STAT 102: 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

Fun Stats Stuff!

- Spring Social (S^2): Fri, 02/21, 3:30-4:30 PM in SC 316
  - https://tinyurl.com/gushspringsocial25
- Stat Undergrad Research Forum (SURF): Fri, 02/28, 3:30-4:30 PM in SC 316
  - https://tinyurl.com/gushsurf2025
- Women in Data Science Panel (WinDS): Fri, 03/7, 7-8 PM in SC 316
  - https://tinyurl.com/womeninds25

My Office Hours

- Normally on Friday, 3-5 PM in the Inn (Adams D-Hall)
- Switched to Saturday, 1-3 PM for the next two weeks (02/22 and 03/01)
- Come if you have any questions!
- https://docs.google.com/spreadsheets/d/1AgnpomB7qUGtyRTI8Ansothj-2LbZnKJYO7q0br6c18/edit?usp=sharing

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)

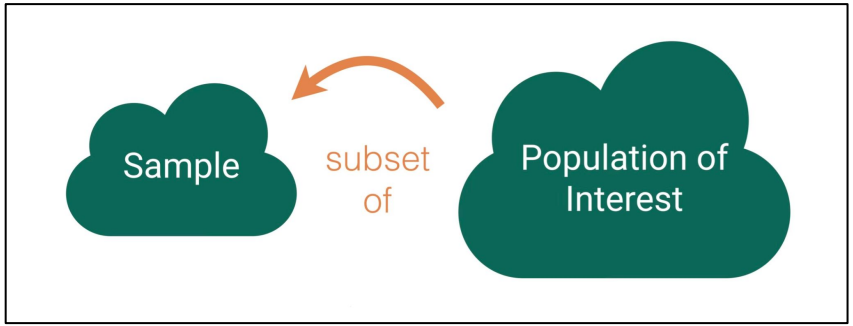
Data Joins: Please Don't Be Scared!

- 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 the data arise
  - Causal relationship can be established with random assignment
- **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.

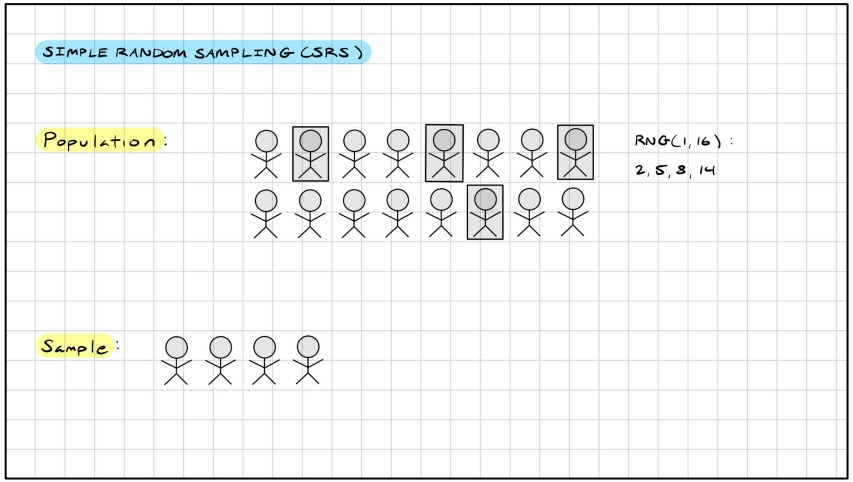
The differences between the two groups after the **experiment** must have been caused by the **treatment/explanatory variable**.

Four Sampling Methods

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

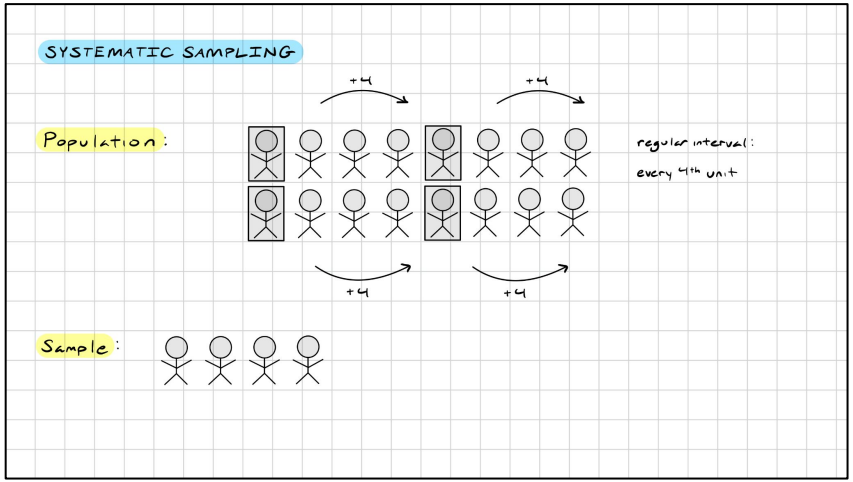
Simple Random Sampling (SRS)

