STAT 216 Activity Coursepack

Fall 2020

Contents

Preface

This coursepack accompanies the textbook for STAT 216: Introduction to Statistics at Montana State University. Each of the activities in this workbook is designed to target specific learning outcomes of the course, giving you practice with important statistical concepts in a group setting with instructor guidance. Bring this workbook with you to class each week, and take notes in the workbook as you would your own notes. A well-written complete workbook will provide an optimal study guide for exams!

Martian Alphabet

Learning Outcomes

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

Activity

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 the right.

Steps of Statistical Investigation

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!). To answer any research question, we must design a study and collect data. (This will normally be provided for you in class.) 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. To answer the research question we will simulate what *could* have happened in our class given random chance, repeat that many times to understand the expected variability between different "randomly guessing" classes, then comparing our class's observed data to the simulation. This gives us an estimate of how often (or the probability of) our class's result would occur if we were all merely guessing, allowing us to determine if we as a class can in fact read Martian.

Let's explore the data. **Observational units** or **cases** are the subjects data is collected on. In a data set the rows will represent a single observational unit.

1. What are the observational units in this 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. We will look at two types of variables: **quantitative** and **categorical**. Each column in a data set will represent a different variable.

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

Categorical variables are data that are in groups or categories such as eye color, state of residency, or whether or not a student is considered in-state. Categorical variables with a natural ordering are considered ordinal variables while those without a natural ordering are considered a nominal variable. All variables will be treated as nominal for analysis.

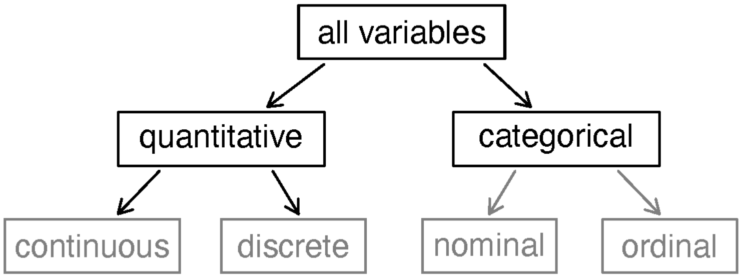


Figure 1: Types of variables.

3. Identify the variable we are collecting on each observational unit in this study, i.e., what are we measuring on each student?

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 is considered in-state** can be measured on each individual student. In a class of 50 students we can calculate the proportion of students who are considered in-state, the summary statistic. Make sure you wrote the variable in question 3 as a variable **NOT** a summary statistic.

4. Is the variable identified in question 3 categorical or quantitative?

5. Were you correct or incorrect in identifying Bumba?

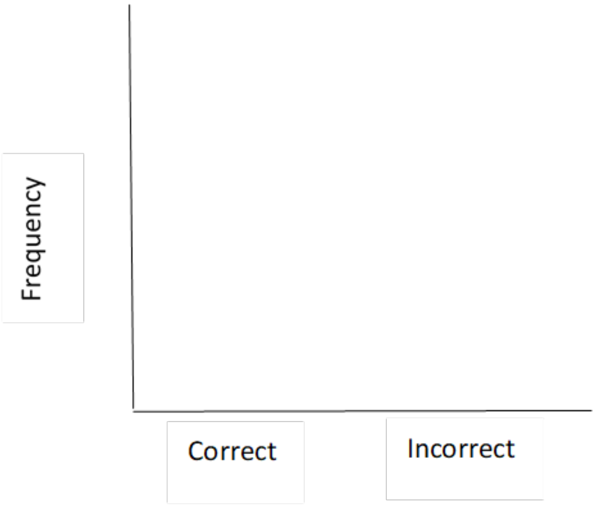
We will now collect the data from the entire class.

6. How many people in your class were correct in identifying Bumba? Using the class size from question 2, calculate the proportion of students who correctly identified Bumba.

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

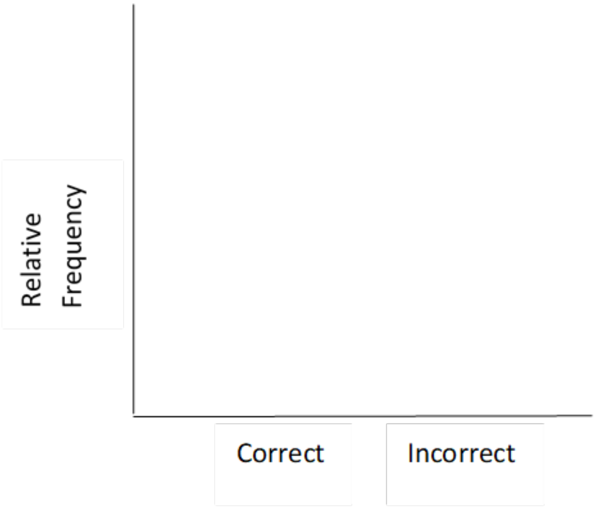
Looking at the data set and the summary statistics is only one way to display the data. We will also want to create a visualization or picture of the data. A **frequency bar plot** is used to display categorical data as a count or frequency. Since our variable has two levels, correct or incorrect, we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

7. Plot the observed class data using a frequency bar plot.



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.



