WEEK 1

Basics of Data

1.1 Week 1 Reading Guide: Basics of Data

Sections 1.1 (Case study) and 1.2 (Data basics)

Videos

- Stat 216 Course_Tour
- Instructor bio
- 1.2.1_1.2.2
- 1.2.3_1.2.4_1.2.5

Vocabulary

Data:

Summary statistic:

Case/Observational unit:

Variable:

Quantitative variable:

Discrete variables:

Examples of discrete variables using the `County` data:

Continuous variables:

Examples of continuous variables using the `County` data:

Example of a number which is NOT a numerical variable:

Categorical variable:

Ordinal variable:

Example of an ordinal variable using the `County` data:

Nominal variable:

Examples of nominal variables using the `County` data:

Note: Ordinal and nominal variables will be treated the same in this course. We recommend taking more statistics courses in the future to learn better methods of analysis for ordinal variables.

Data frame:

Scatterplot:

Each point represents:

Positive association:

Negative association:

Association or Dependent variables:

Independent variables:

Explanatory variable:

Response variable:

Observational study:

Experiment:

Placebo:

Notes

Big Idea: Variability is inevitable! We would not expect to get *exactly* 50 heads in 100 coin flips. The statistical question then is whether any differences found in data are due to random variability, or if something else is going on.

The larger the difference, the **less we believe the difference was due to chance.**

In a data frame, rows correspond to ____________________________

and columns correspond to ___________________________.

How many types of variables are discussed? Explain the differences between them and give an example of each.

True or False: A pair of variables can be both associated AND independent.
True or False: Given a pair of variables, one will always be the explanatory variable and one the response variable.
True or False: If a study does have an explanatory and a response variable, that means changes in the explanatory variable must **cause** changes in the response variable.

True or False: Observational studies can show a naturally occurring association between variables.

Example (Section 1.1 — Case study: Using stents to prevent strokes)

1. What is the principle question the researchers hope to answer? (We call this the **research question**.)

2. When creating two groups to compare, do the groups have to be the same size (same number of people in each)?

3. What are the cases or observational units in this study?

4. Is there a clear explanatory and response variable? If so, name the variable in each role and determine the type of variable (discrete, continuous, nominal, or ordinal).

5. What is the purpose of the control group?

6. Is this an example of an observational study or an experiment? How do you know?

7. Consider Tables 1.1 and 1.2. Which table is more helpful in answering the research question? Justify your answer.

8. Describe in words what is shown in Figure 1.1. Specifically, compare the proportion of patients who had a stroke between the treatment and control groups after 30 days as well as after 365 days.

9. Given the notion that the larger the difference between the two groups (for a given sample size), the less believable it is that the difference was due to chance, which measurement period (30 days or 365 days) provide stronger evidence that there is an association between stents and strokes, or that the differences are not due to random chance?

10. This study reported finding evidence that stents *increase* the risk of stroke. Does this conclusion apply to all patients and all stents?

11. This study reported finding evidence that stents *increase* the risk of stroke. This conclusion implies a causal link between stents and an increased risk of stroke. Is that conclusion valid? Justify your answer.

1.2 Activity 1: Martian Alphabet

1.2.1 Learning outcomes

- Describe the statistical investigation process.
- Identify observational units, variables, and variable types in a statistical study.

1.2.2 Terminology review

Statistics is the study of how best to collect, analyze, and draw conclusions from data. Today in class you will be introduced to the following terms:

- Observational units or cases
- Variables: categorical or quantitative
- Proportions
- Graphs: frequency bar plot and relative frequency bar plot
- Distribution

For more on these concepts, read Sections 1.2 and 2.1 in the textbook.

1.2.3 General information on Friday labs

Each Friday you will complete a lab in class with your group. Questions are selected from each lab to be turned in on Gradescope (one submission per group). The questions to be submitted on Gradescope are bolded in the lab. As you work through the lab with your group have the Gradescope lab assignment open so that you can answer those questions as you go. Today's activity is Lab 0 in Gradescope for practice submitting as a group.

1.2.4 Can you read "Martian?"

How well can humans distinguish one "Martian" letter from another? In today's activity, we'll find out. When shown the two Martian letters, Kiki and Bumba, write down whether you think Bumba is on the left or on the right.

1. Were you correct or incorrect in identifying Bumba?

Steps of the statistical investigation process

Step 1: The first step of any statistical investigation is to *ask a research question*. In this study the research question is: Can we as a class read Martian? (We will refine this later on!).

Step 2: To answer any research question, we must *design a study and collect data*. For our question, the study consists of each student being presented with two Martian letters and asking which was Bumba. Your responses will become our observed data that we will explore.

Observational units or **cases** are the subjects data are collected on. In a spreadsheet of the data set, each row will represent a single observational unit.

2. What are the observational units in this study?

3. How many students are in class today? This is the **sample size**.

A **variable** is information collected or measured on each observational unit or case. Each column in a data set will represent a different variable. Today we are only measuring one variable on each observational unit.

4. **Identify the variable we are collecting on each observational unit in this study, i.e., what are we measuring on each student?** *Hint*: Your answer to question 1 is the outcome for the variable measured on one observational unit.

