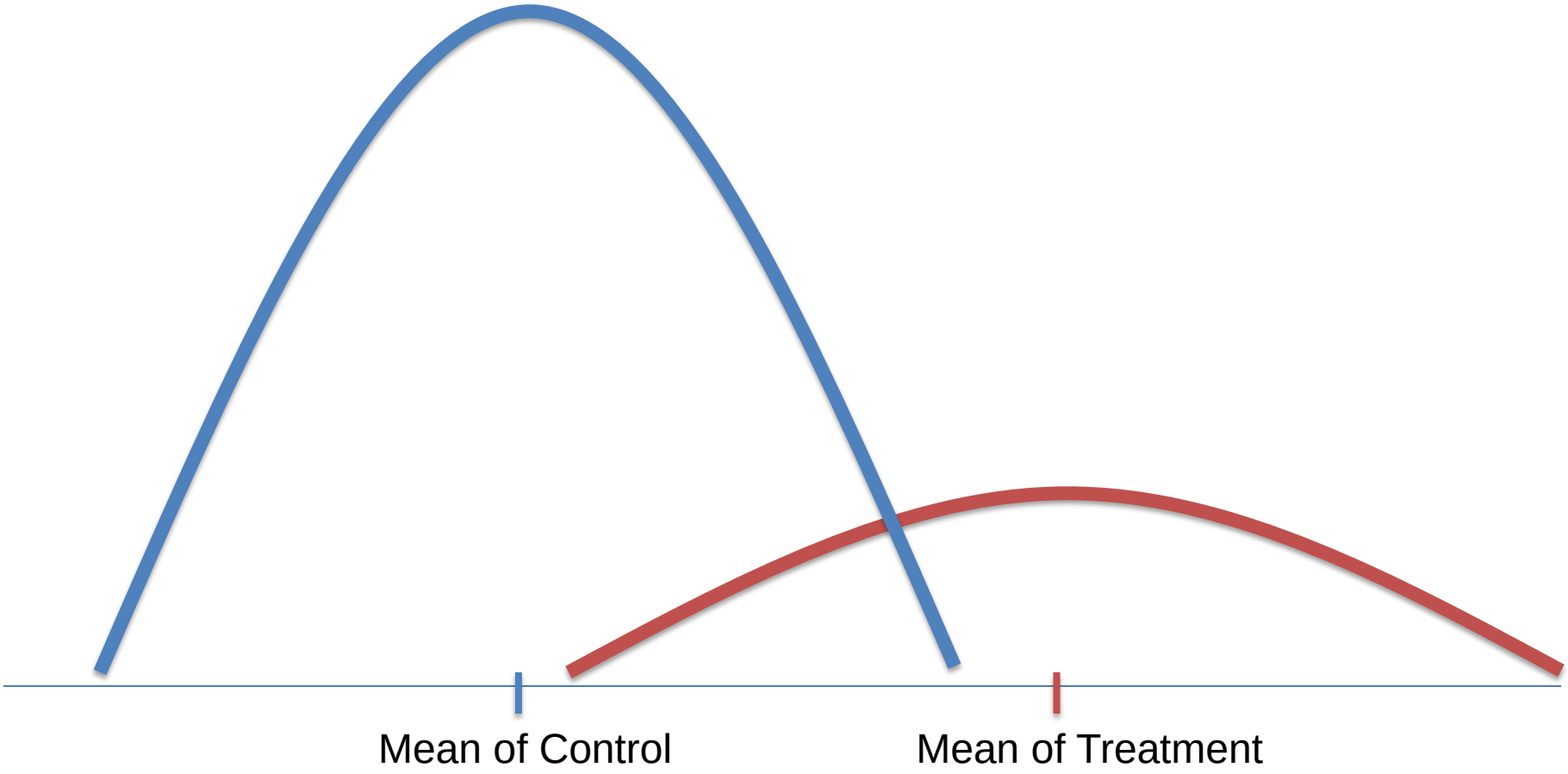
# Sample Size



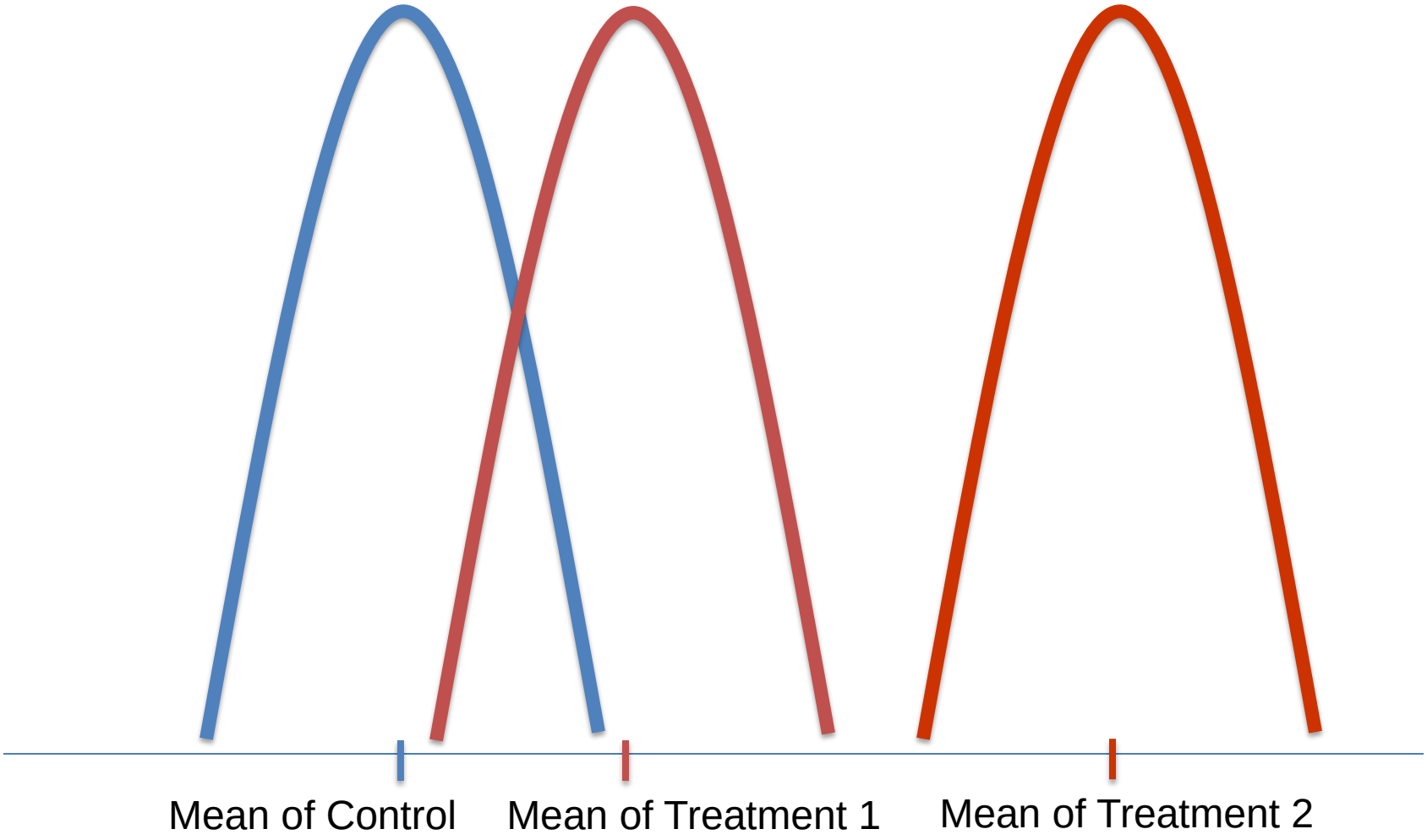
# Sample Size



# Sample Size



# Sample Size



# Sample Size

$$N = \left[ \frac{4\sigma^2 (z_{\alpha/2} + z_{\beta})^2}{D^2} \right] \left[ 1 + \rho(m-1) \right]$$

- Sigma ($\sigma^2$) is the variance in population outcome metric
- Basically means how wide of a range of differences you expect in the outcome that you will measure.
- This can be difficult to calculate – the best way is if you have data collected previously (national household survey, project assessment, piloting data, etc).
- If not, estimations can be made using "(high-low)/4" as a rule of thumb.

# Sample Size

$$N = \left[ \frac{4\sigma^2 (z_{\alpha/2} + z_{\beta})^2}{D^2} \right] \left[ 1 + \rho(m-1) \right]$$

- D is the effect size or how much of an impact your project will have.
- Trade off between sample size and effect – the smaller an effect the bigger a sample size that you will need.
- **Be careful about picking too big of an effect size as you are setting yourself up for failure.**

# Sample Size

$$N = \left[ \frac{4\sigma^2 (z_{\alpha/2} + z_{\beta})^2}{D^2} \right] \left[ 1 + \rho(m-1) \right]$$

- Z's are from standard normal cumulative distribution function and they relate to the certainty of your conclusions.
- The values of z are taken from a table depending on the values of α and β.
- α relates to "type I error" and β relates to "type II error"

# Type I Error (α)

- Significance level: Probability that you will falsely conclude that the program has an effect when in fact it does not.
- Type I error: Conclude that there is an effect, when in fact there are no effect.
- You select level of 5%, you can be 95% confident in the validity of your conclusion that the program had an effect
- For policy purpose, you want to be very confident of the answer you give: the level will be set fairly low.
- The more confident you want to be in your answer, the lower level you will need to select and the bigger your sample will need to be.

