RANDOMIZED TRIALS

Technical Track Session II

Sergio Urzua

University of Maryland

Randomized trials

- Evidence about counterfactuals often generated by randomized trials or experiments
  - Medical trials
- Eliminates common biases (or confounders) when done properly
  - Selection bias
  - Trends concurrent with intervention
- Therefore, often considered the gold standard of estimating causal impacts

Randomized trials

- Not magic
  - Still subject to basic constraints of statistics
    - Need large samples
    - Drop out, non-compliance a problem
    - Though not biased, estimated parameters might differ from desired parameters
  - Sometimes not politically feasible

Outline

1. Randomization solves selection bias

2. What should be the unit of randomization?
   i. Bias
   ii. Statistical power
   iii. Externalities

3. How do you actually randomize?

4. Stratification (what is it, why do we need it)

5. Difference between random sampling and randomization

6. Other issues
   i. Attrition
   ii. Compliance (both for subjects and implementers)
   iii. Estimated parameters

7. Non-randomized methods

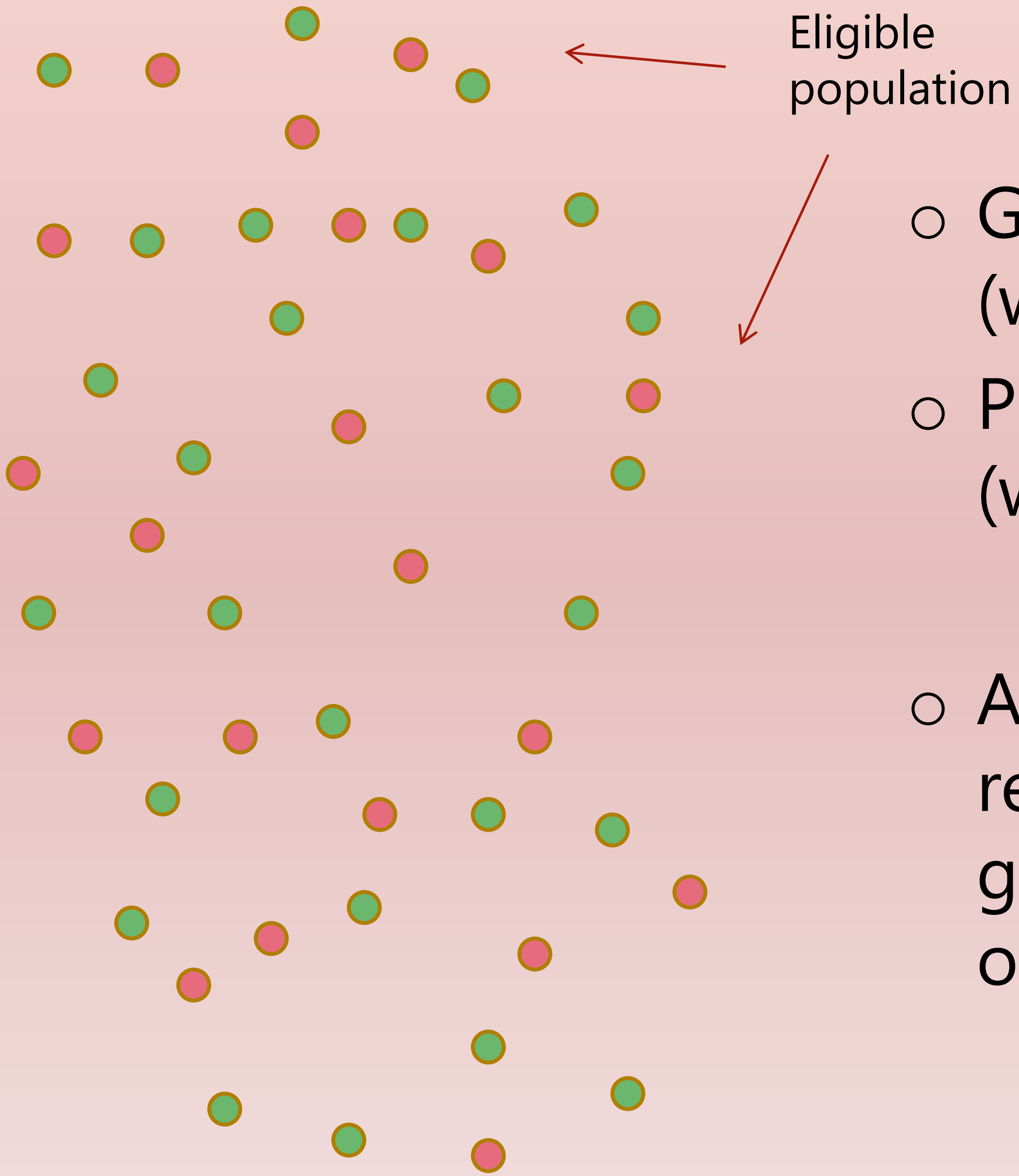
Randomized trials overcome potential confounders

