Chapter 1: Introduction to Data

Math 140

Hitchman

based on content in OpenIntro Stats, 4th Ed

January 3, 2023

Welcome!

- **Probability** - mathematics used to determine the likelihood of certain events

Welcome!

- **Probability** - mathematics used to determine the likelihood of certain events
- **Statistics** - the science of gathering information from data. We draw conclusions about a *population* by analyzing a *sample*.
- Statistics, as a field, uses probability, but probability is its own field.

Welcome!

- **Probability** - mathematics used to determine the likelihood of certain events
- **Statistics** - the science of gathering information from data. We draw conclusions about a *population* by analyzing a *sample*.
- Statistics, as a field, uses probability, but probability is its own field.
- Skills taught in this class can be used to answer questions (responsibly) in almost any field of interest!

Welcome!

- **Probability** - mathematics used to determine the likelihood of certain events
- **Statistics** - the science of gathering information from data. We draw conclusions about a *population* by analyzing a *sample*.
- Statistics, as a field, uses probability, but probability is its own field.
- Skills taught in this class can be used to answer questions (responsibly) in almost any field of interest!
- We use RStudio in this class - a powerful tool for gathering information from data.

An example

Scene: Roll a fair 6-sided die five times. Hey! I rolled a 4 all five times!

An example

Scene: Roll a fair 6-sided die five times. Hey! I rolled a 4 all five times!

- A question in probability: What is the likelihood of rolling five 4s in a row if the die is fair?

An example

Scene: Roll a fair 6-sided die five times. Hey! I rolled a 4 all five times!

- A question in probability: What is the likelihood of rolling five 4s in a row if the die is fair?
- A question in statistics: Is this die fair?

A second example

Scene: I find four phones in the classroom after class. I "randomly" return them to the 4 students next class, and I happen to give each person the correct phone!

A second example

Scene: I find four phones in the classroom after class. I "randomly" return them to the 4 students next class, and I happen to give each person the correct phone!

- What is the likelihood that if I randomly return the phones each student gets the correct phone?

A second example

Scene: I find four phones in the classroom after class. I "randomly" return them to the 4 students next class, and I happen to give each person the correct phone!

- What is the likelihood that if I randomly return the phones each student gets the correct phone?
- Did I really return the phones at random, or did I actually use some inside info?

A third example: A case Study

This is section 1.1 in our text.

Section 1.1
A Case Study

Case Study: Using stents to prevent strokes

Identify a question

Stents are known to reduce the risk of heart attack. Many doctors have hoped that there would be similar benefits for patients at risk of stroke.

Does the use of stents reduce the risk of stroke?

Case Study: Using stents to prevent strokes

Collect Relevant Data on the Subject

451 at-risk patients were randomly assigned to one of two groups

- **Treatment Group**: Received a stent and medical management. The medical management included medications, management of risk factors, and help in lifestyle modification.
- **Control group**: No stent, but received the same medical management as the treatment group.

Why two groups?

Case Study: Using stents to prevent strokes

Collect Relevant Data on the Subject

451 at-risk patients were randomly assigned to one of two groups

- **Treatment Group**: Received a stent and medical management. The medical management included medications, management of risk factors, and help in lifestyle modification.
- **Control group**: No stent, but received the same medical management as the treatment group.

Why two groups?
The control group provides a reference point against which we can measure the medical impact of stents in the treatment group.

Case Study: Using stents to prevent strokes

Collecting data in control and treatment groups

- Researchers randomly assigned 224 patients to the treatment group and 227 to the control group.
- They studied the effect of stents at two time points: 30 days after enrollment and 365 days after enrollment.
- Patient outcomes are recorded as "stroke" or "no event", representing whether or not the patient had a stroke at the end of a time period.

Case Study: Using stents to prevent strokes

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

Case Study: Using stents to prevent strokes

Summarizing the data at 365 days

group	no event	stroke
control	199	28
treatment	179	45

Thoughts?