9. The next step is to analyze the data. 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 of you will flip a coin one time to simulate your "guess". Let Heads = correct, Tails = incorrect. What was the result of your 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 data 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 seem to give enough information. The differences seen could just be due to that set of coin flips! Let's simulate another class. Each student should flip your 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 unfortunately 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.

    Probability of correct guesses: ________

    Sample size: ________

    Number of repetitions: ________

14. Sketch the distribution displayed by your instructor here, being sure to label each axis appropriately.

15. Is your class particularly good or bad at Martian? How can you use the plot in question 14 to tell?

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.

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?

Take Home Messages

1. In this course we will learn how to evaluate a claim by comparing observed results (classes' "guesses") 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 create data (via computer simulation) 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.

Additional Notes

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

Study Design

Learning Outcomes

- Explain the purpose of random sampling and its effect on scope of inference
- Explain the purpose of random assignment and its effect on scope of inference
- Identify whether a study is observational or an experiment
- Identify confounding variables in observational studies and explain why they are confounding
- Identify the types of bias present in a study

Terminology Review

Statistics is the study of how best to collect, analyze, and draw conclusions from data. Statistical inference will allow us to make a statement about a population parameter based on a sample statistic.

Some terms covered in this activity are…

- Population
- Sample
- Parameter
- Statistic
- Selection Bias
- Response Bias
- Non-response Bias
- Scope of Inference
- Explanatory Variable
- Response Variable
- Confounding Variable
- Experiments
- Observational Study

To review these concepts see Section 1.3 to 1.6 in the textbook.

Types of Bias

There are two parts to study design: how the sample was selected and how the study was conducted. First we will look at sampling and types of bias.

In these next questions, identify the target population, the sample, the variable, 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.

   Target population:

   Sample:

   Variable:

   Type of Bias:

2. PEW Research surveys US adults about many different topics. Recently a survey was conducted to assess current presidential approval. A random sample of 6395 US adults was taken. Of those surveyed, 42% say they agree with President Trump on many or nearly all of the top issues facing the country today.

   Target population:

   Sample:

   Variable:

   Type of Bias:

3. A television station is interested in predicting whether or not voters in its listening area are opposed to legalizing marijuana for adult use. It asks its viewers to phone in and indicate whether they are in favor of this or opposed to this. Of the 2241 viewers who phoned in, forty-five percent were opposed to legalizing marijuana.

   Target population:

   Sample:

   Variable:

   Type of Bias:

4. To gauge the interest in a new swimming pool, a local organization stood outside of the Bogart Pool 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?”

   Target population:

   Sample:

   Variable:

   Type of Bias:

5. The Bozeman school district is interested in surveying parents of students about their opinions on returning to school this fall following the COVID-19 pandemic. They divided the school district into 10 divisions based on location and randomly surveyed 20 households within each division.

   Target population:

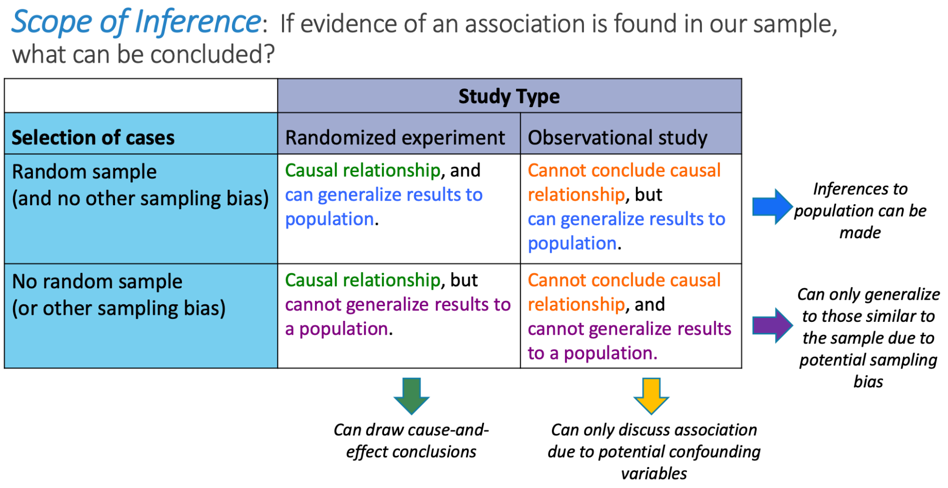
   Sample:

   Variable:

   Type of Bias:

Study Design

The two main study designs we will cover are observational studies and experiments. Both the sampling method and the study design will help to determine the **scope of inference** for a study. Remember that only in a randomized experiment can we conclude a **causal** (cause and effect) relationship between the explanatory and response variable.