- Let's return to earlier examples:
  - Health insurance
  - Conditional cash transfers

- Bias 1: Selection bias
  - Participants might be innately different from non-participants

- Consider a simple lottery
  - Take all eligible people in population of interest
  - Place all names on slips of paper in a jar
  - Pick half of the slips of paper out of jar
  - Chosen names get intervention, those not chosen do not

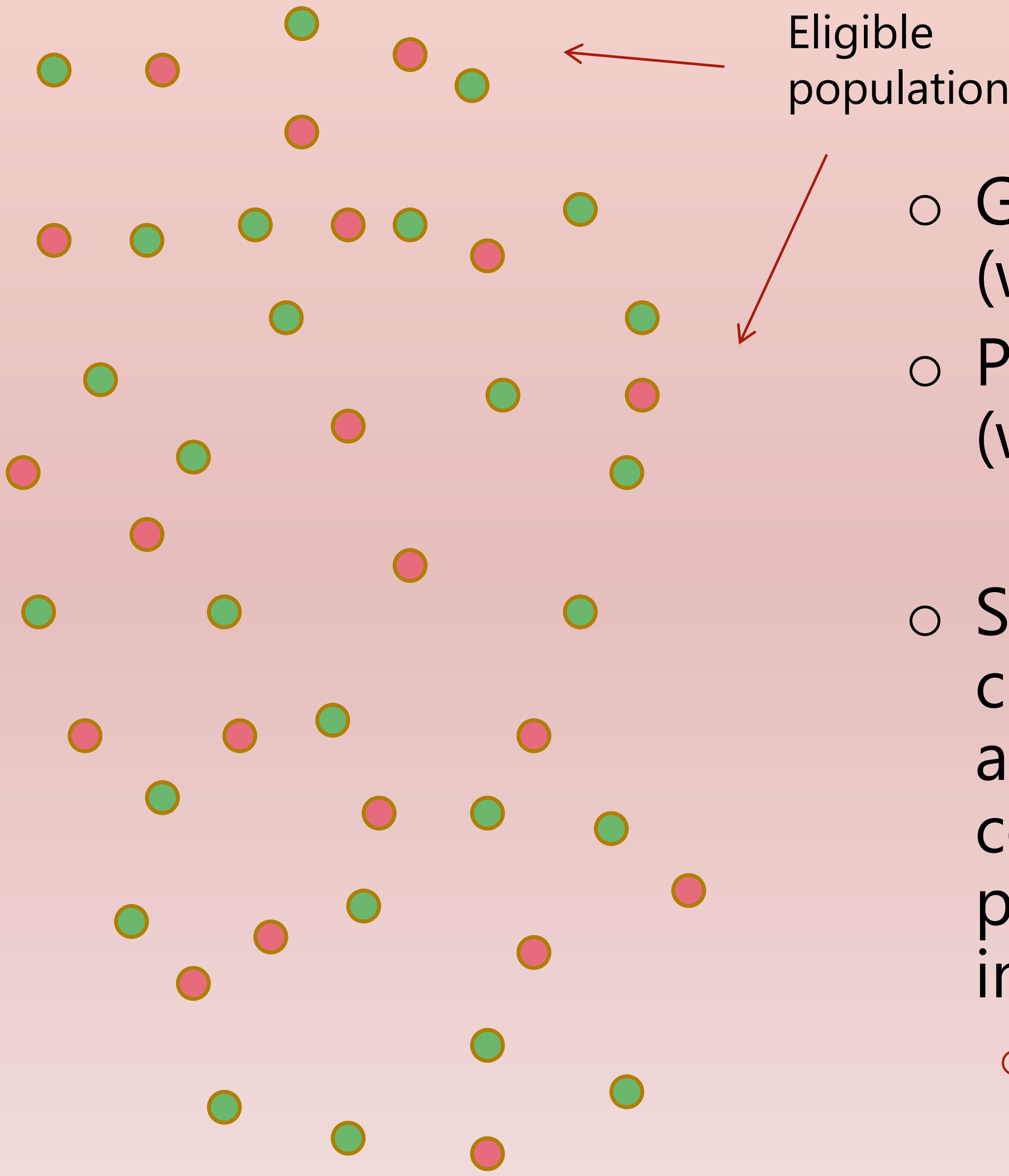
Bias 1: Selection bias



- Green = treatment (with intervention)
- Pink = comparison (without intervention)

- Assume this array represents geographical spread of sample population

Bias 1: Selection bias



- Green = treatment (with intervention)
- Pink = comparison (without intervention)

