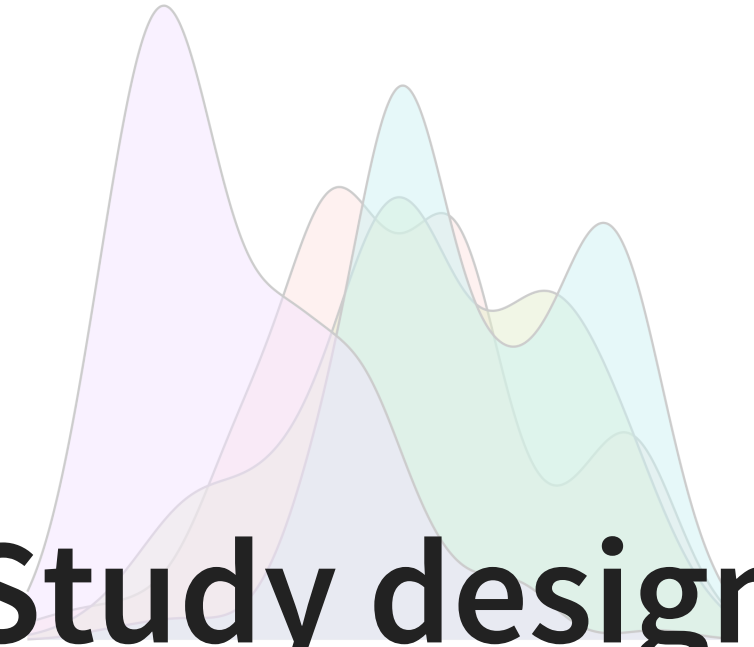
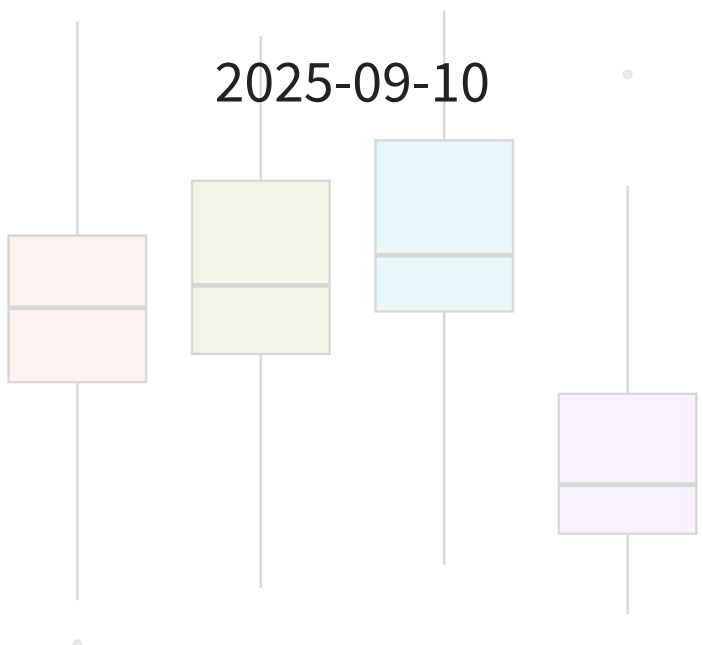


Study design

2025-09-10



Housekeeping

- Please bring your laptops tomorrow installed with R and RStudio!
  - If you need assistance, there are installation help hours from 7-9pm tonight in the Q-Center
- TA hours start this week!
- Optional reading: Chapter 2 Study Design from OpenIntro textbook

Sampling from a population

Variables in statistics

What is a variable?

