STAT 100: Week 4

Ricky's Section

Introductions and Attendance

Introduction: Name

Question of the Week: What is your screen time? This is related to statistics, I promise.

Important Reminders

Anonymous Feedback

https://docs.google.com/forms/d/e/1FAIpQLSfKv_FGvs0oqm-IvtxKx3Vf6bBzSJE2jamK1gklAzL6NkXE8w/viewform

Labs

- Labs are designed to provide practice, so there are a lot of problems
- Similar to worksheets in MATH 1/MATH 21
- Don’t worry—you’re NOT behind if you don’t finish the lab!

One-on-One Office Hours

- Make sure to utilize these resources if you'd like more one-on-one time
- Conceptual help
- Study tips/strategies

Note Taking

- My suggestion: Annotate the slideshow during lecture/section
- Afterwards, update your Google Doc with the important stuff (code, definitions, images)

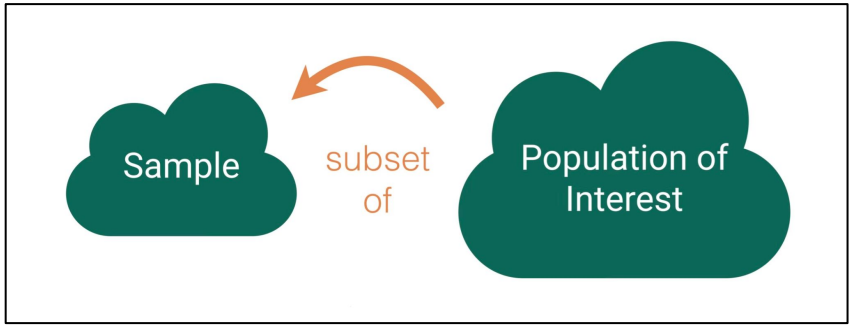
Data Joins: Please Don't Be Scared!

- Data joins are very small in the grand scheme of things, so don't worry!
- P-sets and exams are open-book, so what matters most is having good notes
- Notes should explain the process in a way that makes sense to you

Content Review: Week 4

What Is Sampling?

- **Sample**: Subset of **population of interest**, whatever that may be (ideally, it's **representative** of the population)
- **Census**: When there is data for whole population (**everyone is represented**)
  - Often, it's hard to get a census



What Is Bias?

- **Sampling bias**: When **sampled** units are different from **non-sampled** units on the **variable(s) of interest**
  - *Ex: If I ask Harvard students for their screen time via Instagram poll, those who are sampled probably have higher screen times*
- **Nonresponse bias**: When **respondents** are different from the **non-respondents** on the **variable(s) of interest**
  - *Ex: If I ask Harvard students for their screen time, those with higher screen times may be embarrassed and decline to answer*

Observational Study vs. Experiment

- **Experiment**: Researchers directly influence how data arises
  - Causal relationship can be established
- **Observational study**: Researchers only observe and record data without interfering
  - “Correlation does not mean causation”

Principles of an Experiment

- **Control group**: Group of subjects who get **no treatment**
- **Experimental group**: Group that does get **treatment**
- **Random assignment**: Subjects are randomly assigned to either the **control group** or the **experimental group**
- **Confounding variable**: Third variable that is associated with both the **explanatory variable** and **response variable** *(e.g., genetics on sunscreen use and skin cancer)*

Principles of an Experiment

- **Placebo**: Fake treatment to control for **placebo effect**
  - *If given a sugar pill (placebo), someone may start to feel better because they believe it is medicine*
- **Blinding**: When **subjects** don't know the **group assignments** (**control** vs. **experimental**)
  - *If given a pill, the subject wouldn't know whether it's medicine or sugar/placebo*
- **Double blinding**: When **both subjects and researchers** don't know (not always possible)
  - *All the pills are mixed, so researchers can't tell whether they're giving out medicine or sugar/placebo*

Why can experiments establish causal relationships?

