STAT011 Statistical Methods I

Lecture 8 Experimental and Sampling Design

Lu Chen
Swarthmore College
2/14/2019

Review

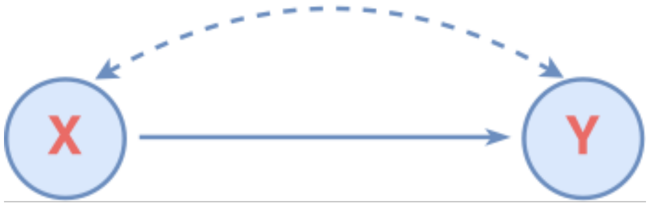
- Relationship between a quantitative variable and a categorical variable
  - Summary statistics
  - Boxplot
- Association and causation
  - Examples of relationships
    - Simpson's paradox
- Lurking variable
- Types of associations

Relationships

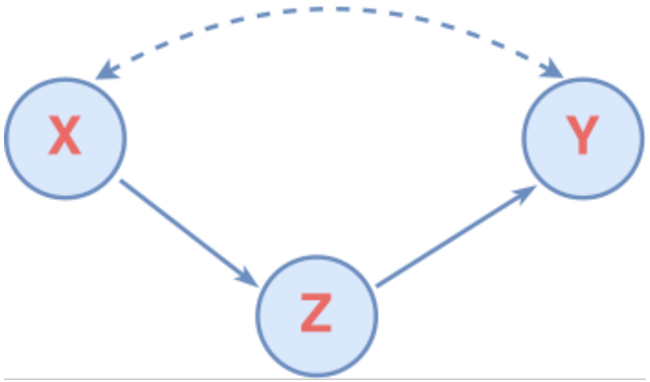
1. Student height vs. shoe length
   - Both caused by genetics and nourishment
2. UFO count vs. temperature
   - Temperature → outdoor activities → UFO reports
3. Traffic death vs. government trust score
   - Both caused by government regulations
4. Respondents' answers vs. wording of survey question
   - Wording → respondents' answers
5. Kidney stones surgery result vs. treatment
   - Surgery result caused by both treatment and severity of disease
6. Coffee consumption vs. class year
   - Class year → workload → coffee consumption

Types of associations

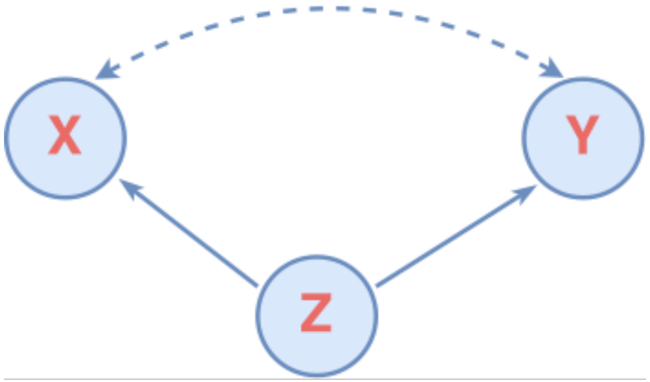
Direct Causation



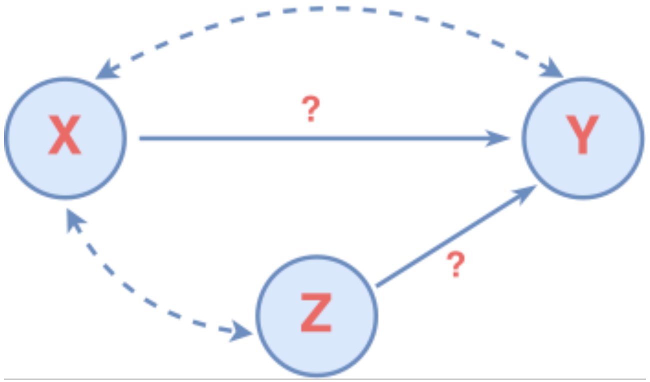
Mediation



Common Response



Confounding



Outline

- Data collection
  - Anecdotal data
  - Available data
  - Experiments
  - Observational studies
- Design of experiments
  - Types of experiments
  - Randomization
  - Principles and cautions
- Sampling design
  - Simple random samples (SRSs)
  - Stratified random samples
  - Multistage random samples

Data Collection

Anecdotal data represent individual cases, which often come to our attention because they are striking in some way. These cases are not necessarily representative of any larger group of cases.

- Your roommate: I have a date today.
- The person sitting on your right talking to someone else: I have a date today.
- Your sister: I have a date today.

Available data are data that were produced for some other purpose but that may help answer a question of interest.

- Internet
- Databases. UFO reporting, U.S. Bureau of Labor Statistics, National Center for Education Statistics, DATA.GOV, Swarthmore college common data set

Data Collection

> In an **observational study** we observe individuals and measure variables of interest but do not attempt to influence the responses.

