WEEK 1

Basics of Data

1.1 Activity 1: Intro to Data

1.1.1 Learning outcomes

- Creating a data set

1.1.2 Terminology review

Statistics is the study of how best to collect, analyze, and draw conclusions from data. This week 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.

1.1.3 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.

1.1.4 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 6–14*
5. Communicate the results and answer the research question. *Weeks 6–14*
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. 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?
- Num_Credits: how many credits are you taking this semester?
- Dominant_hand: are you left or right-handed?
- Hand_span: what is the width of your dominant hand from the tip of your thumb to the tip of your pinky with your hand spread out measured in cm?
- Grip_dominant: what is the grip strength measured in lbs for your dominant hand?
- Grip_nondominant: what is the grip strength measured in lbs for your non-dominant hand?

1.1.5 Take-home messages

1. 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.
2. 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.

1.1.6 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.

1.2 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 natural ordering

    Examples: 'Scale' questions, years of schooling completed

  - Nominal:levels of the variable do not have a natural ordering

    Examples: hair color, eye color, zipcode

- Quantitative variable:

  - Continuous variables: value can be any value 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 specific values, with jumps between

    Examples: SAT score

    - number of car accidents

Example for class discussion: The Bureau of Transportation Statistics ("Bureau of Transportation Statistics" 2019) 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

- The summary statistic and type of plot used depends on the type (categorical or quantitative) 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 for class discussion:

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 ("Physician's Health Study," n.d.), 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 _______________.

Further analysis of class data set

1. What are the observational units or cases for the data collected in class on day 1?

2. How many observations are reported in the data set? This is the **sample size**.

3. The header for each column in the data set describes each variable measured on the observational unit.For each column of data, fill in the following table identifying the type of each variable, and if the variable is categorical whether the variables is binary and if the variable is quantitative the units of measure used.

Column	Type of Variable	Binary?	Units?
Major			
Residency			
Num Credits			
Dominant hand			
Hand Span			
Grip strength dominant hand			
Grip strength non-dominant hand			

4. 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.