Question:

Why can experiments establish causal relationships?

Due to **random assignment**, those in **control group** should be very similar to those in **experimental group**. Thus, **confounding variables** have been eliminated/minimized.

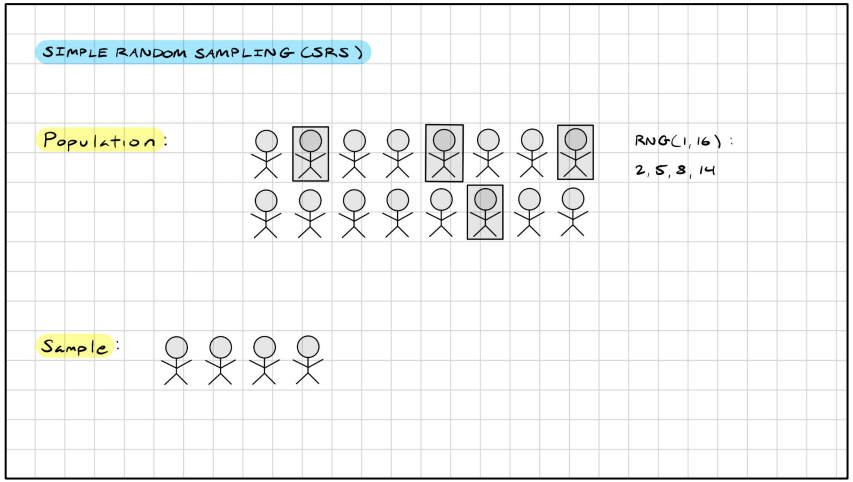
The differences between the two groups after the **experiment** must have been caused by the **treatment/explanatory variable**.

Four Sampling Methods

- There are four main methods for **random** sampling:
  - **Simple random sampling**
  - **Systematic sampling**
  - **Cluster sampling**
  - **Stratified sampling**

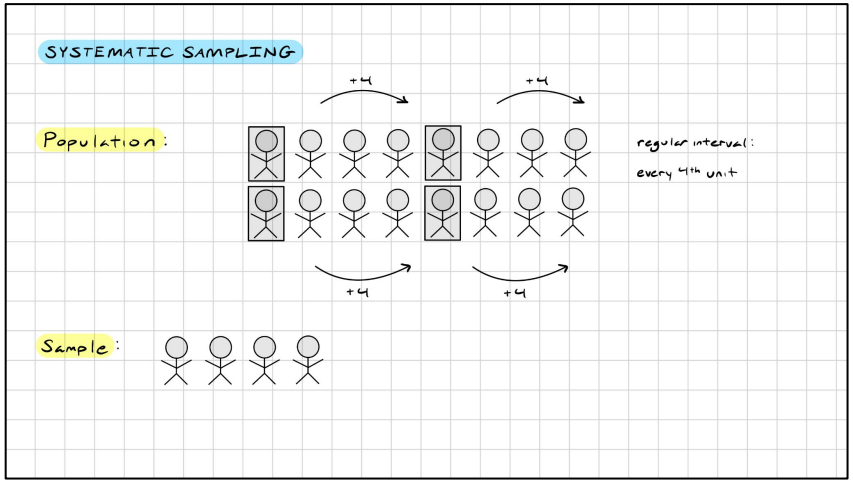
Simple Random Sampling (SRS)

- **Simple random sampling**: Every unit has an equal chance of being selected via **random mechanism** (all units must be listed out in a **sampling frame**)
  - *Ex: To determine smartphone usage within Harvard students, number every student (HUID) and then draw random numbers to determine which ones to sample*



