STOR 155.010 Lecture 03
Sampling Strategies & Experiments

Aniruddhan (Ani) Ganesaraman

Fall 2026
UNC Chapel Hill

25 August 2026

Overview

Census

- Wouldn't it be better to just include everyone and "sample" the entire population? This is called a **census**.
- There are problems with taking a census:

Census

- Wouldn't it be better to just include everyone and "sample" the entire population? This is called a **census**.
- There are problems with taking a census:
  1. Difficult to complete a census: individuals who are hard to locate or hard to measure (pattern in missing values).
  2. Populations rarely stand still, so it is never possible to get a perfect measure.
  3. Monetary and physical costs are very high.
- This necessitates **sampling**.

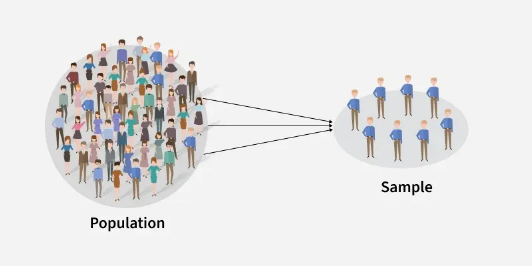
Exploratory analysis to inference

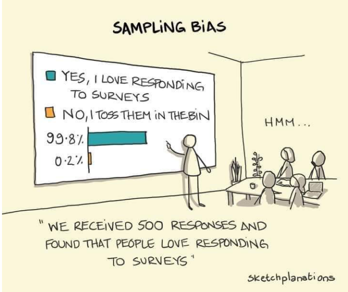
- Sampling is natural.
- Cooking analogy: taste a spoonful to get a sense of the whole dish.
- Tasting the spoonful and noticing it's not salty enough is exploratory analysis.
- Concluding the *entire pot* needs salt is inference.
- For the inference to be valid, the spoonful (sample) must be representative of the pot (population).
    - Tasting only from the surface, while salt has settled at the bottom $\rightarrow$ not representative.
    - Stirring thoroughly before tasting $\rightarrow$ more likely representative.

Common issues with sampling

- **Non-response:** Few people respond $\rightarrow$ sample may no longer represents the population.
  - *Example:* Class survey handed out, but only 1–2 students fill it out.
- **Voluntary response:** Sample self-selects from people with strong opinions.
  - *Example:* Course survey $\rightarrow$ unhappy students more likely to respond.
- **Convenience sample:** Sample favors whoever is easiest to reach.
  - *Example:* Polling people walking by the Pit at noon.

Sampling bias





Observational studies

- Researchers collect data in a way that does not directly interfere with how the data arise (e.g., questionnaire).
- Results of an observational study can generally be used to establish an **association** between the explanatory and response variables.
- They **cannot** be used to establish **causation**.

Obtaining good samples

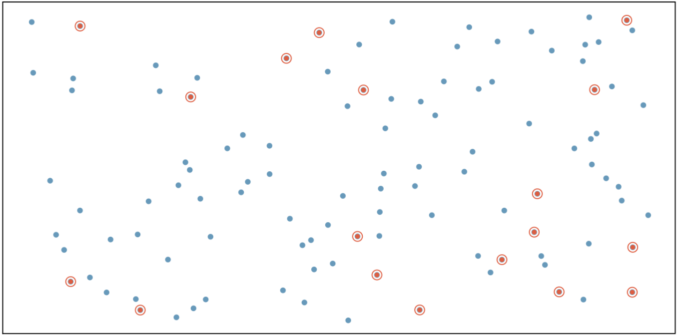
- Almost all statistical methods are based on the notion of implied **randomness**.

- If observational data are not collected in a random framework from a population, these statistical methods (estimates and errors associated with the estimates) are not reliable.

- Most commonly used random sampling techniques are simple, stratified, cluster, and multistage sampling.

Simple random sample

Randomly select cases from the population, where there is no implied connection between the points that are selected.

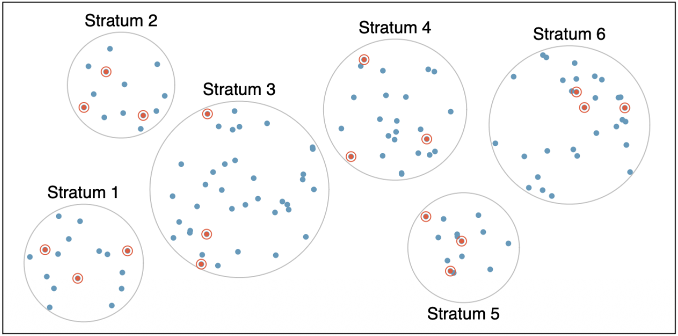


Example

- Objective: study the salaries of MLB players.
- Each player is a member of one of the league's 30 teams.
- Suppose we want to take a random sample of 120 players.
- Execution: write all the players' names on slips of paper, mix them inside a bucket, and select 120 of them.

Stratified sample

Strata are made up of similar observations. We take a simple random sample from each stratum.