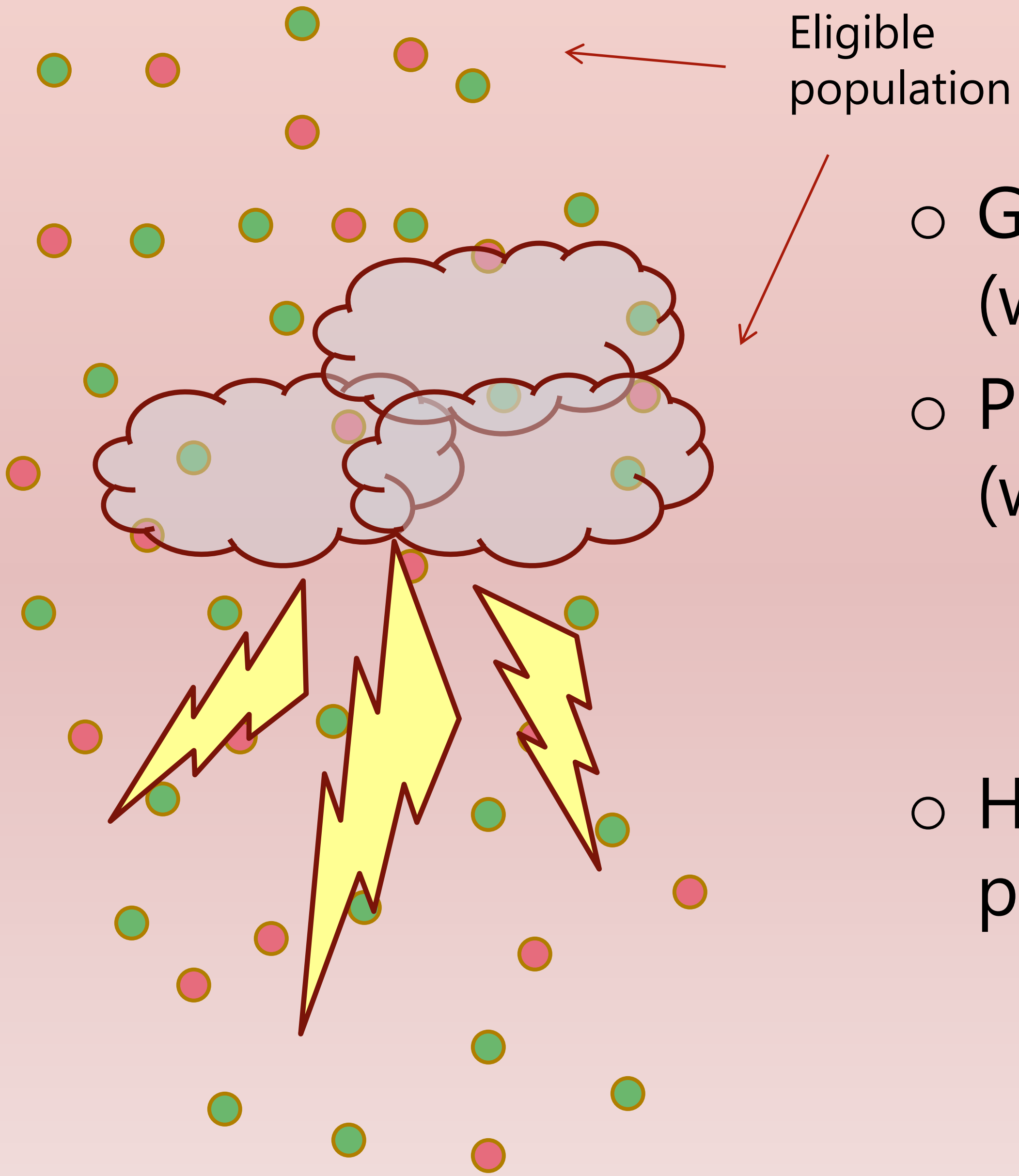
- Should average characteristics differ across treatment and comparison groups prior to the intervention?
  - No.

Bias 1: Selection bias

- Average characteristics should be the same for treatment and comparison groups prior to the intervention
  - Expenditure
  - Health status
  - Motivation to send children to school
  - Fear of dogs
  - Everything!