- **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*



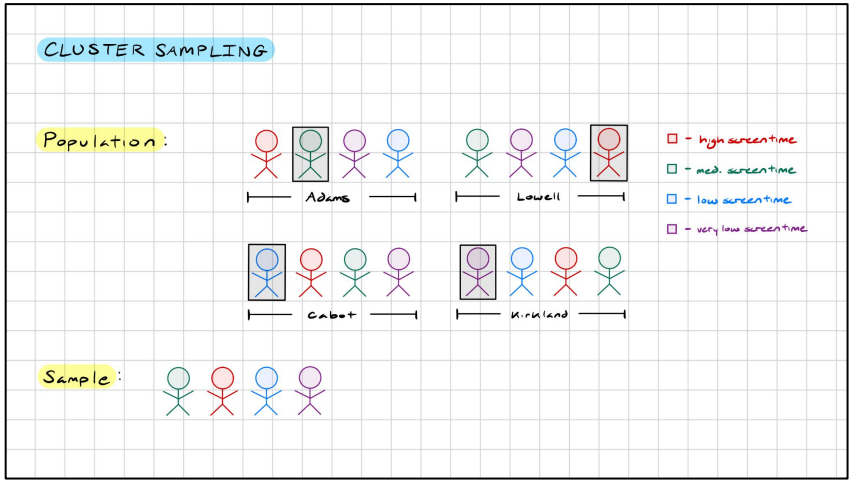
Systematic Sampling

- **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*



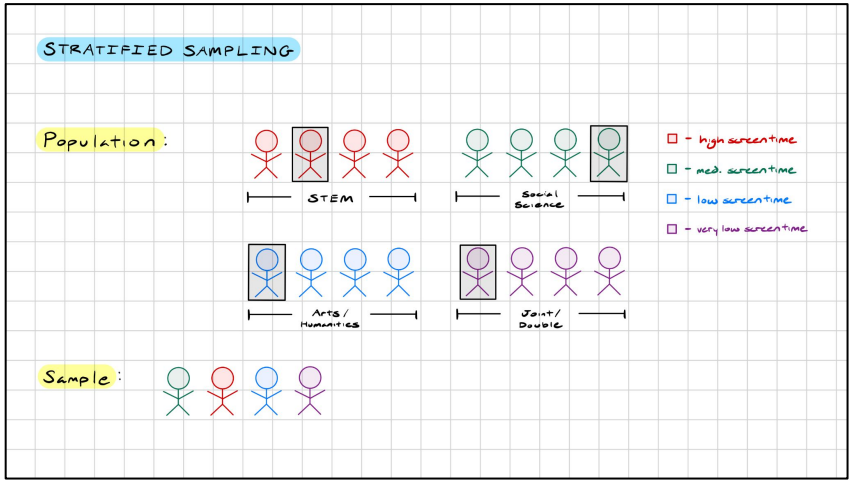
Cluster Sampling

- **Cluster sampling**: Divide **population** into **homogeneous groups/clusters** and 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*