# Type II Error (β)

- Power: Probability to find a significant effect if there truly is an effect
- Type II error: Fail to reject that the program had no effect when it fact it does have an effect
  - Common values used are 80% or 90%.
- One minus the power is the probability to be disappointed. (So if you pick a power of 80%, there is a 20% chance that even though your project does have an impact, the evaluation will fail to detect it.)
- The more power you want your test to have, the larger a sample size you will need.

# Sample Size

$$N = \left[ \frac{4\sigma^2 (z_{\alpha/2} + z_{\beta})^2}{D^2} \right] \left[ 1 + \rho(m-1) \right]$$

- This part of the equation relates to how many clusters and households you select into your sample.
- Rho (ϱ) is the intracluster correlation coefficient. This is a measure of how similar your observations within each PSU tend to be.
- $m$ is the number of observations in each cluster (take).
- The more similar households are to each other and the more households you have in each cluster, the higher overall sample size you will need.

# Sample Size

$$N = \left[ \frac{4\sigma^2 (z_{\alpha/2} + z_{\beta})^2}{D^2} \right] \left[ 1 + \rho(m-1) \right]$$

- The reason that ϱ raises your sample size is because to more alike this are within a cluster, the less likely they are to be representative of the whole area.
- $\varrho$'s are generally high for infrastructure projects, because either the whole village has assess to a road or water source, or the whole village does not.
- $\varrho$'s for contraceptive projects tend to be low, because while a neighbor's actions might influence a woman, decisions about children are generally made within the family.

# Sample Size

$$N = \left[ \frac{4\sigma^2 (z_{\alpha/2} + z_{\beta})^2}{D^2} \right] \left[ 1 + \rho(m-1) \right]$$

- Final note: Beware the square!
- There is not a 1 to 1 relationship between sample size and most of the terms that are used to calculate it.
- So halving the size of the effect that you are looking for will raise required sample size by 4 times.

# Sample Size

$$N = \left[ \frac{4\sigma^2 (z_{\alpha/2} + z_{\beta})^2}{D^2} \right] \left[ 1 + \rho(m-1) \right]$$

- Two of the items in this formula are "fixed" – specifically the population variance ($\sigma^2$) and the intracluster correlation effect ($\varrho$). Nothing can be done in the design stage to change these values.
- There is some scope to change the "z" values but it is limited. Most credible impact evaluations will not dip below a 90% confidence or an 80% power.
- That leaves only the effect size ($D$) and the cluster size ($m$) as parameters that can be manipulated.

# Panel data

$$N = \left[ \frac{4\sigma^2 (z_{\alpha/2} + z_{\beta})^2}{D^2} \right] [1 + \rho(m-1)][1-R]$$

- The main benefit of panel data is that reduces sources of variation down to the level of the unit of observation.
- The correlation in the indicator of interest between the baseline and endline is *R*. The higher the correlation the more of a benefit from using panel data.
- In a cross-sectional survey, the value of *R* is zero.

# Panel Data

- Like many aspects of sample design, *R* will have to be estimated in advance.
- Sample values of *R* include:
  - Households being poor: 0.3
  - Children 6-15 years attending school: 0.6
  - Children 1-3 years being fully immunized: 0.85
  - Stunting among children 1-3 years: 0.3
- Therefore using panel data reduces the sample requirement by 30% for a poverty study and by 85% for an immunization study.

# Sample Size

| Decrease Necessary Sample Size |
| --- |
| Lower Variance |
| Bigger Effect Size |
| More Clusters |
| Panel Data |
| |
| **Increase Necessary Sample Size** |
| Higher Confidence ($\alpha$) |
| More Power ($\beta$) |
| Clusters More Similar |
| More Observations Per Cluster |

# Other Considerations

- Stratification
  - Partition sample to ensure sufficient number of observations in all categories
- Oversampling
  - Larger proportion of observations from certain strata than proportion in overall population
- Sample Weights
  - Used to account for oversampling when making inferences about overall population

# Other Considerations

- Treatment Arms: Formula above refers to only 2 treatment arms - having multiple treatment arms in a program increases the required sample size quickly.
- The sample size calculations give you the total sample size for a two-arm evaluation. If you decide you want to add a third arm – you will need another 50% jump in the sample.

# Non Random Sample Design

- Randomization in an impact evaluation is not always possible – may want to consider other designs such as Propensity Score Matching or Regression Discontinuity.
- In PSM, basic rule of thumb is to collect as many observations as possible to get best match for treatment.
  - See David McKenzie's blog from November 2011 for more details http://blogs.worldbank.org/impactevaluations/node/693
- In RDD, design effects dramatically increase sample size. Individual calculations necessary but can be estimated at roughly 3-4 times random sample.

# Final Consideration

- In reality there are generally finite resources available to do impact evaluations. Many time we end up doing a series of calculations varying the components to see just how much power and certainty we can afford.
- While there is a certain amount of guesswork involved in calculations, it is important to spend effort on them:
  - Avoid launching studies that will have no power at all: waste of time and money, potentially harmful
  - Devote the appropriate resources to the studies that you decide to conduct (and not too much)