We will look at two types of variables: **quantitative** and **categorical** (see Figure 1.1).

Quantitative variables are numerical measurements that can be discrete (whole, non-negative numbers) or continuous (any value within an interval). The number of pets one owns would be a discrete variable as you can not have a partial pet. GPA would be a continuous variable ranging from 0 to 4.0.

The outcome of a categorical variable is a group or category such as eye color, state of residency, or whether or not a student lives on campus. Categorical variables with a natural ordering are considered ordinal variables while those without a natural ordering are considered nominal variables. All categorical variables will be treated as nominal for analysis in this course.

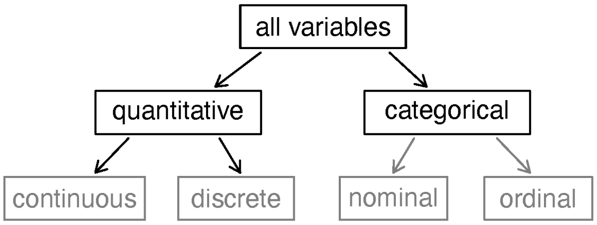


Figure 1.1: Types of variables.

5. Is the variable identified in question 4 categorical or quantitative?

Step 3: Once we have collected data, the next step is to *summarize and visualize the data.*

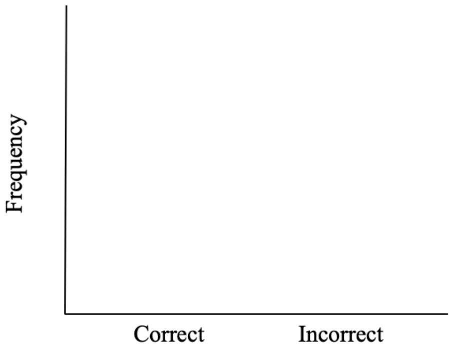
6. How many people in your class were correct in identifying Bumba? Using the class size from question 3, calculate the proportion of students who correctly identified Bumba.

$$\text{proportion} = \frac{\text{number of students who correctly identified Bumba}}{\text{total number of students}}$$

The proportion in question 6 is called a **summary statistic**—a single value that summarizes the data set. It is important to note that a variable is different than a summary statistic. A *variable* is measured on a *single observational unit* while a summary statistic is calculated from a group of observational units. For example, the variable "whether or not a student lives on campus" can be measured on each individual student. In a class of 50 students we can calculate the proportion of students who live on campus, the summary statistic. Look back and make sure you wrote the variable in question 4 as a variable, NOT a summary statistic.

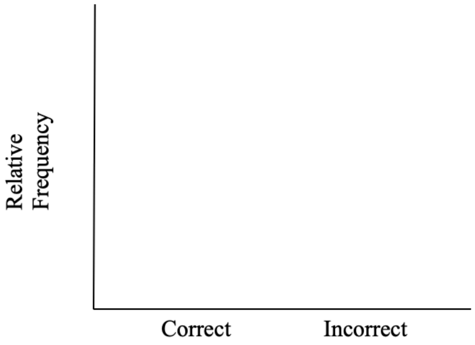
Looking at the data set and the summary statistic is only one way to display the data. We will also want to create a visualization or picture of the data. A **frequency bar plot** is used to display categorical data as a count or frequency. Since our variable has two levels or outcomes, correct or incorrect, we will create two bars—one for each level.

7. Plot the observed class data using a frequency bar plot. Be sure to add a scale to the y-axis.



We can also visualize the data as a proportion in a **relative frequency bar plot**. Relative frequency is the proportion calculated for each level of the categorical variable.

8. Plot the observed class data using a relative frequency bar plot. Be sure to add a scale to the y-axis.



Step 4: The next step is to *use statistical analysis methods to draw inferences from the data.* To answer the research question, we will simulate what *could* have happened in our class given random chance, repeat many times to understand the expected *variability* between different "randomly guessing" classes, then compare our class's observed data to the simulation. This gives us an estimate of how often (or the probability of) the class's result would occur if students were all merely guessing, allowing us to determine if the data provides evidence that we as a class can in fact read Martian.

9. If humans really don't know Martian and are just guessing which is Bumba, what are the chances of getting it right?

   How could we use a coin to simulate each student "just guessing" which Martian letter is Bumba?

   How could we use coins to simulate the entire class "just guessing" which Martian letter is Bumba?

   How many people in your class would you expect to choose Bumba correctly just by chance? Explain your reasoning.

10. Each student will flip a coin one time to simulate your "guess" under the assumption that we can't read Martian. Let Heads = correct, Tails = incorrect. What was the result of your one simulation?

    What was the result from your class's simulation? What proportion of students "guessed" correctly in the simulation?

11. If students really don't know Martian and are just guessing which is Bumba, which seems more unusual: the result from your class's **simulation** or the observed proportion of students in your class that were correct (this is your summary statistic from question 6)? Explain your reasoning.

