WEEK 1

Basics of Data

1.1 Reading Guides

Reading guides are designed to be completed while reading the required sections in the course textbook to aid students in taking notes. These reading guides are not turned in in class but will be useful in understanding key concepts each week. Solutions to the reading guides will be posted on D2L.

1.2 Week 1 Reading Guide: Basics of Data

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

Videos

- Stat 216 Course_Tour
- Instructor bio
- 1.2.1and1.2.2
- 1.2.3to1.2.5

Vocabulary

Data:

Sample size:

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 (quantitative) 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:

Summary statistics:

Scatterplot:

Each point represents:

Positive association:

Negative association:

Associated or Dependent variables:

Independent variables:

Explanatory variable:

Response variable:

Observational study:

Randomized 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 (categorical or quantitative).

5. What is the purpose of the control group?

6. Is this an example of an observational study or a randomized 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 1.2. 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.

Section 2.1 (Sampling principles and strategies)

Videos

- 2.1

Vocabulary

(Target) Population:

Sample:

Statistic:

Parameter:

Anecdotal evidence:

Bias:

Selection bias:

Non-response bias:

Response bias:

Convenience sample:

Simple Random Sample:

Non-response rate:

Representative:

Notes

Ideally, how should we sample cases from our target population? What sampling method should be used?

Notes on types of sampling bias

- Someone must first be *chosen* to be in a study and refuse to participate in order to have **non-response bias**.
- There must be a valid reason for someone to lie or be untruthful to justify saying **response bias** is present. Yes, anyone could lie at any time to any question. Response bias is when those lies are *predictable and systematic* based on outside influences.

True or False: Convenience sampling tends to result in non-response bias.

True or False: Volunteer sampling tends to result in response bias.

True or False: Random sampling helps to resolve selection bias, but has no impact on non-response or response bias.

1.3 Activity 1: Intro to Data

1.3.1 Learning outcomes

- Identify observational units, variables, and variable types in a statistical study.
- Identify biased sampling methods.

1.3.2 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

For more on these concepts, read Chapter 1 in the textbook.

1.3.3 General information on the Coursepack

Information is provided throughout each activity and lab to guide students through that day's activity or lab. Be sure to read ALL the material provided at the beginning of the activity and between each question. At the end of each activity is a section called *Take-home messages* that contains key points from the day's activity. Use these to review the day's activity and make sure you have a full understanding of that material.

1.3.4 Steps of the statistical investigation process

As we move through the semester we will work through the six steps of the statistical investigation process.

1. Ask a research question.
2. Design a study and collect data.
3. Summarize and visualize the data. *Weeks 3–4*
4. Use statistical analysis methods to draw inferences from the data. *Weeks 6–13*
5. Communicate the results and answer the research question. *Weeks 6–13*
6. Revisit and look forward.

Today we will focus on the first two steps.

Step 1: The first step of any statistical investigation is to *ask a research question.* As stated in the textbook, "with the rise of data science, however, we might not start with a research question, and instead start with a data set." Today we will create a data set by collecting responses on students in class.

Step 2: To answer any research question, we must *design a study and collect data.* Our study will consist of answers from each student. 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.

1. What are the observational units or cases for today's study?

2. 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. The rows in a data set represent the observational units.

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.

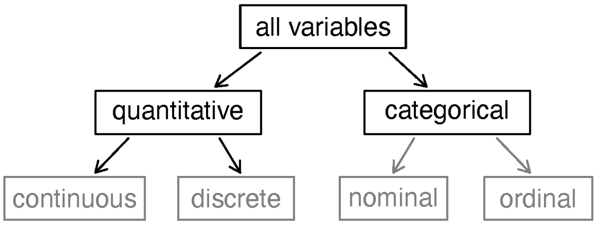

Figure 1.1: Types of variables.

3. One person from each group open the Google sheet linked in D2L and fill in the responses for the following questions for each group member. When creating a data set for use in R it is important to use single words or an underscore between words. Each outcome must be written the same way each time. Make sure to use all lowercase letters to create this data set to have consistency between responses. Do not give units of measure with the numerical values for the length of forearm. For `Residency` use in_state or out_state as the two outcomes.

- Major: what is your declared major?

- Residency: do you have in-state or out-of-state residency?

- Forearm_Length: what is the length of your arm in inches from the end of your elbow to the end of your index finger?

- Num_Credits: how many credits are you taking this semester?

4. The header for each column describes each variable measured on the observational unit. When writing a variable we need to specify what we are measuring. For example, the column header `Residency` in our data set represents the variable *whether a student has in-state or out-of-state residency* not *what state a student is from.* For each column of data, fill in the following table to write out the variable we are collecting on each observational unit in this study and the type of each variable.