- So prior to a health insurance intervention, average expenditure ($\bar{e}$) should be identical in treatment and comparison groups

Bias 2: Common trends



- Green = treatment (with intervention)
- Pink = comparison (without intervention)
- Heavy rains or other program

Bias 2: Common trends

- When treated units selected randomly, rain shock common to both treatment and comparison groups

- What happens when we look at health expenditures of both groups after the intervention?
  - Average outcome for treatment group = ē + impact of health insurance + impact of rains
  - Average outcome for comparison group = ē + impact of rains
  - Difference between treatment and comparison = [ē + impact of health insurance + impact of rains] - [ē + impact of rains] = **impact of health insurance**

Randomization and selection bias more generally

$$\begin{aligned}\bar{\delta} &= E_U[Y_1(u) \mid D=1] - E_U[Y_0(u) \mid D=0] \\ &= E_U[Y_1(u) \mid D=1] - E_U[Y_0(u) \mid D=0] + E_U[Y_0(u) \mid D=1] - E_U[Y_0(u) \mid D=1] \\ &= E_U[Y_1(u) - Y_0(u) \mid D=1] + \underbrace{E_U[Y_0(u) \mid D=1] - E_U[Y_0(u) \mid D=0]}_{\text{Selection bias: Difference in average untreated outcomes between treatment and comparison groups}}\end{aligned}$$

Randomization solves selection bias

- Randomization ensures that
  - Treatment and comparison groups differ in expectation only through exposure to treatment
  - Therefore, in absence of treatment, outcomes should have been the same for both groups
  - Therefore,

$$E_U[Y_0(u) \mid D=1] - E_U[Y_0(u) \mid D=0] = 0$$

Randomization solves selection bias

- Since selection bias is equal to zero, T (an indicator for D=1) is an unbiased estimator of treatment impact

$$y_u = \alpha + \beta T + \varepsilon_u$$

- Control variables
  - Should not affect bias since in expectation treatment and comparison groups should be balanced on controls
  - Can increase precision of estimated impact

Can this be done in practice?

- A few examples implemented in developing countries
  - Textbooks, deworming drugs, contract teachers, performance pay for teachers, merit based scholarships, HIV/AIDS education, school uniforms, health insurance, conditional cash transfers, vouchers to learn HIV results, vouchers for private school, iron supplementation, information about returns to schooling, gender/caste of village leader, fertilizer, micro-credit, school report cards, community score cards, school based management, school meals, savings products, computers in the classroom, interest rates, prices for malaria medicines, prices for mosquito nets, ……..

- See websites of SIEF, Poverty Action Lab, Innovations for Poverty Action and **Development Impact** for more information on studies

The unit of randomization: Why it matters so much

Unit of randomization

- Determines
  1. Extent to which randomization solves selection bias
  2. Statistical power
  3. Ability to measure externalities

Unit of randomization and bias

- Extreme example
  - 1 treatment district and 1 comparison district
- What happens if only 1 district suffers a shock (positive or negative)?
  - Cannot disentangle treatment effect and effect of shock
  - Treatment and comparison district unlikely to be balanced on average traits (law of large numbers cannot apply)
- These concerns still apply when $N^{Treatment} = 5$ and $N^{Comparison} = 5$

Unit of randomization and statistical power

- When do we have enough units?
- Depends on
  - Underlying variance of outcome of interest both across units and within units
    - If underlying variance is high, will need a large sample to separate signal (treatment impact) from noise
    - The more correlated are units within unit of randomization (e.g. households within a village), the more the unit of randomization becomes the effective sample size
- Too few units can lead to low statistical power
  - Perhaps the true treatment impact is non-zero, but your estimates are so noisy (imprecise) that you cannot distinguish them from zero
  - Will not learn anything useful from impact evaluation
    - *Impact could be a 50% improvement or it could be zero – I can't really tell.*
- Therefore, large geographical units not ideal candidates for unit of randomization

Unit of randomization and externalities

- What if we believe that our treatment causes externalities? I.e. controls may be impacted by treatment of others

- Examples
  - Deworming medicine
  - Information campaign

- We might underestimate true treatment impact if individuals randomly selected to receive treatment since comparison group also indirectly benefits

- What can we do?

Unit of randomization and externalities

- We can we do?
  - Randomize at a more aggregate level, and
  - Make sure to measure degree of connectedness among units within treatment and comparison group
- Deworming example
  - Randomize at level of school, not individual, so everyone in treated school can receive medicine
  - Compare average outcomes across T and C schools
  - Measure comparison schools' physical distance from treatment schools
    - Since worms spread through contact with contaminated fecal matter and since open defecation common, schools closer to treated schools should be more likely to experience positive externalities
  - Measure social networks
    - Since intervention randomized, percentage of network that is treated may also be random. Those with more treated networks should also experience more externalities

How do you actually randomize?

How to randomize?

- Randomize participation
  - Units are either in treatment or comparison group
- Randomize order of participation
  - All units eventually treated, but in the interim, later treatment units serve as comparison for early treatment units
- Randomize inducement for participation
  - More on this in later presentations
  - Also called an *encouragement design*

How to randomize?

- But how do we actually do this?