- Prospective and retrospective studies
- Example: most of our class examples; lung cancer vs. smoking

> In an **experiment** we deliberately impose some treatment on individuals and we observe their responses.

- Experimental units: the individuals in the experiment; subjects
- Treatments: the experimental conditions applied to the units
- Outcomes: the measured variables that used to compare the treatments
- Example: surgery result vs. treatment; respondents' answers vs. wording of survey question

How to establish causal relationship

Only experiments can establish a causal relationship between two varaibles.

- Design of experiments helps to control for confounder(s).
- But sometimes experiments are not possible/ethical. For example, to establish the causal relationship between smoking and lung cancer, it is impossible to design an experiment that intentionally assigns some people to smoke.
- If an experiment is not feasible due to ethical reasons, one can use observational studies to establish a causal relationship, but with the following **conditions**:
- 1. The association is strong

  2. The association is consistent

  3. Higher doses are associated with stronger responses

  4. Cause precedes the effect in time

  5. Cause is plausible (reasonable explanation/animal experiments)

How to establish causal relationship

When confounder(s) can hardly be controlled

- Example: Autism vs. genetics/environment
- **Twin study**

  Identical twins - share exactly the same genes and grown-up environment

  Non-identical twins - share half of the genes and the same grown-up environment
- Sibling study, adoption study

Design of Experiments

Three factors: Experimental units, treatments, outcomes.

Example 1: PGS and wording of survey questions

- Experimental units: students
- Treatments: wording
  Treatment levels: positive, negative
- Outcomme: whether the students are interested in having PGS (Yes or No)

Example 2: Sleep quality vs. light and noise

- Experimental units: recruited subjects
- Treatments: light and noise
  Treatment levels: light off, noise off; light off, noise on; light on, noise off; light on, noise on.
- Outcome: quality of sleep

Design of Experiments - Types of experiments

Comparative experiments

- More than one treatment: Control group (placebo group) vs. treatment group

Matched pairs design

- Patients matched by age, doctor, severity of disease

Block design

- Patients grouped by gender: a block of males and a block of females

Note:

- An experiment could be a comparative, matched pairs and block design at the same time.

Question:

- What if there are unknown confounders or the confounders can hardly be controlled?

Design of Experiments - Randomization

- Assign the units to each of the treatment group randomly.
- Without randomization, it may result in **bias**.
- Applied to comparative experiments, matched pairs design and block design.

Bias

> The design of a study is biased if it systematically favors certain outcomes

- Samples from a population are not randomly picked
- Patients with severer conditions are assigned to the treatment group

Design of Experiments - Randomization in R

```r
# Comparative experiments using 20 subjects
sample(x = 1:20, size = 10, replace = FALSE)
```

```
##  [1] 18  7  8 16 13  6  2 17 12  5
```

```r
# Matched pairs design
# Suppose the 20 subjects are matched and result in 10 pairs
sample(x = 1:2, size = 10, replace = TRUE)
```

```
##  [1] 1 2 1 2 2 2 2 2 1 1
```

```r
# Block design
sample(x = 1:10, size = 5, replace = FALSE) # for block 1
```

```
## [1] 8 4 1 3 9
```

```r
sample(x = 1:10, size = 5, replace = FALSE) # for block 2
```

```
## [1]  9 10  5  8  2
```

Randomization in R

In R, the random generation process is **random** so that the same code gives different results **every time**. In scientific studies, the researchers usually want to have reproduceble randomizaiton results. So the `set.seed()` function is often used before any randomization.

```r
sample(x = 1:10, size = 5, replace = FALSE)
```

```
## [1] 7 1 8 5 6
```

```r
sample(x = 1:10, size = 5, replace = FALSE)
```

```
## [1]  4  7 10  6  5
```

```r
set.seed(214)
sample(x = 1:10, size = 5, replace = FALSE)
```

```
## [1] 4 9 5 3 6
```

```r
set.seed(214)
sample(x = 1:10, size = 5, replace = FALSE)
```

```
## [1] 4 9 5 3 6
```

Randomization in R

```
set.seed(214)
sample(x = 1:10, size = 5, replace = FALSE)
```

```
## [1] 4 9 5 3 6
```

```
sample(x = 1:10, size = 5, replace = FALSE)
```

```
## [1] 7 9 3 5 8
```

```
set.seed(214)
sample(x = 1:10, size = 5, replace = FALSE)
```

```
## [1] 4 9 5 3 6
```

```
sample(x = 1:10, size = 5, replace = FALSE)
```

```
## [1] 7 9 3 5 8
```

```
set.seed(2019)
sample(x = 1:10, size = 5, replace = FALSE)
```

```
## [1] 8 7 3 5 1
```

Design of Experiments - Principles & Cautions

Principles

- Compare two or more treatments. This will control the effects of lurking variables on the response variable.
- Randomize - use impersonal chance to assign experimental units to treatments.
- Repeat each treatment on many units to reduce chance variation in the results.

