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 (https://mtstateintrostats.github.io/IntroStatTextbook/index.html) 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

Textbook Sections 1.1: Case study and 1.2: Data basics

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.

1.3 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, class ranking, 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.

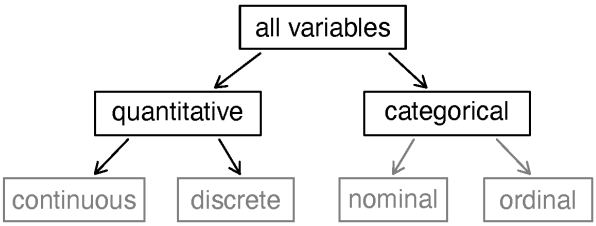

Figure 1.1: Types of variables.

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 for numerical values within the data set. 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 Lecture Notes Week 1: Intro to data

Read through Sections 1.2.1 – 1.2.5 in the course textbook prior to coming to class on Friday using the reading guides at the beginning of week 1 material.

Data basics: Sections 1.2.1 – 1.2.2

Data: _____ used to answer research questions

Observational unit or case: the people or things we _____ data from

Variable: what is measured on each _____ or _____.

Types of variables

- Categorical variable:

  Ordinal: levels of the variable have a _____ ordering

  Examples: 'Scale' questions, Years of schooling completed

  Nominal:levels of the variable do _____ have a natural ordering

  Examples: hair color, eye color, zipcode

- Quantitative variable:

  Continuous variables: value can be any _____ within a range.

  Examples: percentage of students who are nursing majors, average hours of exercise per week; distance or time (measured with enough precision)

  Discrete variables: can only be _____ values, with jumps between

  Examples: years of schooling completed; SAT score, number of car accidents

Example: The Bureau of Transportation Statistics collects data on all forms of public transportation. The data set seen here includes several variables collect on flights departing on a random sample of 150 US airports in December of 2019.

```
airport <- read.csv("data/airport_delay.csv")
glimpse(airport)
#> Rows: 150
#> Columns: 19
#> $ airport             <chr> "ABI", "ABY", "ACV", "ACY", "ADQ", "AEX", "ALB", "~
#> $ city                <chr> "Abilene", "Albany", "Arcata/Eureka", "Atlantic Ci~
#> $ state               <chr> " TX", " GA", " CA", " NJ", " AK", " LA", " NY", "~
#> $ airport_name        <chr> " Abilene Regional", " Southwest Georgia Regional"~
#> $ hub                 <chr> "no", "no", "no", "no", "no", "no", "no", "no", "n~
```

```
#> $ international       <chr> "no", "no", "no", "yes", "no", "yes", "yes", "yes"~
#> $ elevation_1000      <dbl> 1.7906, 0.1932, 0.2223, 0.0748, 0.0787, 0.0881, 0.~
#> $ latitude            <dbl> 32.4, 31.5, 41.0, 39.5, 57.7, 31.3, 42.7, 35.2, 45~
#> $ longitude           <dbl> -99.7, -81.2, -124.1, -74.6, -152.5, -92.5, -73.8,~
#> $ arr_flights         <int> 195, 81, 215, 293, 54, 282, 943, 410, 53, 32314, 6~
#> $ perc_delay15        <dbl> 16.410256, 13.580247, 23.255814, 15.358362, 12.962~
#> $ perc_cancelled      <dbl> 0.5128205, 0.0000000, 4.1860465, 0.6825939, 14.814~
#> $ perc_diverted       <dbl> 0.00000000, 0.00000000, 2.32558139, 0.68259386, 0.~
#> $ arr_delay           <int> 1563, 1244, 4763, 2905, 329, 1293, 15127, 9705, 25~
#> $ carrier_delay       <int> 459, 890, 1613, 476, 180, 302, 5627, 2253, 439, 10~
#> $ weather_delay       <int> 21, 43, 549, 124, 1, 58, 2346, 168, 1236, 13331, 2~
#> $ nas_delay           <int> 257, 39, 154, 771, 51, 112, 2096, 616, 746, 45674,~
#> $ security_delay      <int> 0, 0, 0, 25, 0, 0, 44, 0, 0, 375, 0, 83, 0, 23, 0,~
#> $ late_aircraft_delay <int> 826, 272, 2447, 1509, 97, 821, 5014, 6668, 108, 10~
```

- What are the observational units?

- Identify which variables are categorical.

- Identify which variables are quantitative.

Exploratory data analysis (EDA)

Summary statistic: a number which _____ an entire data set

- Also called the _____ _____

  Examples:

  proportion of people who had a stroke

  mean (or average) age

- Summary statistic and type of plot used depends on the type of variable(s)!

Roles of variables: Sections 1.2.3 – 1.2.5

Explanatory variable: predictor variable

- The variable researchers think *may be* _____ the other variable.

- In an experiment, what the researchers _____ or _____.
- The groups that we are comparing from the data set.

Response variable:

- The variable researchers think *may be* _____ by the other variable.
- Always simply _____ or _____; never controlled by researchers.

Examples:

Can you predict a criminal's height based on the footprint left at the scene of a crime?

- Identify the explanatory variable:

- Identify the response variable:

Does marking an item on sale (even without changing the price) increase the number of units sold per day, on average?

- Identify the explanatory variable:

- Identify the response variable:

In the Physician's Health Study, male physicians participated in a study to determine whether taking a daily low-dose aspirin reduced the risk of heart attacks. The male physicians were randomly assigned to the treatment groups. After five years, 104 of the 11,037 male physicians taking a daily low-dose aspirin had experienced a heart attack while 189 of the 11,034 male physicians taking a placebo had experienced a heart attack.

- Identify the explanatory variable:

- Identify the response variable:

Relationships between variables

- Association: the _____ between variables create a pattern; knowing something about one variable tells us about the other.
  - Positive association: as one variable _____, the other tends to _____ also.
  - Negative association: as one variable _____, the other tends to _____.
- Independent: no clear pattern can be seen between the _____.