**Lab 1: Chs. 2-3**

**Graphical and Numerical Summaries of Data**

The purpose of this section of the activity is to practice producing and interpreting summaries of data in R. The types of summaries that you use will depend on the type of data you have and what you want to learn. Often graphical summaries are most easily digestible, though numerical summaries are helpful as well. As you’ve seen and read, we have different numerical and graphical summaries for categorical and quantitative data, and still different summaries for univariate versus bivariate or multivariate data.

**Summaries of Quantitative Data**

**Example #1 (Submit all graphics/summaries from console by copying into this document, as well as written interpretations requested):** In the RStudio website folder for today, I have placed two copies of the data so we can get more practice with various ways of getting data into R. One of the datasets is called summaries\_ex1.R, and the other has a .csv extension in place of the .R extension. You can load the .csv data file the same way we loaded data last week (File<Import Dataset<From Text (base)…) The load command is another way to bring a .R data file into R. Type

load(file.choose())

into the console, and a window will be launches where you can navigate to the .R datafile on the server. The data now exists in R, and you can see it in the Environment tab of the upper-right hand corner pane in RStudio (it’ll be called ex1).

The data are from a survey of students from PSU in 2006. The columns give data for each student’s estimates of (1) how many minutes they talk on a cell phone each week, (2) how many minutes they exercise each week, and (3) how much they spent on textbooks for that semester, and in column (4) they report their sex.

First, I want you to create a “dotplot” of the phone time usage variable. There are several ways to produce these plots in R, but the one that looks most like what we see in the text is from a package called BHH2 using the function dotPlot (remember R is case sensitive!). To do this and create the plot, use the following commands:

library(BHH2)

with(ex1,dotPlot(Phone)) ##with works like attach, but it is temporary!

##could also use dotPlot(ex1$Phone), or just attach(ex1) and then call dotPlot()

PASTE DOTPLOT HERE

In the resulting plot, the horizontal axis will give “phone time” values.” Each dot locates a student’s value. From the dotplot, about what was the longest amount of time reported by any student?

|  |
| --- |
| Ans: |

As before, we can use the command summary(ex1$Phone) to get the mean and the 5-number summary. If we want to know how many people were in the dataset, we can either find out how many rows were in the entire data frame with any one of:

dim(ex1) ##this returns 2 numbers: the number of rows first, and then number of columns

nrow(ex1) ##this just gives the number of rows

length(ex1$Phone) ##because this single column is no longer a data frame, use length

##dim would return NULL, because it is not the appropriate data structure

N = number of students in the dataset =

|  |
| --- |
| Ans: |

The mean time reported =

|  |
| --- |
| Ans: |

The median time reported =

|  |
| --- |
| Ans: |

**Using the 5-number summary, answer the following:**

a. In the context of this situation, write a sentence that interprets the median. HINT: What does this tell us about half of the students?

|  |
| --- |
| Ans: |

b. What is an interval that describes the lowermost 1/2 of the students’ phone time amounts?

|  |
| --- |
| Ans: (,) |

c. What is an interval that describes the lowermost 1/4 of the students’ phone time amounts?

|  |
| --- |
| Ans: (,) |

d. What is an interval that describes the uppermost 1/4 of the students’ phone time amounts?

|  |
| --- |
| Ans: (,) |

e. What is an interval that describes the middle 1/2 of the students’ phone time amounts?

|  |
| --- |
| Ans: (,) |

f. How do the mean and median compare? Why? How does this relate to the shape of the data as seen in the dotplot?

|  |
| --- |
| Ans: |

Now suppose we want to compare how male and female students respond to this question differently. We might want to separately obtain the number of males and females and get their respective means. Use the following code:

library(psych)

with(ex1,describeBy(Phone,Sex))

|  |  |  |
| --- | --- | --- |
|  | n | mean |
| Female |  |  |
| Male |  |  |

It would also be convenient to produce a nice graphical summary comparison. To create a side-by-side boxplot of phone time by sex, use:

with(ex1,boxplot(Phone~Sex))

