Day 16 of #100daysofmathandstats: Data sampling Concepts

By Harsh Kathiriya

Outline

- Population vs Sample
- Observational study vs experimental study
- Sampling methods

Population and sample

- Question to research on:
  - **Are consumers of certain alcohol brands more likely to end up in the emergency room with injuries**?
- If we want to research on this question, then there are two ways in which research could happen
  1. Take the whole data (aka population)
  2. Take the subset of the data (aka sample)

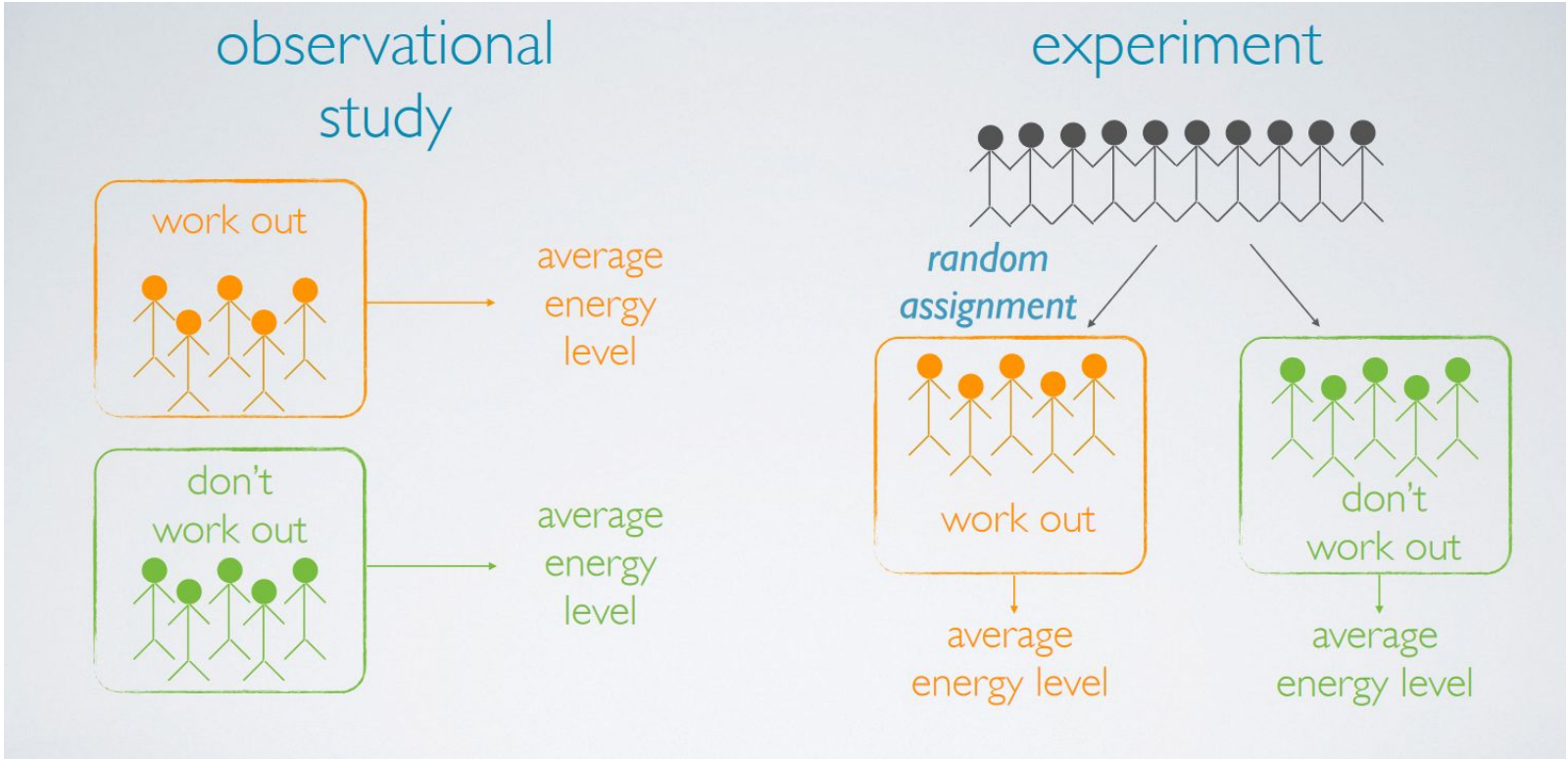
Population and sample (Contd…)

- **Population:**
  - Take everyone whoever took alcohol
- **Sample:**
  - ER patients at the local hospital at Atlanta in the US
- If someone choose to take sample and do research then it’s more likely to generalize to the residents of Atlanta since the samples of that would be easily accessible to research.

Observation study vs Experimental study

- Observational study:
  - collect data in a way that does not directly interfere with how the data arise (“observe”)
  - only establish an **association**
  - It uses **past data and data collected during study**
- Experimental:
  - randomly assign subjects to treatments
  - establish **causal connections**
  - It uses only **experimental data**

Observation study vs Experimental study (Contd…)



Sampling methods

1. Simple random sampling
2. Stratified sampling
3. Cluster sampling
4. Multistage sampling

Simple random sampling

- Random sampling is a process in which **each available member of the population being sampled** has an equal chance of being chosen for the sample at each draw.
- The sample that results is called a **simple random sample.**
- Simple random sampling is probably the most intuitive form of random sampling.

Example of Simple random sampling

- Consider the salaries of Major League Baseball (MLB) players, where each player is a member of one of the league's 30 teams.
- To take a simple random sample of 120 baseball players and their salaries
    - we could write the names of that season's several hundreds of players onto slips of paper, drop the slips into a bucket, shake the bucket around until we are sure the names are all mixed up, then draw out slips until we have the sample of 120 players

Stratified Sampling

- Stratified sampling is a **divide-and-conquer sampling** strategy.
- The population is divided into **groups called strata.**
- The strata are chosen so that **similar cases are grouped** together, then a second sampling method, usually simple random sampling, is employed within each stratum
- Stratified sampling is especially useful when the cases in **each stratum are very similar** with respect to the outcome of interest

Example of Stratified Sampling

- In the baseball salary example, the teams could represent the strata, since some teams have a lot more money (up to 4 times as much!).
- Then we might randomly sample 4 players from each team for a total of 120 players.
- The downside is that analyzing data from a stratified sample is a more complex task than analyzing data from a simple random sample.

Cluster Sampling and Multistage clustering

- Cluster sampling:
  - In a cluster sample, we **break up the population into many groups,** called clusters. Then we sample a fixed number of clusters and include all observations from each of those clusters in the sample.
- Multistage sampling:
  - A multistage sample is like a cluster sample, but rather than keeping all observations in each cluster, we collect **a random sample within each selected cluster.**

Cluster Sampling and Multistage clustering (Contd…)

- Sometimes cluster or multistage sampling can be **more economical** than the alternative sampling techniques.
- Also, unlike stratified sampling, these approaches are most helpful when there is a lot of **case-to-case variability within a cluster** but the clusters themselves don't look very different from one another.
- A downside of these methods is that **more advanced techniques** are typically required to analyze

Thank you

Github Link: https://github.com/harsh9898/100daysofstatandmath

Don’t forget to post your queries or feedbacks on the post.

Share or like for the benefit of others.

Source for some of the content: https://www.openintro.org/book/os/