- Lots of research questions revolve around asking how variable x affects variable y
- **Response variable** y: the primary variable of interest, i.e. the variable whose behavior we want to understand
- **Explanatory variable** x: the variable(s) that (might) affect the response variable
  - In scientific studies, explanatory variables can often be controlled/observed by the researcher ahead of time

A “good” sample

- The way we sample data from a population can directly influence the quality of that sample.
- What are desirable characteristics of a sample?
- **Representative**: the sample roughly “looks like” the target population
  - i.e. the characteristics of participants in the sample are similar to those of the population
- If we have a representative sample *and* if our analysis is sound, then our results should be **generalizable** to the target population
  - i.e. we can use results from a sample to draw conclusions about a specific population

Bias in a sample

- **Biased** samples occur when the methods used to obtain data result in inaccurate depictions of the population. This is bad!!
  - Can occur if a sample is not representative
- Bias in a sample can arise due to many causes. Here are just a few:
  - **Selection bias**: systematic tendency in procedure that causes some members of population to be more likely to be included than others
  - **Non-response bias**: the values of the response variable of non-respondents differ systematically from those that do respond
  - **Response bias**: systematic favoring of certain response variable values that occurs when people don't answer truthfully (e.g. lying)
- Any type of bias *could* lead to our sample being non-representative (and therefore, non-generalizable findings)

Example: *Literary Digest* poll

- In 1936, Franklin D. Roosevelt (Democrat) was running for re-election as president again Alfred Landon (Republican)
- *Literary Digest* magazine conducted a survey which received 2.4 million respondents (largest number of people every replying to a poll at that time)
  - Prediction: overwhelming victory for Landon (predicted FDR would only get **43%** of popular vote)
- Sampling methods: *Literary Digest* mailed questionnaires to 10 million people. Addresses were obtained from telephone books and club membership lists. *Note: only ~25% of households had telephones.*
- Do you think the sample was representative? If not, what sorts of biases might have occurred and why? Based on your answer, do you believe the results from the survey were generalizable to whole population?

Example: *Literary Digest* poll (cont.)

- Selection bias?
  - Only wealthy households were included in the survey.
  - This wouldn't necessarily be bias, EXCEPT for the fact that poor people overwhelmingly favored FDR and the rich favored Landon –> yes selection bias
- Non-response bias?
  - Note that ~75% of those who received a survey did not respond! Wouldn't matter if there wasn't a difference in the opinions of respondents vs non-respondents. But among those who responded, over half favored Landon
  - If the majority of non-respondents also favored Landon, then no non-response bias. Otherwise, yes non-response bias.
- Response bias?
  - No good reason to lie -> no response bias
- Actual result: FDR won by a landslide! (**62%** to 38%)

Sampling methods

Convenience sampling

The worst kind of sampling (but often the easiest)!

- **Convenience sampling** takes place when cases that are easily accessible are more likely to be included in the sample
- Example:
  - Population: students enrolled in statistics courses at Middlebury
  - Sample: students in STAT 201 AZ
- Issue: typically does not yield a representative sample

Probability sampling

- Sampling methods that include a bit of randomness can help reduce the chance of bias
- **Probability/random sampling**: any sampling method where the selection from the population is based on random selection/chance
  - No one has full discretion about who is included in the sample
- **Random sampling *usually* (but not always) yields a representative sample**
- Examples included here: simple random, stratified, cluster

1. Simple random sampling (SRS)

- In a **simple random sample**, each case is chosen entirely by chance from the population, and each member of the population has an equal chance of being sampled
  - Knowing that an individual was sampled does not provide useful information about which other cases are included
  - Any given fixed-size subset of the population is equally likely to be chosen
- Consider again the research question: What proportion of current Middlebury professors attended a liberal arts college?

  How might I obtain a simple random sample of 25 professors?

2. Stratified sampling

- Assume that the population is/can be broken up into several different, distinct sub-populations or **strata**
  - Cases grouped into a strata should be similar to each other
- Then take a (simple) random sample from ***each*** stratum ("divide and conquer")
  - How many from each stratum? Typically use a sampling fraction that is proportional to entire population!
  - E.g. if population of trees on Middlebury campus are 80% deciduous and 20% coniferous and we want to sample $n = 10$ trees total, we should *randomly* sample ___ deciduous and ___ coniferous trees
- What are some pros/cons?

