Introduction to Statistics

Dajiang Liu

Basic Information for PHS525

- Course title: Biostatistics for Laboratory Scientists
- Instructors:
  - Dr. Dajiang Liu: dajiang.liu@psu.edu
    - Office: HCAR 2020
    - Tel: 717-531-4178
  - Dr. Huamei Dong:
- Course website: ANGEL: https://angel.psu.edu

Introduction to Data

- Go over course syllabus
- Brief introduction to R

Data Basics

- We already use statistics unintentionally in daily life
  - Examples:
    - Flip coins
    - Casino
    - Observe variations in experiments
  - Do you get data from your experiment, and how would you approach it?

Data Basics

- A guiding example: Effectiveness of stent on stroke prevention
  - Consider each patient individually can be time-consuming
    - Cannot see the big picture underlying the data

Patient	group	0-30 days	0-365 days
1	treatment	no event	no event
2	treatment	stroke	stroke
3	treatment	no event	no event
⋮	⋮	⋮	
450	control	no event	no event
451	control	no event	no event

Table 1.1: Results for five patients from the stent study.

A guiding Example

- Data summary
  - Q1: What is the proportion of people that develop stroke after treatment in 30 days?
  - Q1: What is the proportion of people that develop stroke without treatment in 30 days?
  - Is the treatment effective?

	0-30 days		0-365 days	
	stroke	no event	stroke	no event
treatment	33	191	45	179
control	13	214	28	199
Total	46	405	73	378

Table 1.2: Descriptive statistics for the stent study.

Data Basics

- Variables, cases and data matrix

- Examples: Emails received in 2012 (Table 1.3)

	spam	num_char	line_breaks	format	number
1	no	21,705	551	html	small
2	no	7,011	183	html	big
3	yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮	⋮
50	no	15,829	242	html	small

Table 1.3: Four rows from the email50 data matrix.

Cases

Variable

Types of Variables

- Numerical variables
  - Continuous:
    - Temperature, height, BMI
  - Discrete:
    - Pain levels: 0-10
    - Cigarettes per day; drinks per week
- Categorical data
  - Regular categorical
    - Jobs: graduate students, medical students, professors
  - Ordinal:
    - How would you rate this professor: Outstanding, Good, Okay, Poor etc.

Relations between Variables

- Display relations between variables
  - Scatterplot

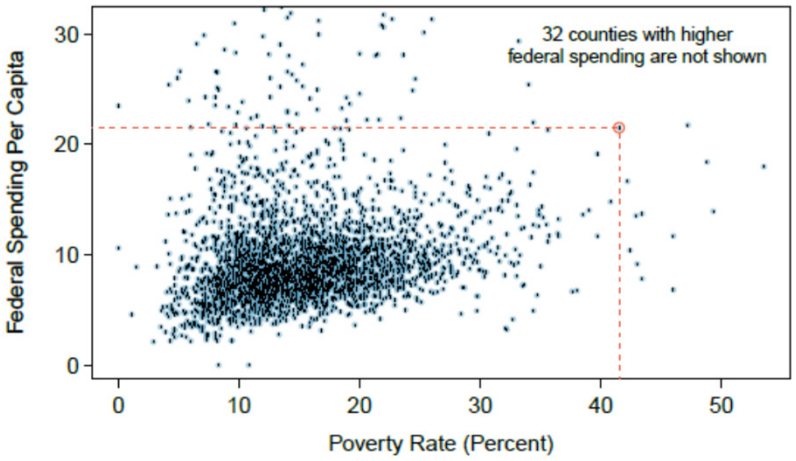


Figure 1.8: A scatterplot showing `fed_spend` against `poverty`. Owsley County of Kentucky, with a poverty rate of 41.5% and federal spending of $21.50 per capita, is highlighted.