Case Study: Using stents to prevent strokes

Summarizing the data at 365 days

group	no event	stroke
control	199	28
treatment	179	45

Thoughts?
Proportion of patients in control group with a stroke in a year:

$$28/(199+28) \approx 0.123, \quad \text{or} \quad 12.3\%.$$

Proportion of patients in treatment group with a stroke in a year:

$$45/(179+45) \approx 0.201, \quad \text{or} \quad 20.1\%.$$

Case Study: Using stents to prevent strokes

Form a Conclusion

Does the use of stents reduce the risk of stroke?

Case Study: Using stents to prevent strokes

Form a Conclusion

Does the use of stents reduce the risk of stroke?

- The researchers expected to find that stents helped reduce the rate of stroke.
- Perhaps it's the other way around?

Case Study: Using stents to prevent strokes

Two possible conclusions

1. Stents *do help* reduce the rate of stroke, and we just happened to observe a sample with an unusually high number of strokes
2. Stents *do not help* reduce the rate of stroke.
3. Other possibilities?

Key Statistical Question

Is the observed difference due to chance, or do the data show a "real" difference?

Simulation: A tool with which to address this key question

Intermission: A first look at RStudio as a simulation tool

- Simulation results

What is Statistics?

We may place statistics in the context of an investigative process:

1. Identify a question or problem.
2. Collect relevant data on the topic.
3. Analyze the data.
4. Form a conclusion.
5. Make decisions based on the conclusion.

What is Statistics?

We may place statistics in the context of an investigative process:

1. Identify a question or problem.
2. Collect relevant data on the topic.
3. Analyze the data.
4. Form a conclusion.
5. Make decisions based on the conclusion.

Statistics as a subject focuses on making steps 2-4 objective, rigorous, and efficient.

What is Statistics?

In other words, statistics has three primary components:

- How best can we collect data?
- How should it be analyzed?
- And what can we infer from the analysis?

Common terms to know

- We are often interested in *population parameters*.
- Complete populations are difficult to collect data on, so we use *sample statistics* as *point estimates* for the unknown population parameters of interest.
- *Error* in the estimate = difference between population parameter and sample statistic
- *Bias* is systematic tendency to over- or under-estimate the true population parameter.
- *Sampling error* describes how much an estimate will tend to vary from one sample to the next.
- Much of statistics is focused on understanding and quantifying sampling error, and *sample size* is helpful for quantifying this error.

Two commonly studied population parameters

Population Mean

- Birth weight of newborns in a certain community
- Money spent by households on Halloween candy
- Percentage of monthly income spent on internet and cell service

Two commonly studied population parameters

Population Mean

- Birth weight of newborns in a certain community
- Money spent by households on Halloween candy
- Percentage of monthly income spent on internet and cell service

Population Proportion

- Are you going to vote for the Green candidate for governor?
- Do you subscribe to Netflix?
- What is your eye color?

Section 1.2
Data Basics

Storing Data

Definition

A **data matrix** is a 2-dimensional array, in which each row corresponds to an **observational unit** (individual cases) and each column corresponds to a **variable** (that is being measured).

Storing Data

	Name	height (in)	Age (yrs)	FavNum	TimeinUD (ep)	TeleAbil
1	Dustin	61.3	12	π	0	no
2	Will	61.9	12	4	7.2	no
3	Lucas	63.8	13	8	0	no
4	Eleven	62.1	11	315	nan	yes
5	Mike	64.3	13	11	0	no