12. While your observed class data is likely far different from the simulated "just-guessing" class, comparing our class data to a single simulation does not provide enough information. The differences seen could just be due to the randomness of that set of coin flips! Let's simulate another class. Each student should flip their coin again. What was the result from your class's second simulation? What proportion of students "guessed" correctly in the second simulation? Create a plot to compare the two simulated results with the observed class result.

13. **We still only have a couple of simulations to compare our class data to. It would be much better to be able to see how our class compared to hundreds or thousands of "just-guessing" classes. Since we don't want to flip coins all class period, your instructor will use a computer simulation to get 1000 trials. Fill in the following blanks to describe how we would create a simulation of random guessing with 1000 trials (repetitions).**

    Probability of correct guesses: ________

    Sample size: ________

    Number of repetitions: ________

14. Sketch the distribution displayed by your instructor here. Label each axis appropriately.

What does one dot on the plot above represent in context of the problem?

15. Is your class particularly good or bad at Martian? Use the plot in question 14 to explain your answer.

16. Is it *possible* that we could see our class results just by chance if everyone was just guessing? Explain your reasoning.

17. **Is it *likely* that we could see our class results just by chance if everyone was just guessing? Explain your reasoning.**

Step 5: The next step in the statistical investigation process is to *communicate the results and answer the research question.*

18. Does this activity provide strong evidence that students were not just guessing at random? If so, what do you think is going on here? Can we as a class read Martian?[1]

Introduction to `R`

In Stat 216 we will use the statistical package `R` to analyze data through the IDE (integrated development environment) RStudio. Though it is possible to download `R` and RStudio on your own computer, we will use this program through the MSU RStudio server: https://rstudio.math.montana.edu/.

Read through the preliminaries chapter in the textbook and watch the video "Starting with `R`" before completing the following questions.

The RStudio workflow operates best by the use of "Projects." You should create a separate project for each activity or assignment in this course that requires the use of `R`. To get started with this activity, follow these steps:

- Log onto the RStudio server using your NetID and password: https://rstudio.math.montana.edu/.
  - Please note: Your netID password expires every 6 months. It is HIGHLY recommended that you reset your netID password BEFORE attempting to login to the Rstudio server. You can reset your netID password in the MSU password portal (https://pwreset.montana.edu/react/).
- In the top right corner, you will see a dropdown menu next to "Project" that currently says "(None)." Click on this menu and choose "New Project." (Alternatively, you can click the "File" menu in the top left and select "New Project.")
  - A "New Project Wizard" window should pop up: click "New Directory," then click "New Project."
  - Give your project directory a name (e.g., Activity1). *Do not use spaces or other characters in the name.*
  - Click "Browse" and choose a location where you would like to save your project (you can create a new folder if desired). Note that this location is on your server account, not on your computer.
  - Leave all other boxes unchecked, and click "Create Project." (Now, if you click on the home icon in the top right, you will see your RStudio account, and the project should be listed under "Projects.")
- Download the Martian Alphabet `R` script file from D2L.
- Click "Upload" in the "Files" tab in the bottom right window of RStudio. Click "Choose File," and navigate to the folder where the Martian Alphabet `R` script file is saved. Then click "Open"; then click "Ok."
- You should see the uploaded file appear in the list of files. Click on the filename to open the file.

[1] Reference for "Martian alphabet" is a TED talk given by Vilayanur Ramachandran in 2007. The synesthesia part begins at roughly 17:30 minutes: `http://www.ted.com/talks/vilayanur_ramachandran_on_your_mind`.

In the Martian Alphabet R script file, highlight the lines of code that starts with library and click "Run." This will load the **package** (or library) catstats needed for this activity; each package is a collection of R functions. We review a few of these packages here.

- Throughout the semester we will use the package tidyverse to allow us to use chaining (see Section 1.7 in the textbook for more on this symbol %>%.) Contained in tidyverse is the package ggplot2, used to create graphs in RStudio.
- The package mosaic contains the favstats() function to find summary statistics for quantitative variables.
- We will use the package catstats, starting in Chapter 5 (and in this activity), to create simulations for statistical inference.

These packages are already installed in the RStudio server, but you need to use the library() function to call the package into your R environment. We will only use the package catstats for this activity.

The # sign is not part of the R code. It is used by these authors to add comments to the R code and explain what each call is telling the program to do. R will ignore everything after a # sign when executing the code.

In the Martian Alphabet R script file for the one_proportion_test() function arguments, enter your class size (Q3 from the in-class activity) for sample_size and the number of students who were correct in identifying Bumba (Q6 from the in-class activity) for as_extreme_as argument. Highlight lines 3 – 8 and click run.

Is the distribution created from this code similar to what you saw in class in Q14?

1.2.5 Take-home messages

1. In this course we will learn how to evaluate a claim by comparing observed results (classes' "guesses" when asked to identify Bumba) to a distribution of many simulated results under an assumption like "blind guessing."

2. Blind guessing between two outcomes will be correct only about half the time. We can simulate data using a computer program to fit the assumption of blind guessing.

3. Unusual observed results will make us doubt the assumptions used to create the simulated distribution. A large number of correct "guesses" is evidence that a person was not just blindly guessing.

1.2.6 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity and material covered, and to write down the names and contact information of your teammates.