Data Collection Principle

- Population and Samples
  - Population is what a research project targets
    - What is the average height for a Ph.D. student?
    - What is the population?
  - It is nearly impossible to get the knowledge of the entire population
  - Using a small fraction of cases to represent the entire population
    - Is the data actually representative?
    - Avoid anecdotal samples:
      - Terry Tao got a full professorship at the age of 24, so it must be easy in academia??
  - It is important to obtain random samples that are representative of the entire population!
    - Or it will introduce **bias** to the sample

Sampling from a Population

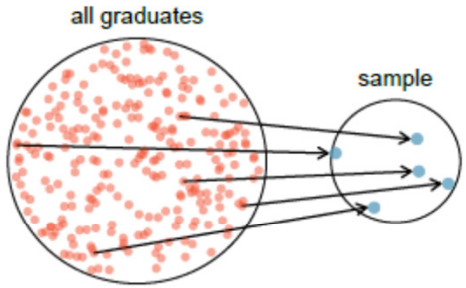


Figure 1.11: In this graphic, five graduates are randomly selected from the population to be included in the sample.

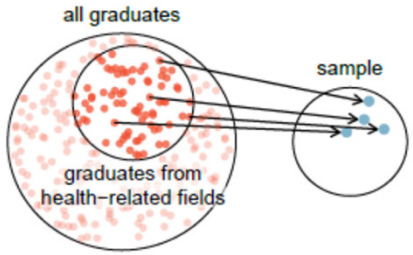


Figure 1.12: Instead of sampling from all graduates equally, a nutrition major might inadvertently pick graduates with health-related majors disproportionately often.

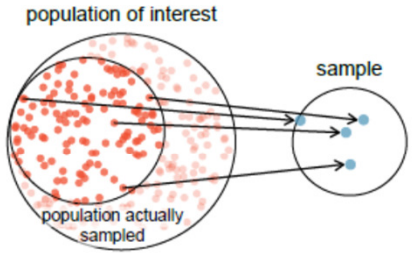


Figure 1.13: Due to the possibility of non-response, surveys studies may only reach a certain group within the population. It is difficult, and often times impossible, to completely fix this problem.

Explanatory and Response Variables

- Explanatory variable may affect response variables

- Sometimes it is easy to decide the “causation”
  - BMI affects your belt size
  - Does your belt size affect BMI?

- Sometimes causations are hard to decide
  - Government spending and poverty
  - Gene expression and disease outcome

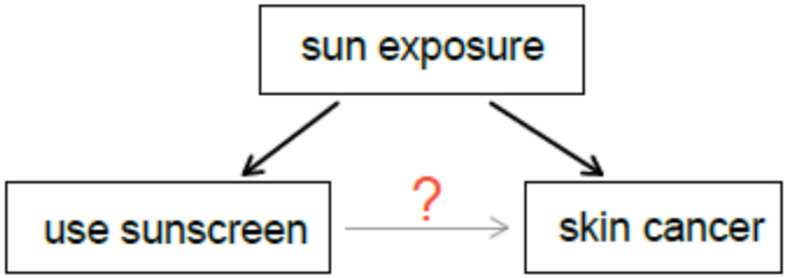
How to distinguish Causality and Association

- Observational study
  - Hard to perform experiment
    - Impossible to knockout human genes in vivo
  - Collect information via
    - Surveys
    - Review records
    - Follow a cohort over a number of years

- Experiment
  - Necessary for understanding causation
  - Clinical trials
    - Is drug A effective for disease B?

Observational Studies

- Use of sunscreen is associated with higher rate of skin cancer
- Is this relation causal?
- Sun exposure is associated with both the use of sunscreens and the rate of skin cancer
  - Which induces the correlations



Strategies for Sampling

- Simple random sampling

- Stratified sampling
  - Randomly divide the population into groups
  - Sample randomly with each group

- Cluster sampling
  - Divide the population into cluster
  - Sample a few clusters

Discussion Problems

- A nationwide survey of adults asks, “How many times per week do you eat in a fast-food restaurant?” Possible answers: 0, 1-3, 4 or more.
  a) Identify the variable.
  b) Is the variable quantitative or qualitative (categorical)?
  c) What is the implied population?