Storing Data

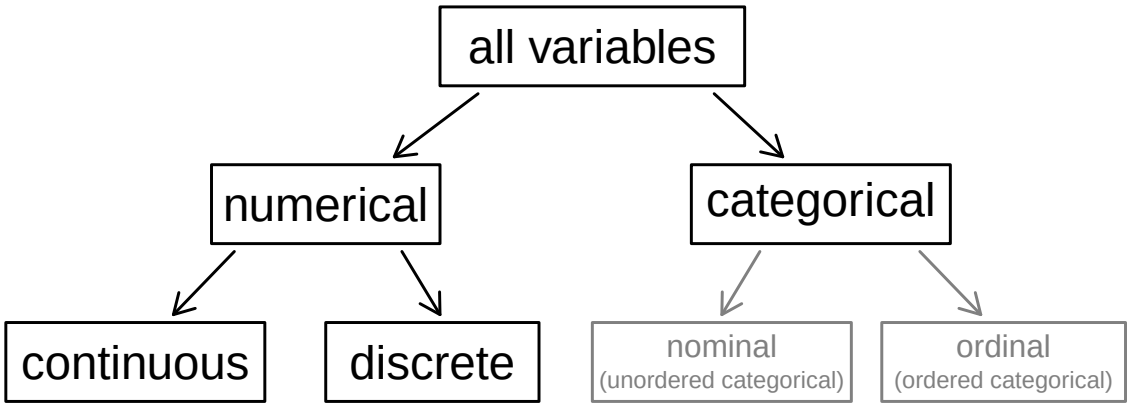
	Name	height (in)	Age (yrs)	FavNum	TimeinUD (ep)	TeleAbil
1	Dustin	61.3	12	π	0	no
2	Will	61.9	12	4	7.2	no
3	Lucas	63.8	13	8	0	no
4	Eleven	62.1	11	315	nan	yes
5	Mike	64.3	13	11	0	no

- 5 observations (characters in *Stranger Things*)
- 6 variables
- Each variable gives a piece of information about each character (name, height, time spent in the upside down (in units of 'episodes'), etc).
- The data matrix in this example has 5 rows and 6 columns

Data Matrix

- Convenient and common way to organize data
- Spreadsheets!
- This structure allows new cases to be added as rows and/or new variables to be added as columns.
- A data matrix is also called a **data frame**.

Types of Variables



Section 1.2.2 has a nice discussion of these terms. Generally speaking, if it makes sense to "do math" on a variable (like add, subtract, find the average), the variable is better thought of as numerical than categorical. A numerical variable is sometimes called a **quantitative variable**.

Figure 1.7 in *OpenIntro Stats*

Example (*Stranger Things*)

	Name	height (in)	Age (yrs)	FavNum	TimeinUD (ep)	TeleAbil
1	Dustin	61.3	12	π	0	no
2	Will	61.9	12	4	7.2	no
3	Lucas	63.8	13	8	0	no
4	Eleven	62.1	11	315	nan	yes
5	Mike	64.3	13	11	0	no

- Categorical (nominal): 'Name', 'TeleAbil'
- Numerical (discrete): 'Age'
- Numerical (continuous): 'height', 'FavNum', 'TimeinUD'
- Note: 'nan' commonly used as a place holder for a missing data value.

Relationships between variables

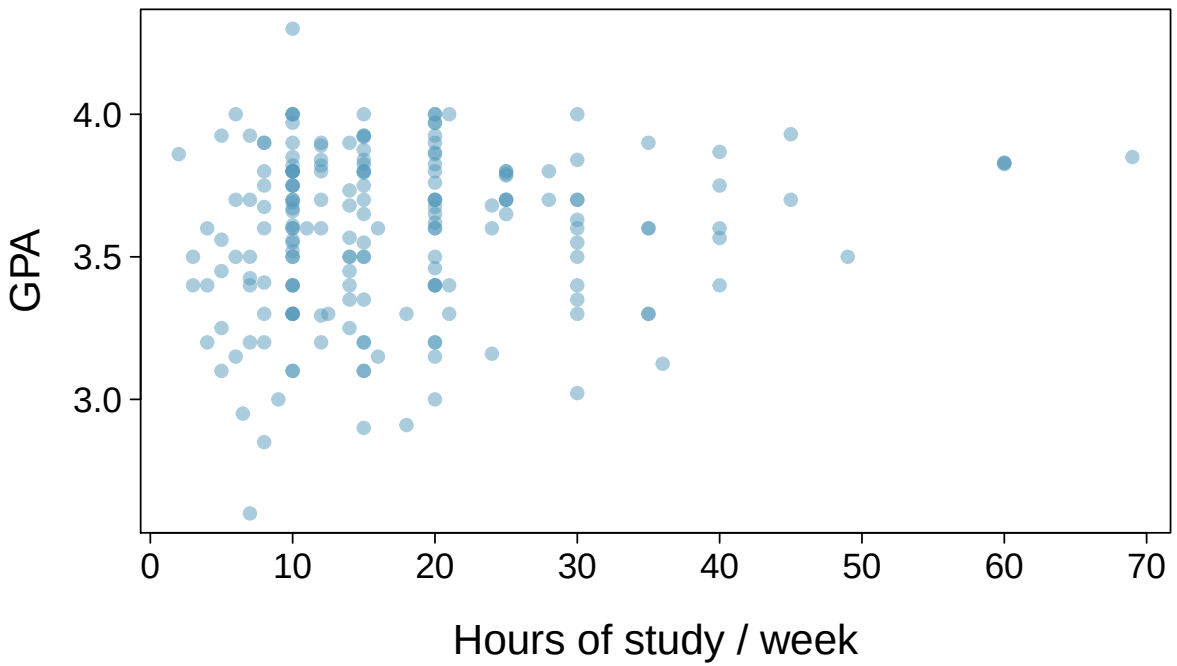
- Often we look for a relationship between two or more variables.
- A **scatterplot** gives a visual description of any association between two numerical variables.
- Categorical variables can be pictured in scatterplots as well, perhaps with colors or dot sizes!