Scope of Inference: If evidence of an association is found in our sample, what can be concluded?

	Study Type	
Selection of cases	Randomized experiment	Observational study
Random sample (and no other sampling bias)	Causal relationship, and can generalize results to population.	Cannot conclude causal relationship, but can generalize results to population.
No random sample (or other sampling bias)	Causal relationship, but cannot generalize results to a population.	Cannot conclude causal relationship, and cannot generalize results to a population.

Inferences to population can be made

Can only generalize to those similar to the sample due to potential sampling bias

Can draw cause-and-effect conclusions

Can only discuss association due to potential confounding variables

For the next exercises, identify the explanatory variable, the response variable, a potential confounding variable, and the study design.

6. The pharmaceutical company, Moderna Therapeutics is working in conjunction with the National Institute of Health towards a vaccine for COVID-19 and has recently begun Phase 3 clinical trials. US Clinical research sites will enroll 30,000 volunteers without COVID-19 to participate. Participants will be randomly assigned to receive either the candidate vaccine or a saline placebo. They will then be followed to assess vaccine related symptoms and development of COVID-19. The trial is blinded, so the investigators and the participants will not know who is assigned to which group.

   Explanatory Variable:

   Response Variable:

   Confounding Variable:

   Study design:

   What is the scope of inference for this study?

7. In another study, a local health department randomly selected 1000 US adults without COVID-19 to participate in a health survey. Each participant was assessed at the beginning of the study and then followed for 1 year. They were interested to see which participants elected to receive a vaccination for COVID-19 and whether any participants developed COVID-19.

   Explanatory Variable:

   Response Variable:

   Confounding Variable:

   Study design:

   What is the scope of inference for this study?

8. What is a potential confounding variable for the study in question 7? Explain how this meets the definition of a confounding variable.

Additional Notes

Use this space to summarize your thoughts and take additional notes on today's activity

Exploratory Data Analysis: Categorical Variables

Learning Outcomes

- Identify and create appropriate summary statistics and plots given a data set or research question
- Plots for a single categorical variable: bar plot
- Plots for association between two categorical variables: segmented bar plot, mosaic plot
- Recognize and simulate probabilities as long-run frequencies
- Construct two-way tables to evaluate conditional probabilities

Terminology Review

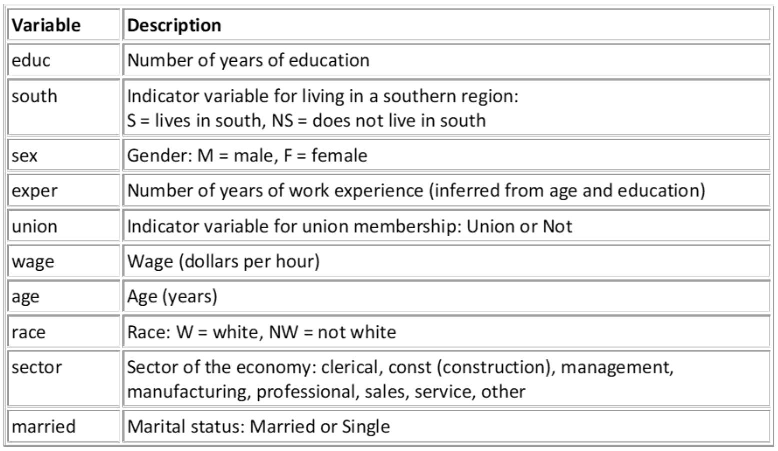
In today's activity we will review summary measures and plots for categorical variables. Some terms covered in this activity are…

- Proportions
- Bar plots
- Segmented bar plots
- Probability
- Conditional Probability
- Two-way tables

To review these concepts see Section 2.1 in the textbook.

