Experimental Design

#CMSC320 #M1

Origins of Data

Observational Studies

Observe a sample of a population without influencing the response of participants (i.e. no treatment applied).

- **Cross sectional:** Looks at data from a single point in time (*Present*)
- **Retrospective studies:** Looks at studies of events in the past (*Past*)
- **Prospective studies:** Researchers follow and observe groups closely (*Future*)

Experiments

Observe effects of treatment after application on subjects.

Requirements for an Experiment

- *Randomly selected* subjects
- Subjects are *representative of the population* being tested on
- Experiment can be *replicated*
- *Controls* for effects of variables
  - (more about controlling experiments in Experimental Design > Collecting Data > Controlling Data)

Synthetic

Data created by experimenter, typically through simulation.

Collecting Data

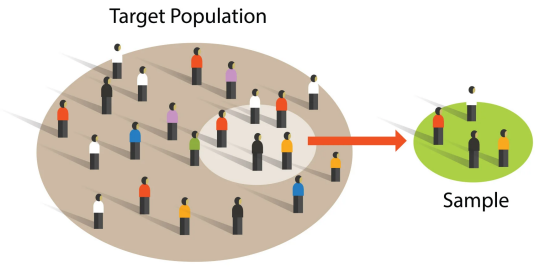
> **Note**
>
> Data must be representative of the population with regards to the question(s) of interest.

Controlling Data

- Ways to control an experiment:
  - **Blinding**: participants are unaware of the kind of treatment they are recieving, if any at all
    - **Double dummy**: a method of blinding where both treatment groups may receive *placebo*
  - **Placebo**: something that appears to the participants to be an active treatment, but does not actually contain the active treatment
  - **Blocking**: arranging experimental units into similar groups based on treatment applied.
    - (See difference between blocking and stratification Experimental Design > Collecting Data > Sampling Techniques > Stratified Sampling)

Sampling Techniques

Well designed sampling incorporates several of the following types of sampling.

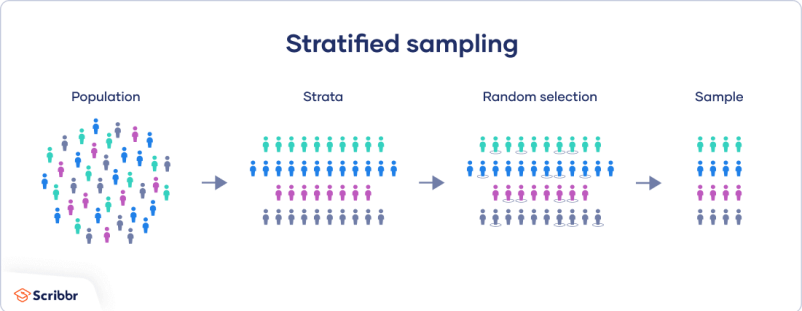


Systematic Sampling

- A probability sampling method where researchers select members of the population at a regular interval.
  - Example: selecting every 15th person on a list of the population.

Stratified Sampling

- In a stratified sample, researchers divide a population into homogeneous subpopulations, called strata, based on specific characteristics (ex. race, gender identity, location, etc.). Every member of the population studied should be in exactly one stratum.

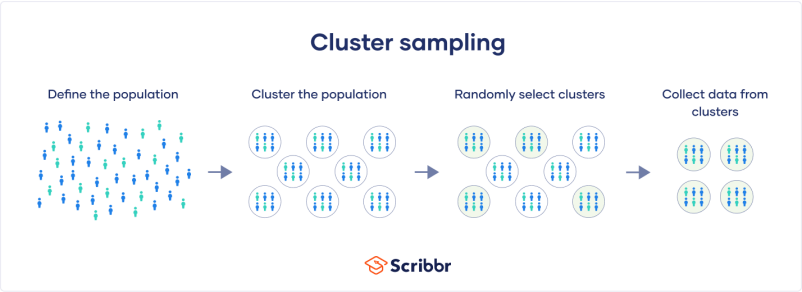


Stratification vs. Blocking

- Stratification groups subjects based on characteristics which the experimenter cannot control (ex. eye color). Blocking groups subjects based on variables the experimenter can control such as the treatments.

Cluster Sampling

- In cluster sampling, researchers randomly divide a population into smaller groups known as clusters. They then randomly select among these clusters to form a sample.
- Cluster sampling is a method of probability sampling that is often used to study large populations, particularly those that are widely geographically dispersed.



Cluster vs Stratification

- In clustering subjects are grouped randomly, while in stratification they are grouped based on shared characteristics.

Multistage Sampling

- In multistage sampling you draw a sample from a population using smaller and smaller groups at each stage.

Convenience Sampling

- Convenience sampling is a method of collecting samples by taking samples that are conveniently located around a location or Internet service. Be careful of using this sampling technique, can introduce a lot of bias.

Error

Types of Error

- **Sampling Error**
  - Unrepresentative sample taken
- **Non-Sampling Error**
  - Errors due to sample data that are incorrectly collected, recorded, or analyzed

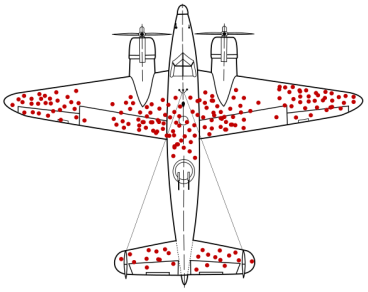
Error in Surveys

- Wording of questions
- Ordering of questions (planting ideas)
- Convenience samples
- Desire of respondents to please
- Non-response bias
- Lizardman constant (around 3% of respondents are just messing around)

Other Concepts

- **Confounding Variable**

    - Is one that affects the response variable and is related to the explanatory variable.
        - Example: People given leeches produce magical tears that heal wounds, the tears would be a confounding variable in an experiment testing leeches' effects on wound healing
- **Survivorship Bias**
    - Survivorship bias or survival bias is the logical error of concentrating on entities that passed a selection process while overlooking those that did not.
        - Example: Cannot test on fatal shots to planes



- Once a rigorous expirement is designed and conducted correctly, an experimenter must accept the results even if they go against their expectations.

Extra Reading

1. Neo, B. (2020, October 27). *Experimental Design in data science*. Medium. Retrieved March 20, 2023, from https://towardsdatascience.com/designing-experiments-in-data-science-23360d2ddf84
2. Wood, R. (2020, January 4). *The three keys to experimental design every data scientist should know*. Medium. Retrieved March 20, 2023, from https://towardsdatascience.com/the-three-keys-to-experimental-design-every-data-scientist-should-know-b0d812d86865