Check out gapminder

Relationships between variables

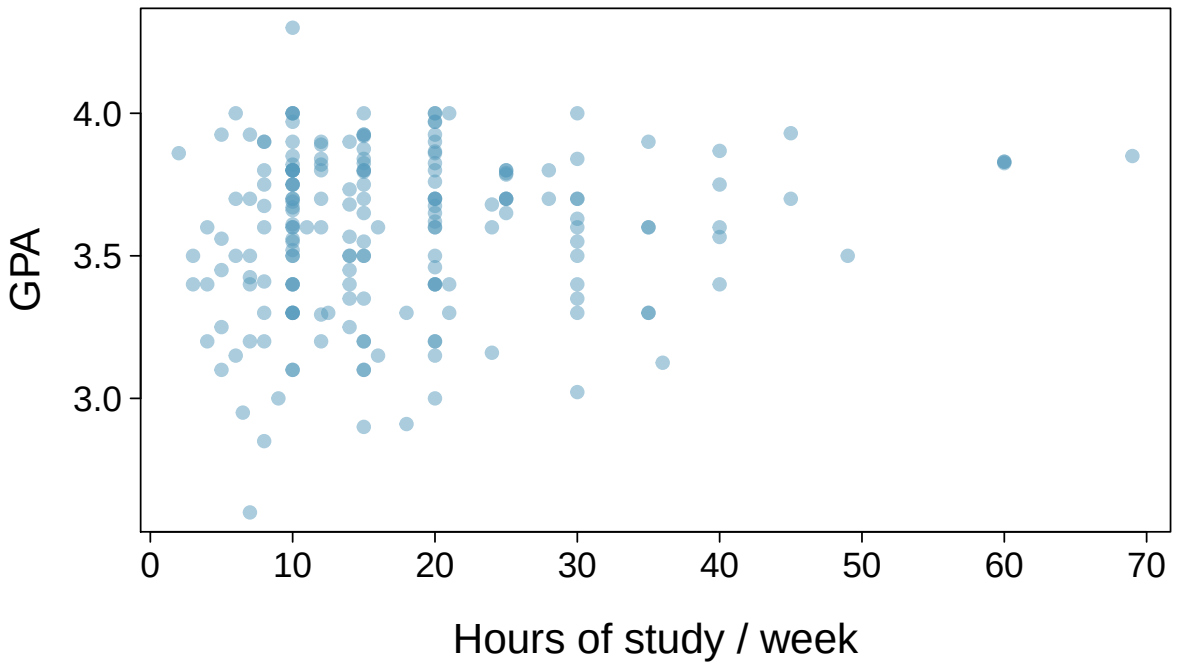
- When two variables show some connection with one another, they are called **associated variables**.
- If two variables are not associated, i.e. there is no evident connection between the two, then they are said to be **independent**.
- Two numerical variables are **negatively associated** if one tends to decrease as the other increases,
- and are **positively associated** if one tends to increase as the other increases.
- If two variables do not appear to be associated, they are said to be **independent**.