Activity

The data set we will use for this activity is from the Current Population Survey in 1985. The CPS is a survey sponsored by the Census Bureau and the Bureau of Labor Statistics to track labor force statistics for the United States population. The following table summarizes the data:



Variable	Description
educ	Number of years of education
south	Indicator variable for living in a southern region: S = lives in south, NS = does not live in south
sex	Gender: M = male, F = female
exper	Number of years of work experience (inferred from age and education)
union	Indicator variable for union membership: Union or Not
wage	Wage (dollars per hour)
age	Age (years)
race	Race: W = white, NW = not white
sector	Sector of the economy: clerical, const (construction), management, manufacturing, professional, sales, service, other
married	Marital status: Married or Single

Vocabulary Review

1. What are the observational units?

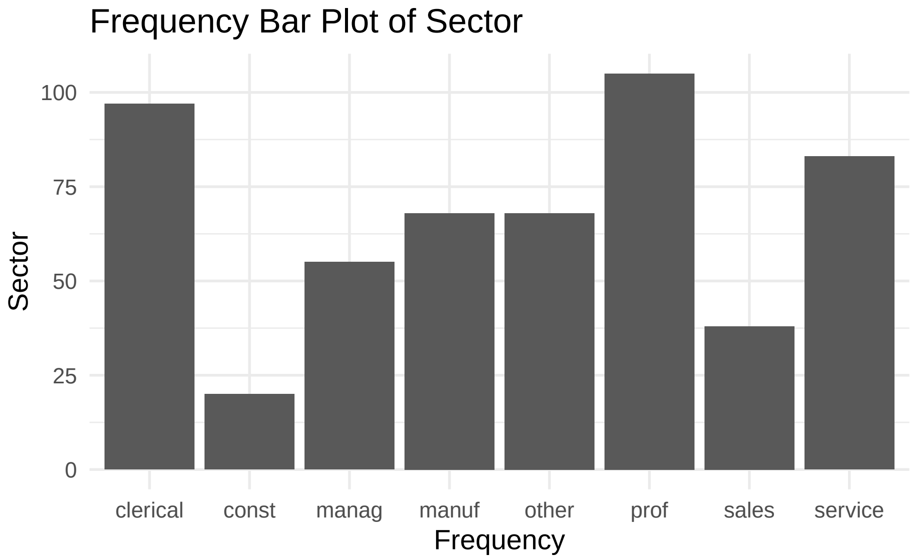
2. Which variables are categorical?

3. What types of plots can be used to display categorical data?

An important part of understanding data is to create visual pictures of what the data represents. In this activity we will create graphical representations of categorical data.

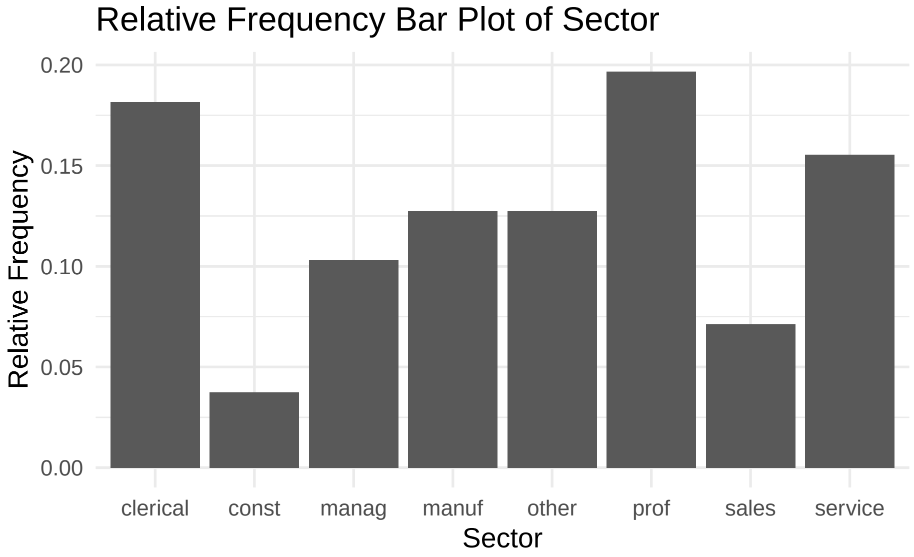
Displaying a single categorical variable.

If we wanted to know how many people in our data set were in each sector, we would create a bar plot of the variable sector.



4. Which Sector has the largest number of people in it?

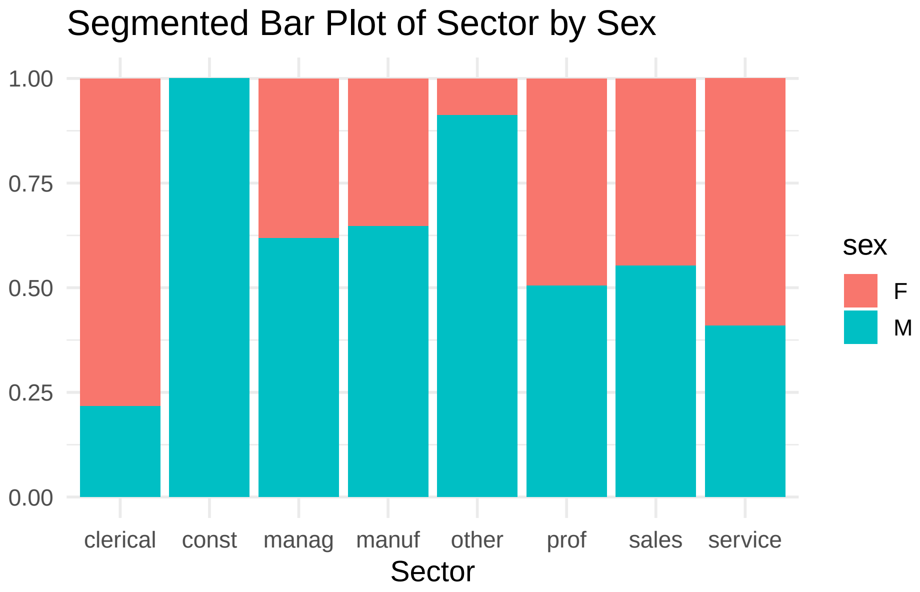
We could also choose to display the data as a proportion in a relative frequency bar plot. To find the relative frequency divide the count in each sector by the sample size. These are sample proportions.