- Many options
  - Flip a coin
  - Public or private lottery (pull names from a jar)
  - Roll dice

How do you actually randomize?

- Software that allows you to generate a random number
  - Faster than above options
  - Can later prove that randomization was legitimate
  - Example: A unit can be in 1 of 4 experimental groups
    - Assign random number to all units
    - First quartile of random number distribution in comparison group, and other quartiles correspond to other 3 experimental groups

Stratification and randomization

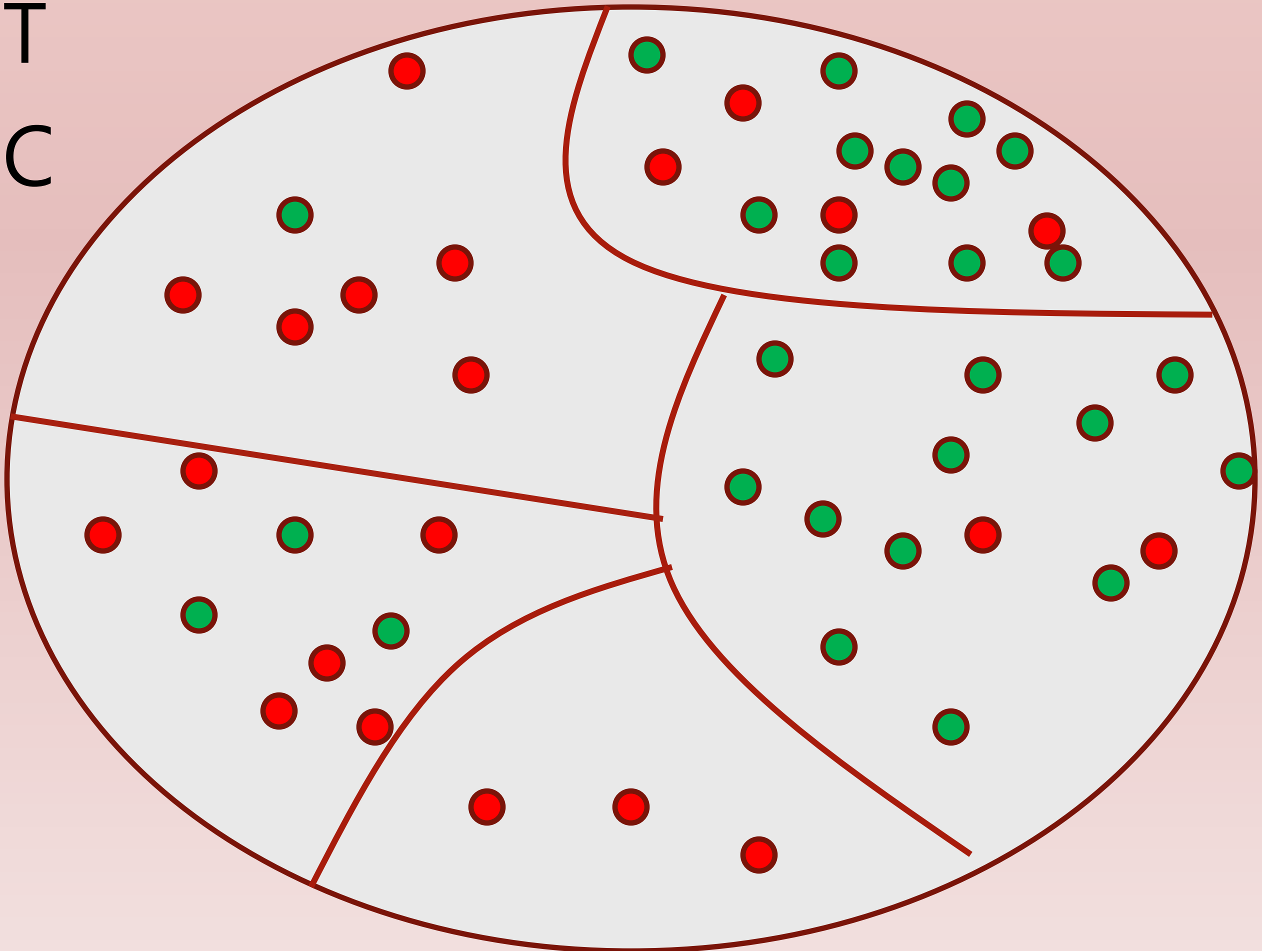
What is stratification?