Relationships among variables



Does there appear to be a relationship between GPA and number of hours students study per week?
Can you spot anything unusual about any of the data points?

Relationships among variables



Does there appear to be a relationship between GPA and number of hours students study per week?

Can you spot anything unusual about any of the data points?

ANSWER: There is one student with GPA > 4.0, this is likely a data error.

Explanatory and response variables

- If we suspect one variable might causally affect another, we label the first variable the **explanatory variable** and the second the **response variable**.

$$\text{explanatory variable} \xrightarrow{\textit{might affect}} \text{response variable}$$

- Labeling variables as explanatory and response does not guarantee a causal relationship, of course. We use these labels only to keep track of which variable we suspect affects the other.

Explanatory and response variables

Example (Migraines and Acupuncture)

- The patients in the treatment group received acupuncture that was specifically designed to treat migraines.
- The patients in the control group received placebo acupuncture (needle insertion at non-acupoint locations).
  24 hours after patients received acupuncture, they were asked if they were pain free.
- What are the explanatory and response variables in this study?

Explanatory and response variables

Example (Migraines and Acupuncture)

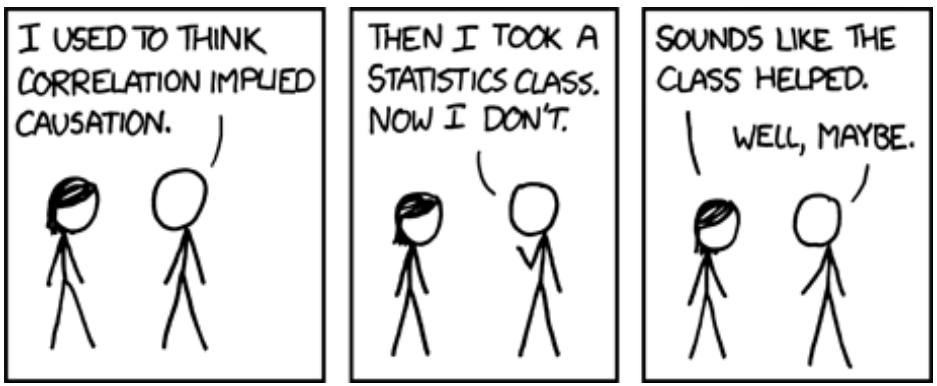
- The patients in the treatment group received acupuncture that was specifically designed to treat migraines.
- The patients in the control group received placebo acupuncture (needle insertion at non-acupoint locations).
  24 hours after patients received acupuncture, they were asked if they were pain free.
- What are the explanatory and response variables in this study?

Explanatory: type of acupuncture; Response: pain free (y or n)

Association does not mean Causation

- Is there an association between the percentage of people in a country not using the internet and the life expectancy? Does failure to use the internet decrease life expectancy?
- gapminder

Association vs. causation



http://xkcd.com/552/

Observational Studies vs Experiments

Two primary types of data collection:

- Researchers perform an **observational study** when they collect data in a way that does not directly interfere with how the data arise.
- When researchers want to investigate the possibility of a causal connection, they conduct an **experiment**, which generally has explanatory and response variables.

Observational Studies vs Experiments

Example (Does drinking alcohol affect body temperature?)

- Researchers give varying amounts of alcohol to volunteer subjects
- They measure the change in each subject's temperature in the 15 minutes after taking the alcohol.
- Observational study or Experiment?

Observational Studies vs Experiments

Example (Does drinking alcohol affect body temperature?)

- Researchers give varying amounts of alcohol to volunteer subjects
- They measure the change in each subject's temperature in the 15 minutes after taking the alcohol.
- Observational study or Experiment?

Experiment. The amount of alcohol consumed is the explanatory variable, change in body temp is the response.

Observational Studies vs Experiments

Example (Do biology majors spend more on textbooks than psychology majors?)

- Find 10 bio students and 10 psych students and compare average costs on books.
- Observational study or Experiment?

Observational Studies vs Experiments

Example (Do biology majors spend more on textbooks than psychology majors?)

- Find 10 bio students and 10 psych students and compare average costs on books.
- Observational study or Experiment?

Observational study. "Major" is the explanatory variable, "cost of books" the response.