5. What features in this plot are same as the frequency bar plot? Which are different?

Displaying two categorical variables

To see the differences in proportion of each sector between males and females we would create a segmented bar plot of sector segmented by sex.



6. Using the segmented bar plot, which sector has about the same proportion of males and females?

7. Which sector has the highest proportion of females?

8. Which variable is the explanatory variable? Which is the response variable?

Probability

9. A study was reported in which ninth grade Minnesota teens were asked whether they had gambled at least once a week in the past year. The sample consisted of 49.1% boys. The proportion of boys who had gambled at least once per week during the past year was 0.229, while among non-boys this proportion was only 0.045.
   Let B = the event the person is a boy, and C = the event the person is a weekly gambler.

a. Draw a segmented bar plot of gambling segmented by sex.

b. Identify what each numerical value represents in probability notation.

0.491 =

0.229 =

0.045 =

c. Create a two-way hypothetical table to represent the situation.

	Total
Total	100,000

d. Find P(B and C). What does this probability represent in the context of the problem?

e. Find the probability that a selected non-gambler is a non-boy. What is the notation this probability?

10. In a computer store, 30% of the computers in stock are laptops and 70% are desktops. Five percent of the laptops are on sale, while 10% of the desktops are on sale. Let L = the event the computer is a laptop, and S = the event the computer is on sale.

a. Identify what each numerical value represents in probability notation.

0.30 =

0.70 =

0.05 =

0.10 =

b. Create a two-way table to represent the situation.

			Total
Total			100,000

c. Calculate the probability that a randomly selected computer will be a desktop, given that the computer is on sale. What is the notation used for this probability?

d. Find $P(S|L^C)$. What does this probability represent in context of the problem?

Additional Notes

Use this space to summarize your thoughts and take additional notes on today's activity

Exploratory Data Analysis: Quantitative Variables

Learning Objectives

- Identify and create appropriate summary statistics and plots given a data set or research question for quantitative data
- Interpret the following summary statistics in context: median, lower quartile, upper quartile, standard deviation, inter-quartile range
- Given a plot or set of plots, describe and compare the distribution(s) of a single quantitative variable (center, variability, shape, outliers)

Terminology Review

In today's activity we will review summary measures and plots for quantitative variables. Some terms covered in this activity are

- Two measures of center
  - Mean
  - Median
- Two measures of spread (variability)
  - Standard deviation
  - IQR
- Boxplots, dotplots, histograms

To review these concepts see Section 2.3 in the textbook.

Movies Released in 2016

A data set was collected on Movies released since 1916 to 2016. Here is a list of some of the variables collected on these movies.

- Year: Year the movie was released
- Budget: The amount of money (in US $ millions) budgeted for the production of the movie
- Revenue: The amount of money (in US $ millions) the movie made after release
- Duration: The length of the movie (in minutes)

- Content Rating: Rating of the movie (G, PG, PG-13, R, Not Rated)
- IMDb Score: User rating score from 1 to 10
- Genre: Category the movie falls into
- Movie Facebook Likes: Number of likes a movie receives on Facebook

Vocabulary Review

1. What are the observational units in this data set?

2. Which of the above listed variables are categorical?

3. Which of the above listed variables are quantitative?

Summarizing a single quantitative variable

The favstats function gives the summary statistics for a quantitative variable. Here we have the summary statistics for the variable 'IMDb'.

```
#>  min  Q1 median  Q3 max     mean        sd  n missing
#>  3.4 5.9    6.6 7.1 8.2 6.459016 0.9218418 61       0
```