Column	Variable	Type of Variable
Major		
Residency		
Forearm Length		
Num Credits		

5. Review the completed data set with your table. Remember that when creating a data set for use in R it is important to use single words or an underscore between words. Each outcome must be written the same way each time to have consistency between responses. Do not give units of measure for numerical values. Write down some issues found with the created class data set.

1.3.5 Take-home messages

1. There are two types of variables: categorical (groups) and quantitative (numerical measures).

2. When creating a data set, each row will represent a single observational unit or case. Each column represents a variable collected. It is important to write each variable as a single word or use an underscore between words.

3. Make sure to be consistent with writing each outcome in the data set as R is case sensitive. All outcomes must be written exactly the same way.

1.3.6 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.

1.4 Week 1 Lab - Sampling Methods

1.4.1 Learning outcomes

- Identify biased sampling methods.

1.4.2 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:

- Types of sampling bias
  - Selection bias
  - Response bias
  - Non-response bias

For more on these concepts, read Chapter 1 and Section 2.1 in the textbook.

1.4.3 General information on labs

On Friday of each week you will complete a lab. Questions are selected from each lab to be turned in on Gradescope. The questions to be submitted on Gradescope are bolded in the lab. As you work through the lab have the Gradescope lab assignment open so that you can answer those questions as you go.

1.4.4 Types of bias

In the next few weeks we will look at how to summarize data both numerically and graphically. For now we will focus on sampling methods and the type of sampling bias that may be present.

- Selection bias: a part of the target population is not included or is underrepresented in the sample
- Non-response or non-participation bias: part of the already selected sample does not respond or chooses not to participate
- Response bias: survey participant gives an untruthful or misleading response

To help determine the type of bias present, it is helpful to think about the observational units, the sample, and the target population represented by the problem. The **target population** is the group of cases that makes up the population the researcher is interested in. If sampling bias is present, than the sample taken will not be representative of the actual target population. In these next questions, identify the target population, the sample selected, the variable collected and its type (categorical or quantitative), and the type of bias present.

1. **To determine if the proportion of out-of-state undergraduate students at Montana State University has increased in the last 10 years, a statistics instructor sent an email survey to 500 randomly selected current undergraduate students. One of the questions on the survey asked whether they had in-state or out-of-state residency. She only received 378 responses.**

   Sample size:

   Sample taken:

   Target population:

   Variable:

   Type of Variable: categorical quantitative

   Justify why there is non-response bias in this study.

2. A television station is interested in predicting whether or not a local referendum to legalize marijuana for adult use will pass. It asks its viewers to phone in and indicate whether they are in favor or opposed to the referendum. Of the 2241 viewers who phoned in, forty-five percent were opposed to legalizing marijuana.

   Sample size:

   Sample taken:

   Target population:

   Variable:

   Type of Variable: categorical quantitative

   Justify why there is selection bias in this study.

3. To gauge the interest in a new swimming pool, a local organization stood outside of the Bogart Pool in Bozeman, MT, during open hours. One of the questions they asked was, “Since the Bogart Pool is in such bad repair, don’t you agree that the city should fund a new pool?”

   Sample size:

   Sample taken:

   Target population:

   Variable:

   Type of Variable: categorical quantitative

   Justify why there is response bias in this study.

   Justify why there is selection bias in this study.

4. **The Bozeman school district was interested in surveying parents of students about their opinions on returning to in-person classes following the COVID-19 pandemic. They divided the school district into 10 divisions based on location and randomly surveyed 20 households within each division. Explain why selection bias would be present in this study design.**

1.4.5 Gas prices

In this part of the lab we will explore two different websites to explore the cost of gas. Open both the Gas Buddy Website (www.gasbuddy.com) and a government website (https://www.eia.gov/petroleum/). Spend some time exploring each site.

5. Choose a city listed on both sites. Write down three gas prices found on Gas Buddy for this city and the reported gas price from the government website for the same city.

6. Compare the two websites.

- How are gas stations selected to appear in each data set?

- Do we know if gas stations were left out for any given time period?

- Can we make claims about what the mean price is for all gas stations in a region? Explain.

7. **Which of the following questions are best answered with the government data, and which with Gas Buddy?**

- How do average gas prices compare across regions of the country?

- Where should I go to buy gas right now?

- What will prices be like in one week? One year?

8. What type(s) of sampling bias may be present? Explain.

1.4.6 Take-home messages

1. There are three types of bias to be aware of when designing a sampling method: selection bias, non-response bias, and response bias.
2. Think about how the sample was selected and the target population when determining if sampling bias exists.
3. It is always important to look at how a sample was selected to determine which group of observational units the results of a study can be generalized to (the target population or observational units similar to the sample).

1.4.7 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.