- Separate units into sub-populations
  - Geographic areas
  - Gender or ethnicity
  - Income level

- <u>Within</u> each strata, randomize treatment
  - Example: Half of women in sample are treated, half are in the comparison

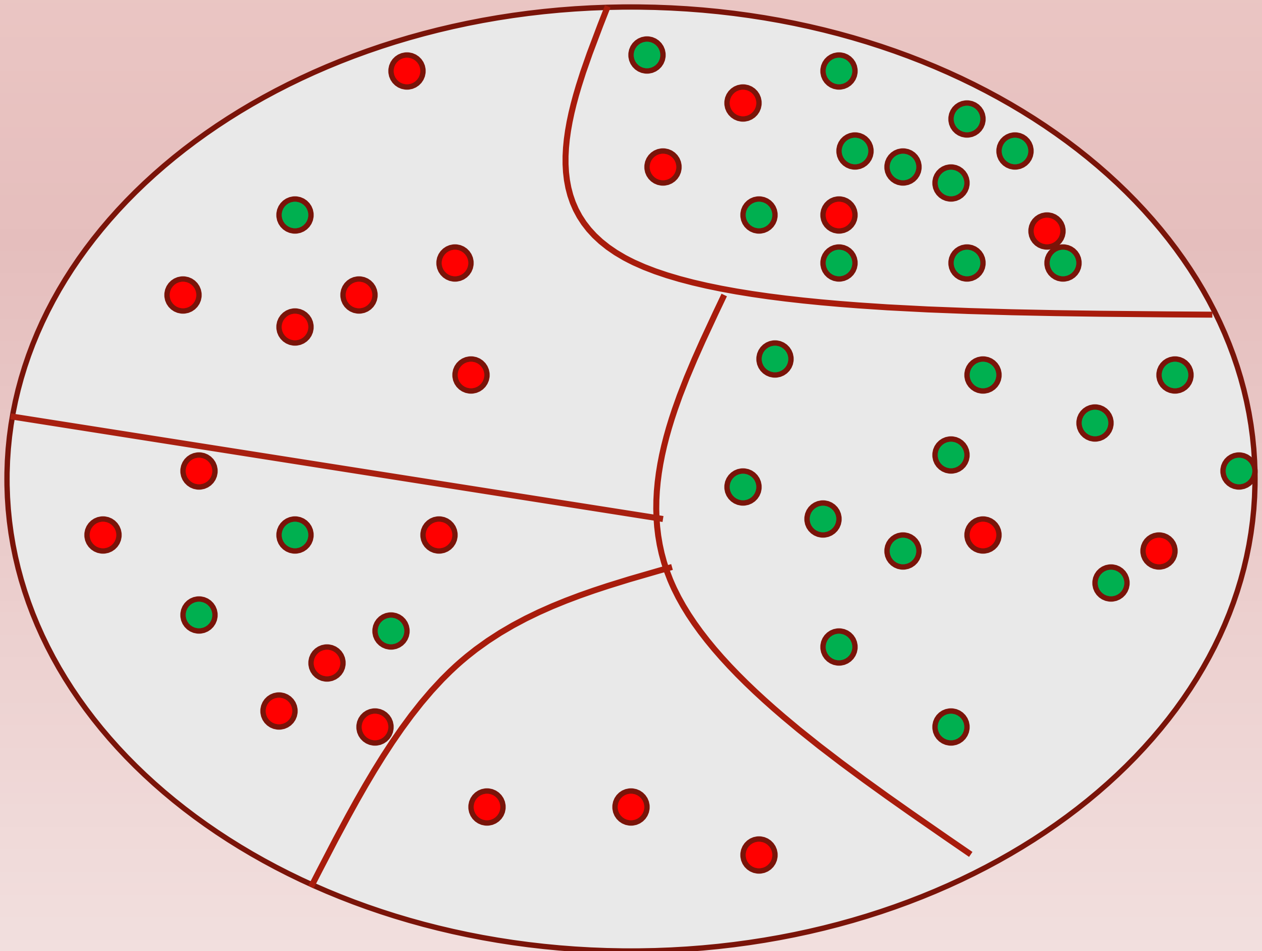
Why do we need strata?

- Geography example
- = T
- = C



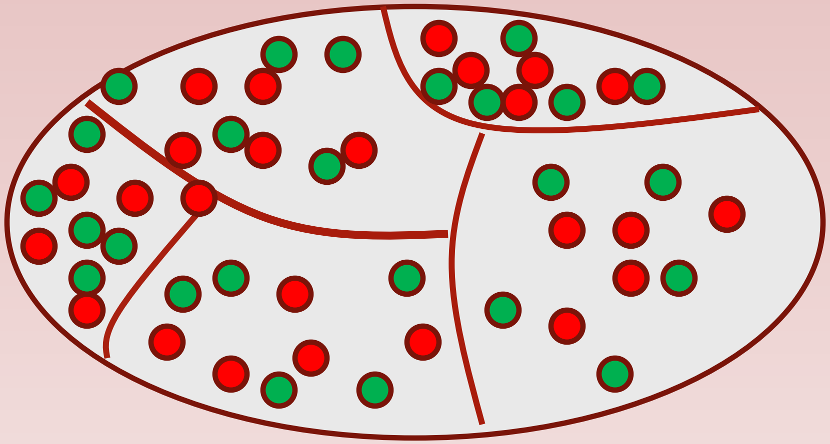
Why do we need strata?