4. Give the values for the two measures of center.

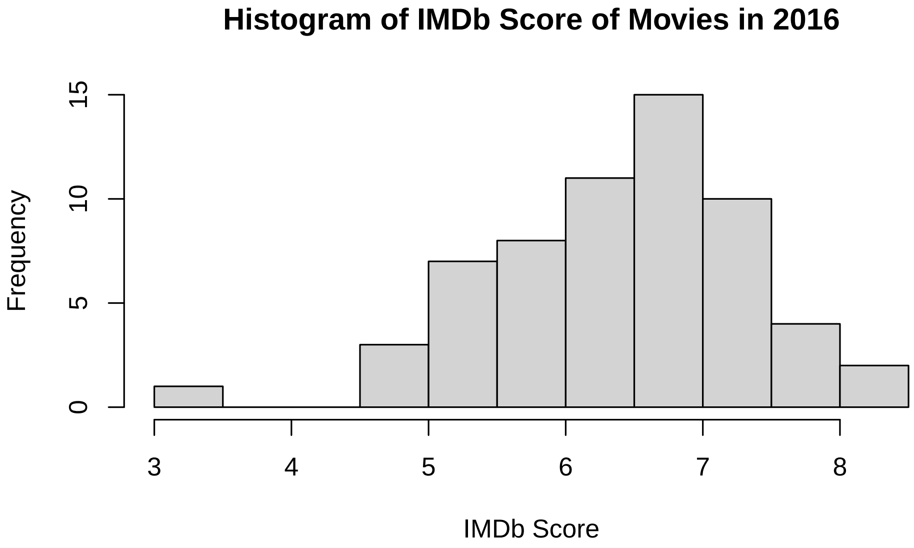
5. Calculate the IQR.

6. Report the value of the standard deviation and interpret this value in context of the problem.

Displaying a single quantitative variable

7. What are the three types of plots used to plot a single quantitative variable?

A histogram of the variable 'IMBd Score' is shown below. Notice that the bin width is 0.5. For example the first bin consists of the number of movies in the data set with an IMBd score of 3 to 3.5. It is important to note that a movie with a IMBd score of 5 will fall into the bin for 5 - 5.5. Visually this shows us the range of IMBd scores for Movies released in 2016.



8. Which range of IMBb scores have the highest frequency?

9. What is the shape of the distribution of IMDb scores?

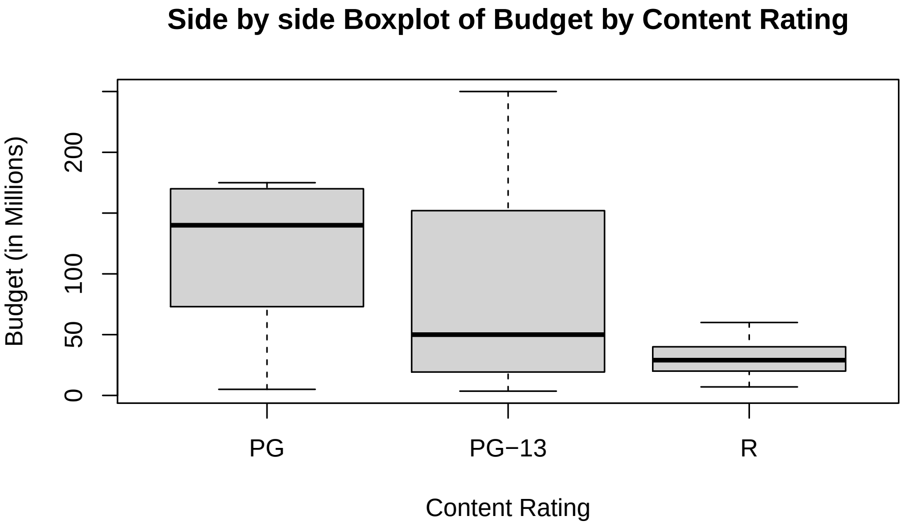
The boxplot is created using the five number summary:

- Minimum value
- Quartile 1 (Q1) - the value at the 25th percentile
- Median - the value at the 50th percentile
- Quartile 3 (Q3) - the value at the 75th percentile
- Maximum value

10. The three smallest IMDb scores in the data set are 3.4, 3.5, and 3.7 and the three largest IMDb scores are 8.5, 8.7, and 9.1. Using the summary statistics above, sketch a boxplot of IMDb Score. Be sure to label the axes.

Displaying a Single Categorical and Single Quantitative Variable

The boxplot of ‘Budget’ in millions by ‘Content rating’ is plotted using the code below. This plot helps to compare the budget for different levels of content rating.



11. Answer the following questions about the boxplots above.

a. Which content rating has the highest center?

b. Which content rating has the largest spread?

c. Which content rating is the most symmetric?

d. Fifty percent of movies in 2016 with a PG-13 content rating fall below what value?

e. What is the value for Q3 for the PG-13 rating? Interpret this value in context.