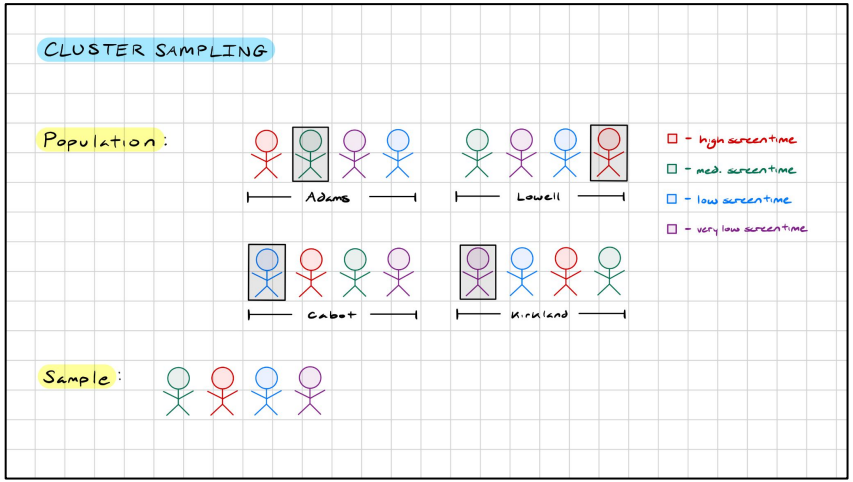
Stratified Random Sampling

- **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*



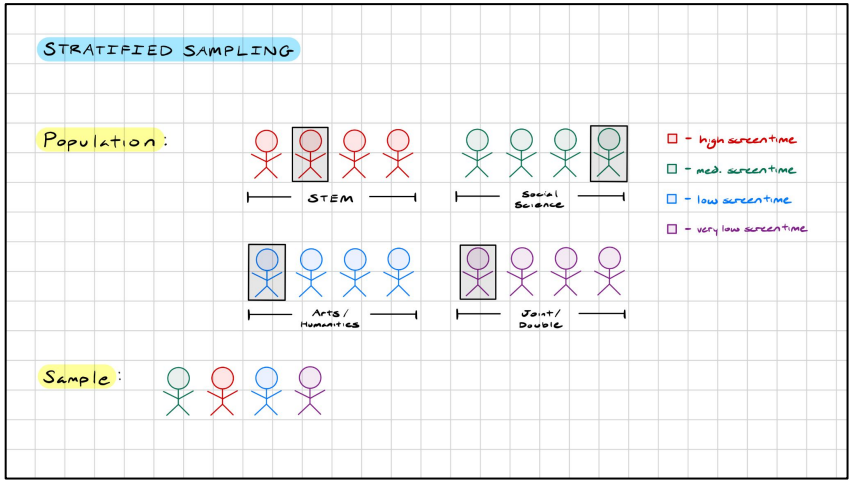
A Clarifying Note...

- "Hetero-" means different, "homo-" means same
- When we say **homogeneous groups**, we mean the GROUPS are homogeneous/similar to EACH OTHER
  - Notice how the PEOPLE within the groups are pretty different from each other, but that's not what we are referring to!



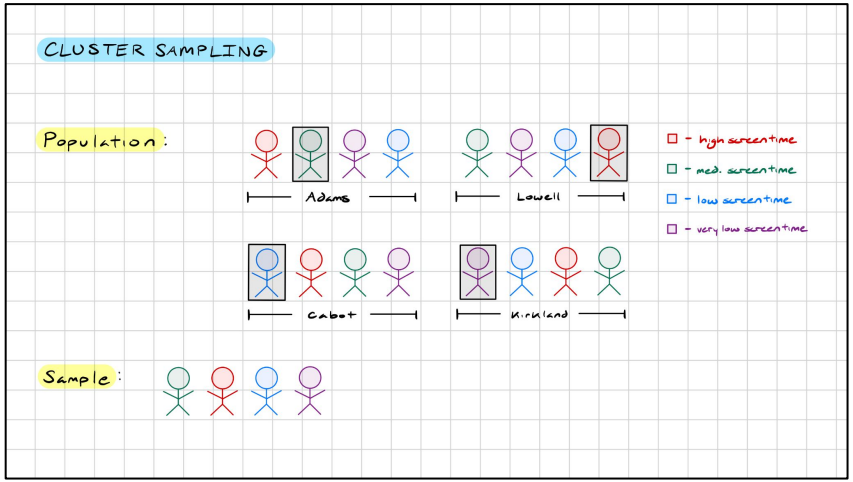
A Clarifying Note...

- Similarly, these are **heterogeneous groups** as in each GROUP is different from the others (in terms of our variable of screen time)
  - For example, the STEM group is primarily high screen time while the arts/humanities is primarily low screen time



And One Last Thing...

- “**Homogeneous in terms of our variable**” is context dependent
- Instead of screen time, let’s say our variable of interest is “Hours Spent at the MAC per Week”
  - Before, houses were homogeneous groups (in terms of screen time), so we can appropriately treat them as clusters
  - Now, they aren’t. Why?



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.

We divide the population of Harvard students based on their home country and (randomly) sample 10 countries. Am I treating countries as clusters or as strata?

Question:

We divide the population of Harvard students based on their home country and (randomly) sample 10 countries. Am I treating countries as clusters or as strata?

Clusters!

This is because we only sample SOME of the countries (if we treated countries as strata, we would need to sample ALL of them).

Treating countries (or any group) as clusters when it's not appropriate to do so results in non-representative data.

Questions?

P-Set 3

Have a great rest of your week!