- What's the impact in a particular region?
- Sometimes hard to say with any confidence



Why do we need strata?

- Random assignment to treatment *within* geographical units
  - Within each unit, ½ will be treatment, ½ will be comparison.
- Similar logic for any other sub-population



Why do we need strata?

- Also allows us to *cleanly* measure heterogeneous treatment impacts
  - Separate impacts for each group
- Also guarantees balance of stratified variables between treatment and control and improves power

Random sampling and randomization:

They are not the same, but both are important

Randomization

- Random assignment of units to treatment and comparison groups
- Treatment impact will be unbiased for that sample

Random sampling

- Randomly choosing units from overall study population to observe

- Could occur before or after assignment of treatment
  - Would occur after if intervention is large and we do not need to survey everyone to estimate treatment impact

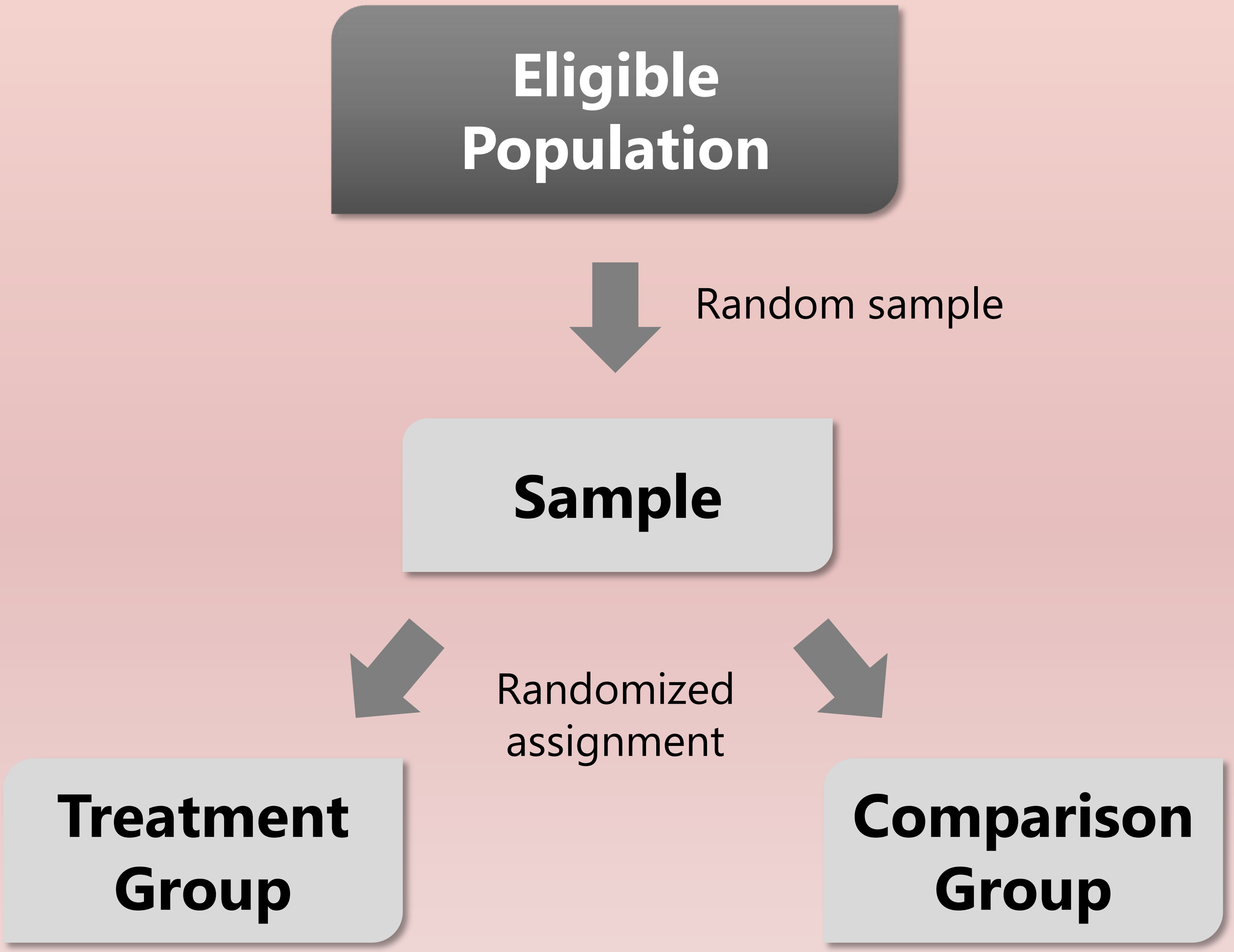
Typical sequencing

First stage

A random sample of units is selected from a defined population.