Cautions

- Control for confounding factors.
- Avoid bias.
- Double-blind design is recommended. Neither the subjects nor the experimenters know which treatment any subject has received.
- Avoid generalization. Statistical analysis of an experiment cannot tell us how far the results will generalize to other settings.

Sampling Design - Population and Sample

The entire group of individuals that we want information about is called the **population**.

- For example: U.S. Census

A **sample** is a part of the population that we actually examine in order to gather information.

- For example: American Community Survey

"The first census after the American Revolution was taken in 1790, under Secretary of State Thomas Jefferson; there have been 22 federal censuses since that time. The current national census was held in 2010; the next census is scheduled for 2020 and will be largely conducted using the Internet. For years between the decennial censuses, the Census Bureau issues estimates made using surveys and statistical models, in particular, the American Community Survey." — Wikipedia

Sampling Design - Population and Sample

Population does not have to be very *large*. Any group of individuals we are interested in can be called a population.

Example: Do Swarthmore students like to study at the library?

- Population: Swarthmore students
- How to obtain a sample?
  - Send out surveys via emails
    - This may result in a biased sample because the students can choose themselves to respond.
  - Give out paper surveys on campus
    - Location matters!

Obtaining a **good** sample that is **representative of the entire population** is essential to statistical analyses.

Bias in sampling

Voluntary response sample

- People choose themselves by responding to a general appeal. For example, people with strong opinions, especially negative opinions, are most likely to respond.
- Voluntary response sample is biased
  - "Are you anti-war?" - anti-war website
  - "Do you believe the existence of UFOs?" - UFO reporting website
  - How much time do you spend surfing the web at work?

Convenience sample

- Some groups in the population are left out of the process of choosing the sample
- Undercoverage may cause bias
  - Internet surveys always ignore those who do not use the internet
  - How much time do you spend surfing the web everyday?

Sampling Design - Simple Random Sample

> A **simple random sample (SRS)** of size n consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance to be the sample actually selected.

- Each of the n individuals in the sample has an equal chance to be chosen from the population.
- Suppose we want to generate a sample of 40 students from 1600 Swarthmore students.

```r
N <- 1600 # population size
n <- 40 # sample size
set.seed(214)
sample(x = 1:N, size = n, replace = FALSE)
```

```
##  [1]  512 1490  826  671 1345 1093 1560  571  988  699 1058 1173 1228  270
## [15] 1067  440   22  695 1545  428    7  864  371 1258  298  940  472  163
## [29] 1587 1512  474  329  206  895 1040   64 1065  209  650 1356
```

Sampling Design - Stratified Random Sample

> To select a **stratified random sample**, first divide the population into groups of similar individuals, called **strata**. Then choose a separate SRS in each stratum and combine these SRSs to form the full sample.

- Suppose we want a class-year-stratified random sample.

```
set.seed(214)
sample(x = 1:400, size = 10, replace = FALSE) # Freshman
```

```
##  [1] 128 372 206 167 334 271 386 141 244 172
```

```
sample(x = 401:800, size = 10, replace = FALSE) # Sophomore
```

```
##  [1] 666 695 708 468 667 510 406 573 783 506
```

```
sample(x = 801:1200, size = 10, replace = FALSE) # Junior
```

```
##  [1]  802 1019  894 1117  875 1036  919  841  868 1177
```

```
sample(x = 1201:1600, size = 10, replace = FALSE) # Senior
```

```
##  [1] 1321 1284 1253 1427 1463 1216 1469 1598 1363 1540
```

Sampling Design - Multistage Random Sample

- In multistage random sampling, we choose the sample in stages. Each stage could be a SRS or stratified random sample.
- Suppose we want a gender-stratified random sample from 5 randomly chosen departments from the 44 departments and programs at Swarthmore.

```r
set.seed(214)
# Stage 1: choose 5 departments/programs
sample(x = 1:44, size = 5, replace = FALSE)
```

```
## [1] 15 41 22 18 34
```

```r
# Stage 2: for each department, choose 4 female and 4 male students
#          Suppose the first department has 50 females and 40 males
sample(x = 1:50, size = 4, replace = FALSE)
```

```
## [1] 35 48 18 30
```

```r
sample(x = 1:40, size = 4, replace = FALSE)
```

```
## [1] 18 26 29 38
```

Sampling Design

Summary

- Data collection
  - Anecdotal data
  - Available data
  - Experiments
  - Observational studies
- Design of experiments
  - Types of experiments
  - Randomization
  - Principles and cautions
- Sampling design
  - Simple random samples (SRSs)
  - Stratified random samples
  - Multistage random samples