3. Cluster sampling

- Divide total population into M distinct groups or **clusters** of roughly equal size
- Perform a (simple) random sample on the M clusters, than sample all individuals within each of the randomly selected clusters
- Discuss the following:
  - Would you prefer the individuals within a cluster to be homogeneous (similar) or heterogeneous (varied)? Why?
  - Would you prefer that cluster A and cluster B be relatively similar or different in terms of their sub-populations?
  - What is the difference between stratified and cluster sampling?

Types of studies

We now know *how* to collect data, but now we turn to examining what *kind* of study we'd like to perform in order to answer the research question.

Experiments vs Observational studies

- **Observational studies** occur when a research *observes* cases without manipulating any variables
- **Experiments** are studies where the researcher *assigns* specific treatments to cases
  - Experiments are often conducted in medical settings, hence the word "treatment"
- Example: I want to design a study to learn if students who take quizzes throughout the semester end up performing better on the final exam.
  - Observational study: students optionally take quizzes
  - Experiment: I choose half of the students to take quizzes and the other half to not take quizzes.
- Are treatments in experiments considered explanatory or response variables?

Experiments: Treatment vs control

- Treatments are typically divided into two categories:
  1. **Control group**: establishes a baseline, and typically receives "zero amount" of the explanatory variable
  2. **Treatment group(s):** receive some "non-zero amount" of the explanatory variable
- Quiz example continued:
  - Control group (i.e. treatment group 0): no quizzes
  - Treatment group 1: takes one quiz
  - Treatment group 2: takes two quizzes
  - How to decide which case gets which treatment?
- Question of ethics!

Randomized experiments

- **Randomized experiment:** researcher *randomly assigns the treatments*
- **Note**: randomized experiment ≠ random sampling
- Continuing example:
  - Randomized experiment is achieved if I use SRS to determine who received which treatment
  - But the students in the experiment were not obtained via SRS
- Why care randomized experiments important? Because of confounders!

Confounding variables

- Oftentimes, we design an experiment to identify a *causal* relationship: does changing x cause a change in y?
- Assessing causality is made difficult by **confounding variables**: other variables that are associated with both x and y
  - Confounders are bad!! Why?
- Example: consider a study that seeks to examine the effect of coffee consumption on heart disease.
  - From each person, we only collect information on the average amount of coffee they consume per day and whether or not they have heart disease.
  - We find a positive association: more coffee $\longrightarrow$ higher risk of heart disease
  - Possible confounder: smoker status. Smokers tend to drink more coffee and tend to have higher rates of heart disease than non-smokers.
  - So the increase in heart disease may be due to smoker status rather than caffeine intake

Principles of experimental design

The following principles help design a generalizable experiment:

1. **Randomized experiment**
    - Helps account for possible confounders! Why?
2. **Control for differences**: ensure that everyone follows the same protocol exactly
3. **Replication**: collect a sufficiently large sample
    - The more cases we observe, the more confidence we have in our findings

Reducing bias in experiments

- Biases can still unintentionally arise in experiments, even if we follow these three principles.
- **Blind experiment** by not allowing participants to know which treatment they are receiving
  - Give a **placebo** (a fake treatment) to those in the control group (e.g. a sugar pill)
- Best practice is to **double-blind** experiments: *both* the patients and the doctors/researchers who interact with patients are unaware of who receives which treatment
- Blinding is not always possible!

Observational studies

- Causal conclusions ***cannot*** be obtained using data from observational studies
  - There are too many confounding variables at play!
- But they are much cheaper, and can be used to identify associations or form hypotheses for future experiments!

Your turn!

In groups, design an experiment that seeks to examine the effect of coffee consumption on heart disease.

- Define your target population and response variable as exactly as possible
- Decide what your treatment(s) will be
- Make sure your experiment meets the three principles of experimental design and is generalizable
  - How will you avoid the confounding variable of smoking?
- Can you also (double) blind your experiment?