Second stage

This sample of units is randomly assigned to treatment and comparison groups.


Eligible Population
Random sample
Sample
Randomized assignment
Treatment Group
Comparison Group

Why two stages?

First stage – Random sampling from population

For external validity

Ensures that the results in the sample will represent the results in the population within a defined level of sampling error

Second stage – Randomized assignation of treatment

For internal validity

Ensures that the observed effect on the dependent variable is due to the treatment rather than to other confounding factors

Other issues:

Attrition, compliance, estimated parameters

Attrition

- Drop out from intervention or survey sample

- Why this matters
  - What if only treatment units experiencing high returns remain in intervention?
    - Will over-estimate impact of intervention
  - What if most desperate members of comparison group migrate to another area?
    - Will under-estimate impact of intervention

- Need to be concerned about
  - Differential attrition across T and C groups
  - Differential attrition across types within an experimental group

(Non)compliance

- Some members of treatment group do not take up the treatment
- Some members of comparison group get the treatment

- Could occur through actions of either experimental units or implementers

- Non-compliance usually not random
  - Interferes with causal inference

- Often difficult to avoid
  - Methods to address this if extent of non-compliance is not large (discussed in later presentation)

Estimated parameters

- Still need to think about what these are even when randomizing!

- Randomization can remove selection bias but we can still estimate something that is
  - Irrelevant
  - Different from what we were intending to estimate

Estimated parameters

- Are we measuring partial or total derivative?

- Example 1: School meals offered in randomly selected schools
  - We are interested in impact of school meals on school attendance

  - What if schools offering school meals raise their (effective) prices after they observe everyone wants to go to their school?
    - Can induce some children to drop out of school

  - We will end up measuring the sum of direct impact on attendance and indirect impact on attendance operating through prices (total derivative)

  - But price variation occurs because some schools do not offer meals
    - Would not occur during scale-up
    - Therefore, we might be more interested in partial derivative

Estimated parameters

- Example 2: Mandated provision of health insurance in formal sector
  - We are interested in impact on service utilization

  - Immediate impact
    - Formal sector firms must provide insurance
    - Increase in insurance coverage and utilization
    - Partial derivative

  - Potential impact over time
    - Reform decreases incentive to be a formal firm
    - Decrease in insurance coverage and utilization
    - Total derivative

  - In this case, we might be more interested in the total derivative
    - Should be incorporated into evaluation design
      - Timing of measurement
      - Units to measure (e.g. firms and households)
      - Variables to measure (e.g. formal sector status, insurance offer by firm)

Estimated parameters

- Hawthorne effects
  - Act of observation or demonstrated interest makes units behave differently
  - Treatment impact = true treatment impact + observation effect
  - Experiments on productivity effects of lighting from 1924-1932 at the Hawthorne Works factory
  - Productivity effects disappeared when study concluded even though intervention remained

- John Henry effects
  - Comparison group alters behavior because they know they are in the comparison group
    - May try to compensate (Folklore: John Henry tries to lay railroad faster than a machine)
    - May become disgruntled

- The effects might not occur during scale up
  - Problem if effect observed in pilots results from Hawthorne or John Henry effects rather than treatment

Randomization and non-randomized methods

- Randomization solves selection bias problem
  - All other methods (even quasi-experimental) will always try to *approximate* randomization

- Randomization does not solve every problem
  - Statistical power
  - Attrition and compliance
  - Potential deviation from estimated parameters and parameters of interest

References

- Esther Duflo, Rachel Glennerster, and Michael Kremer (2007), "Using Randomization in Development Economics Research: A Toolkit," in T.Paul Schultz and John Strauss (eds.) Handbook of Development Economics, Vol 4.

- Edward Miguel and Michael Kremer (2004), "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica*, 72(1)

- Michael Kremer and Edward Miguel (2007), "The Illusion of Sustainability," *Quarterly Journal of Econometrics*," 122(3).

- Michael Kremer and Alaka Holla (2009), "Pricing and Access: Lessons from Randomized Evaluations in Education and Health," in Jessica Cohen and William Easterly (eds.) *What Works in Development? Thinking Big and Thinking Small*, Brookings University Press

- See also websites of
  - SIEF [Spanish Impact Evaluation Fund]
  - J-PAL [Abdul Latif Jameel Poverty Action Lab]
  - IPA [Innovations for Poverty Action]