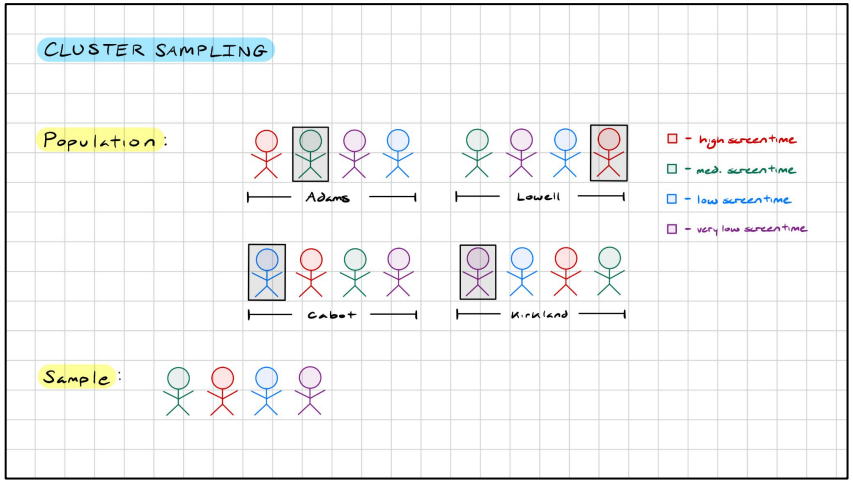
Systematic Sampling

- **Systematic sampling**: Starting point is randomly chosen, and then units are sampled at a **regular interval**
  - *Ex: To determine smartphone usage within Harvard students, number every student (HUID) and then sample every fourth student*



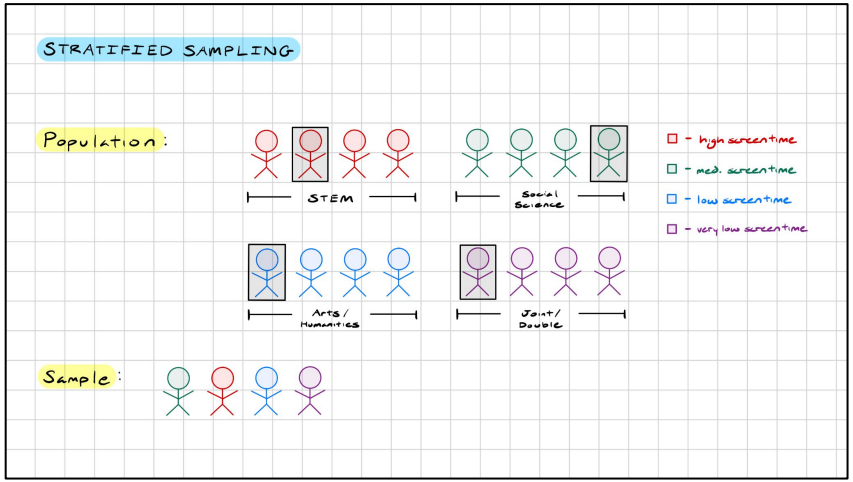
Cluster Sampling

- **Cluster sampling**: Divide **population** into **homogeneous groups/clusters** take a **random sample** within **SOME** of the **clusters** (to be chosen randomly)
  - *Ex: To determine smartphone usage within Harvard students, sample students within four randomly-selected houses*
  - *Here, houses should be homogeneous (in terms of screen time) because houses are randomly assigned*



Stratified Random Sampling

- **Stratified random sampling**: Divide **population** into **heterogeneous groups/strata** and take a **random sample** within **EVERY stratum**
  - *Ex: To determine smartphone usage within Harvard students, sample students within each concentration*
  - *Here, concentrations should be heterogeneous (in terms of screen time) because STEM fields require more technology*



Intuitively, why do we NOT need to sample every cluster?

Question:

Intuitively, why do we NOT need to sample every cluster?

Clusters are relatively **homogeneous** in terms of our **variable**. For example, houses are similar to each other in terms of screen time. Thus, we don't need to sample Leverett if we already sampled Cabot, Adams, and Pfoho.

Conversely, **strata** are defined to be relatively **heterogeneous**, so all groups must be accounted for.

Questions?

P-Set 3

Have a great rest of your week!