Additional Notes

Use this space to summarize your thoughts and take additional notes on today's activity

Exploratory Data Analysis: Multivariate Thinking

Learning Objectives

- Identify and create appropriate summary statistics and plots given a data set with two quantitative variables
- Use scatterplots to assess the relationship between two quantitative variables
- Find the correlation coefficient
- Find the estimated line of regression using summary statistics and R Linear Model Output
- Understand what the slope coefficient represents
- Understand what the coefficient of determination is

Movies Released in 2016

We will revisit the data set used last week collected on Movies released since 1916 to 2016. Here is a reminder of the variables collected on these movies.

- Year: Year the movie was released
- Budget: The amount of money (in US $ millions) budgeted for the production of the movie
- Revenue: The amount of money (in US $ millions) the movie made after release
- Duration: The length of the movie (in minutes)
- Content Rating: Rating of the movie (G, PG, PG-13, R, Not Rated)
- IMDb Score: User rating score from 1 to 10
- Genre: Category the movie falls into
- Movie Facebook Likes: Number of likes a movie receives on Facebook

Terminology Review

In today's activity we will review summary measures and plots for two quantitative variables. Some terms covered in this activity are

- Scatterplot

- Correlation
- Slope
- Line of Regression
- Coefficient of determination

To review these concepts see Chapter 3 in the textbook.

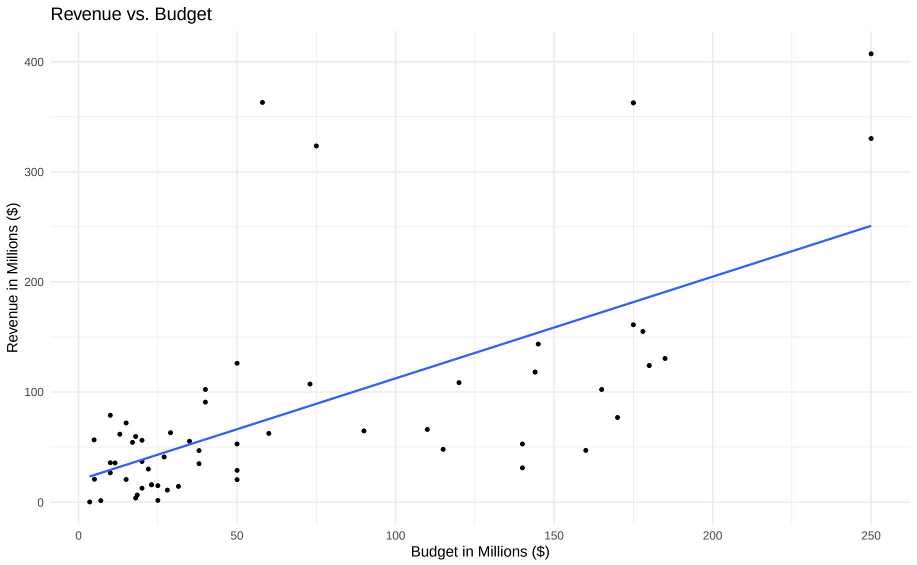
Vocabulary Review

1. What type of plot is used to display two quantitative variables?

2. What summary statistics are used to describe the relationship between two quantitative variables?

We will look at the relationship between ‘Budget’ and ‘Revenue’ for movies released in 2016. This shows a scatterplot of ‘Budget’ as a predictor of ‘Revenue’ (note: both variables are measures in “millions of dollars”.

```
ggplot(data = moviesa,   #This is the data set
       aes(x = budget_mil, y = revenue_mil))+  #Specify variables
  geom_point() +  #Add scatterplot of points
  labs(x = "Budget in Millions ($)",  #Label x-axis
       y = "Revenue in Millions ($)",  #Label y-axis
       title = "Revenue vs. Budget") + #Be sure to tile your plots
  geom_smooth(method = "lm", se = FALSE)  #Add regression line
```



3. Assess the four features of the scatterplot that describe this relationship.
    - Form (linear, non-linear)

    - Direction (positive, negative)

- Strength

- Unusual Observations or Outliers

4. Does there appear to be an association between ‘Budget’ and ‘Revenue’? Explain.

Correlation

Correlation measures the strength and the direction between two quantitative variables. The closer the value of correlation to + or - 1 the stronger the linear relationship. Values close to zero indicate a very weak linear relationship between the two variables. The following output shows a correlation matrix between several pairs of quantitative variables.

```
#>                      budget_mil revenue_mil duration imdb_score
#> budget_mil               1.0000      0.6466   0.5274     0.3081
#> revenue_mil              0.6466      1.0000   0.2516     0.4876
#> duration                 0.5274      0.2516   1.0000     0.2362
#> imdb_score               0.3081      0.4876   0.2362     1.0000
#> movie_facebook_likes     0.6481      0.6710   0.5619     0.3462
#>                      movie_facebook_likes
#> budget_mil                         0.6481
#> revenue_mil                        0.6710
#> duration                           0.5619
#> imdb_score                         0.3462
#> movie_facebook_likes               1.0000
```