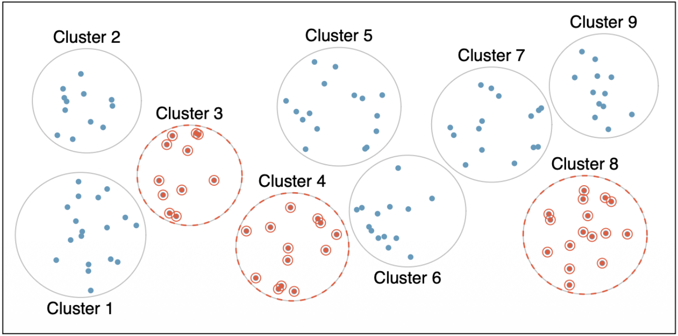


Example

- Back to the MLB players example.
- Goal: sample 120 players.
- Insight: teams might differ a lot in terms of compensation, so it makes sense to use different teams as different strata.
- Stratified sampling: split the players population naturally in 30 teams, and then sample 4 players from each team.

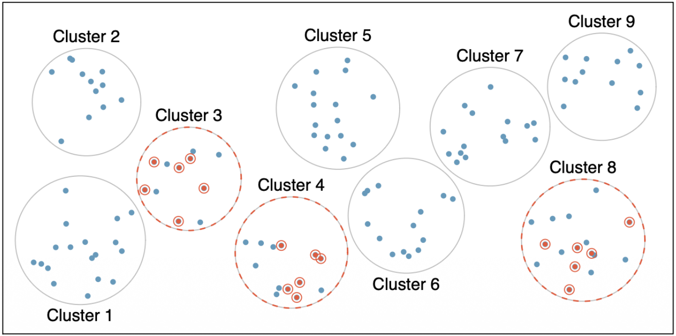
Cluster sample

Clusters are usually not made up of homogeneous observations. We take a simple random sample of clusters, and then sample all observations in that cluster. Usually preferred for economical reasons.



Multistage sample

Like cluster sampling, but we sample within each selected cluster rather than taking everyone.



Overview

Experiments

- Researchers assign treatments to cases: called an **experiment**.
- Assignment involves randomness (e.g., a coin flip decides who gets what): a **randomized experiment**.
- Randomized experiments are the key tool for showing causation.

Principles of experimental design

Four building blocks of a randomized experiment:

1. **Control:** Control for the (potential) effect of variables other than the ones directly being studied.
2. **Randomize:** randomly assign subjects to treatments and randomly sample from population.
3. **Replicate:** use a large enough sample, or repeat the study.
4. **Block:** group subjects by a known factor (suspected to affect the response variable) first, then randomize within each group.

Controlling

- Keep everything the same between groups, except the treatment itself.
- *CFS trial:* both the CBT group and the relaxation group get the same number of sessions, same length, same setting. Only the therapy type differs.

Randomization

- Randomly assigning patients evens out differences we can't directly control. It also prevents accidental bias from entering the study.
- *Example*: Some patients may be more susceptible to a disease than others due to their dietary habits. Randomizing patients into the treatment or control group helps even out such differences, and it also prevents accidental bias from entering the study.

Replication

- More cases $\implies$ a more reliable estimate of the treatment's effect.

Replication

- More cases $\implies$ a more reliable estimate of the treatment's effect.

- *Within a study:* use a large sample.

- *Across studies:* other researchers repeat the CFS trial to confirm the result.

Blocking

- Sometimes a variable, other than the treatment, affects the outcome. Called a **blocking variable**.

Blocking

- Sometimes a variable, other than the treatment, affects the outcome. Called a **blocking variable**.

- Group subjects into **blocks** by this variable first, *then* randomize within each block.

Blocking

- Sometimes a variable, other than the treatment, affects the outcome. Called a **blocking variable**.
- Group subjects into **blocks** by this variable first, *then* randomize within each block.
- *CFS trial:* split patients into mild fatigue and severe fatigue blocks. Randomize each block separately into CBT vs. relaxation.

Blocking - Example

- To design an experiment to investigate if energy gels makes you run faster:
  - Treatment: energy gel
  - Control: no energy gel
- It is suspected that energy gels might affect pro and amateur athletes differently. So we block for pro status:
  - Split sample: pro and amateur
  - Randomly assign pro athletes to treatment and control groups
  - Randomly assign amateur athletes to treatment and control groups
  - Pro/amateur status is equally represented in the resulting treatment and control groups

Difference between blocking and explanatory variables

- **Factors** are conditions we can impose on the experimental units.
- **Blocking variables** are characteristics that the experimental units come with, that we would like to control for.
- Blocking is like stratifying, except used in experimental settings when **randomly assigning**, as opposed to when **randomly sampling**.

Experimental Design terminology

- Placebo: fake treatment, often the control in medical studies.
- Placebo effect: patients improve just from believing they are being treated.
- Blinding: patients don't know if they are in treatment or control.
- Double-blind: neither patients nor researchers know who is in which group.

References

- Sections 1.3 and 1.4 of the textbook
- *6 historic experiments that helped pave the way for modern clinical research*
- *The Fascinating and Extensive History of Clinical Trials*
- *Famous Statistical Blunders in History*
- *Google's 41 shades of blue*