Discussion Problems

- What is the average miles per gallon (mpg) for all new cars? Using *Consumer Reports*, a random sample of 35 new cars gave an average of 21.1 mpg.
  a) Identify the variable.
  b) Is the variable quantitative or qualitative (categorical)?
  c) What is the implied population?

Discussion Problems

- Modern Managed Hospitals (MMH) is a national for-profit chain of hospitals. Management wants to survey patients discharged this past year to obtain patient satisfaction profiles. They wish to use a sample of such patients. Several sampling techniques are described below. Categorize each technique as *simple random sample*, *stratified sample*, *systematic sample*, *cluster sample*, or *convenience sample*.

Discussion Problems

- (a) Obtain a list of patients discharged from all MMH facilities. Divide the patients according to length of hospital stay (2 days or less, 3-7 days, 8-14 days, more than 14 days). Draw simple random samples from each group.

Discussion Problems

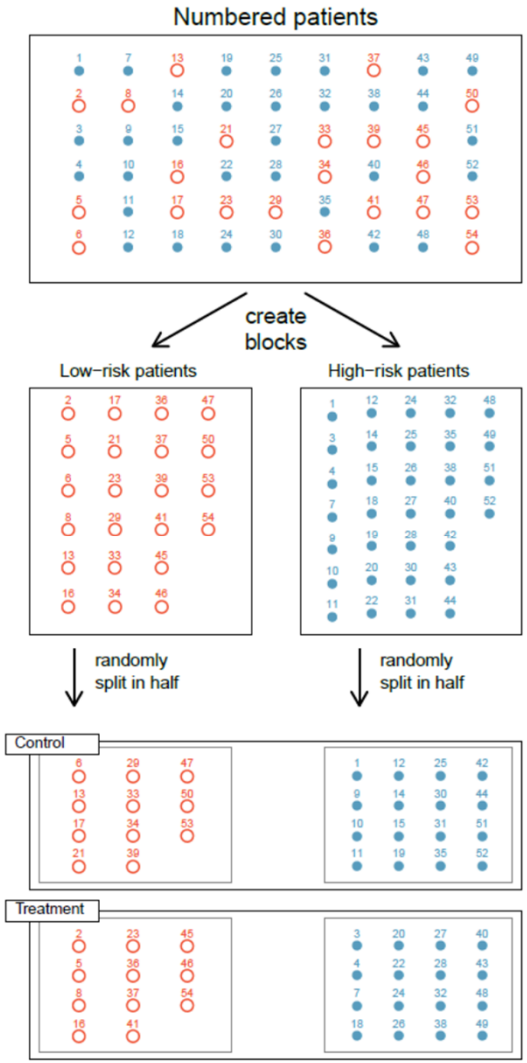
- (b) Obtain list of patients discharged from all MMH facilities. Number these patients, then use a random number table to obtain the sample.
- (c) Randomly select some MMH facilities from each of five geographic regions, and then include all the patients on the discharge list of the selected hospitals.

Discussion Problems

- (d) At the beginning of the year, instruct each MMH facility to survey every 500^{th} patient discharged.
- (e) Instruct each MMH facility to survey 10 discharged patients this week and send in the results.

Procedures for Experiment

- Studies where students assign treatments to cases are called experiments
- Controlling: controlling for all possible confounders
- Randomization: divide samples into groups to reduce the impact of uncontrollable confounders
- Replication: Reproduce your results in a separate experiment
- Blocking: divide samples into groups for different classes of confounders, and randomize within each block


Numbered patients
create blocks
Low−risk patients
High−risk patients
randomly split in half
randomly split in half
Control
Treatment

Data Examination – Numerical Data

- Scatterplot
- Histograms
- Dot plot
- Mean and standard deviation