5. Using the output above, which two variables have the strongest correlation?

6. What is the value of correlation between 'Budget' and 'Revenue'?

7. Based on the value of correlation what would the sign of the slope be? Positive or negative? Explain.

8. Does your answer to question 13 match the direction you choose in question 3?

9. Explain why the correlation values on the diagonal are equal to 1.0.

Slope

The slope measures the change in y for each increase in x by 1. In other words, as the x variable increases by 1 unit, the y variable changes (increase/decreases) by the value of slope.

The linear model function in R gives us the summary for the least squares regression line. The estimate for (Intercept) is the y-intercept for the line of least squares and the estimate for budget is the value of b_1, the slope.

```
#>               Estimate Std. Error  t value     Pr(>|t|)
#> (Intercept) 20.0362329 14.3458255 1.396659 1.677479e-01
#> budget_mil   0.9236972  0.1418579 6.511426 1.806269e-08
```

10. Write out the least squares line using the summary statistics provided.

11. Interpret the value of slope in context of the problem.

12. Using the least squares line from Question 10, predict the revenue for a movie with a budget of 165 million.

Residuals:

The model we are using assumes the relationship between the two variables follows a straight line. The residuals are the errors, or the part that hasn't been modeled by the line.

$$\text{Data} = \text{Model} + \text{Residual}$$

$$\text{Residual} = \text{Data - Model}$$

$$e_i = y_i - \hat{y}_i$$

13. The movie, *Independence Day: Resurgence*, had a budget of 165 million and revenue of 102.315 million. Find the residual for this movie.

14. Did the line of regression overestimate or underestimate the revenue for this movie?

Coefficient of Determination

The coefficient of determination, R^2, can also be used to describe the strength of the linear relationship between two quantitative variables. R^2 describes the amount of variation in the response that is explained by the least squares line with the explanatory variable.

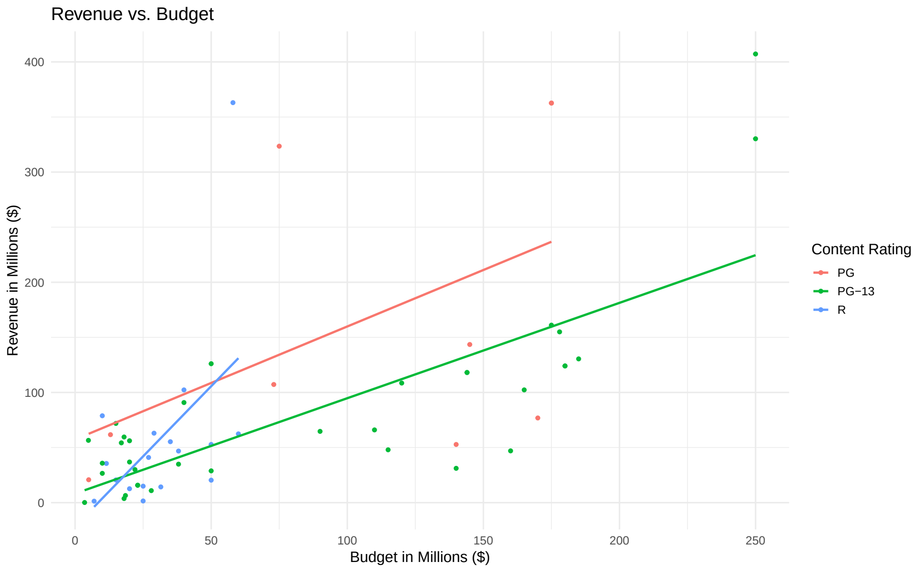
15. Calculate the coefficient of determination between 'Budget' and 'Revenue'.

16. Interpret the coefficient of determination in context of the problem.

Multivariate Plot

In the next plot we are graphing three variables.

```
ggplot(data = moviesa,   #This is the data set
       aes(x = budget_mil, y = revenue_mil, color = content_rating))+  #Specify variables
  geom_point() +  #Add scatterplot of points
  labs(x = "Budget in Millions ($)",  #Label x-axis
       y = "Revenue in Millions ($)",  #Label y-axis
       color = "Content Rating",  #Label legend
       title = "Revenue vs. Budget") + #Be sure to tile your plots
  geom_smooth(method = "lm", se = FALSE)  #Add regression line
```



25. Identify the three varables plotted in this graph.

26. Does the relationship between ‘Budget’ and ‘Revenue’ differ among the different content ratings? Explain.

Additional Notes

Use this space to summarize your thoughts and take additional notes on today’s activity