PASTE BOXPLOT HERE

Comment on the salient features of this plot. What can you see? What did you learn about the comparison of men and women here?

|  |
| --- |
| Ans: |

**Example #2 (Submit all graphics/summaries from console, as well as written interpretations requested for parts a-f, j)**: In the RStudio website folder for today, you’ll find a dataset called summaries\_ex2.R, which you can load again using

load(file.choose())

and then navigating to the .R datafile on the server. The data now exists in R, and you can see it in the Environment tab of the upper-right hand corner pane in RStudio.

The data are from a Stat 200 class survey in the year 2002 (thus some features may be a bit behind the times.) The seven variables in the dataset are:

***Sex*** (male or female)

***EarPrces*** = total # of ear piercings in two ears combined

***Tattoo*** (yes or no to whether student has one)

***CDs*** = number of music CDs owned,

***Height*** = height in inches

***HtChoice =*** how tall the student would like to be in inches

***Friends*** *=* responseto “With whom is it easiest to make friends?” (Opposite sex, same sex, No difference)

**a.** Of the seven variables, which are categorical and which are quantitative?

Categorical =

|  |
| --- |
| Ans: |

Quantitative =

|  |
| --- |
| Ans: |

**b.** Create a histogram of the number of CDs owned, and describe the shape of the histogram (symmetric, skewed to the right, skewed to left, or bimodal?). Use the command

with(ex2, hist(CDs)) ##Those are some thick bars, maybe we need a finer grid

with(ex2, hist(CDs,breaks=seq(0,900,50))) ##better.

PASTE HISTOGRAM HERE

**c.** Obtain the following descriptive statistics for the ***CDs*** data, and give their values below using:

with(ex2,summary(CDs)) ##this gives everything but standard deviation

sd(ex2$CDs) ##There’s the standard deviation!

Mean = | Std dev =

Minimum = | Q1 =

Median = |Q3 = |Maximum =

**d.** Refer to the statistics found in part **c** and the histogram in part **b**. Explain why the mean is larger than the median for these data.

|  |
| --- |
| Ans: |

**e.** Refer to part **c.**

Calculate the range of the data.

|  |
| --- |
| Ans: |

Calculate the IQR (interquartile range) for the data.

|  |
| --- |
| Ans: |

What feature of the data is measured by the standard deviation?

|  |
| --- |
| Ans: |

Complete the following sentence. About ¼ of the students owned more than \_\_\_\_\_ CDs.

|  |
| --- |
| Ans: |

**f.** Let’s create a boxplot that compares the heights of men and women in the class using:

with(ex2,boxplot(Height~Sex))

PASTE BOXPLOT HERE

On the basis of the boxplot, would you say that the shape of height variable is symmetric or skewed.

|  |
| --- |
| Ans: |

What is the median for females?

|  |
| --- |
| Ans: |

What is the median height for males?

|  |
| --- |
| Ans: |

What is an interval of heights that describes the middle 50% of heights for females?

|  |
| --- |
| Ans: (,) |

What is an interval of heights that describes the middle 50% of heights for males?

|  |
| --- |
| Ans: (,) |

On the basis of the boxplot, explain whether the spread of the data (variability) is about the same or different for males and females.

|  |
| --- |
| Ans: |

**Some Graphical/Numerical Summaries of Categorical Data**

**g.** Let’s explore the Tattoo variable. Suppose I want to know both how many students answered the question about tattoos, and the percent who had a tattoo. I could do it with:

t=with(ex2,table(Tattoo))

n=sum(t)

pct=t/n

pct

How many students were in the dataset, and what percentage of the students had a tattoo?

|  |
| --- |
| Ans: |

**h.** Suppose I want to visualize this data. We might use a bar chart or a pie chart. If we do a bar chart, we can choose to either plot the raw frequencies (t) or the percentages (pct). Either way, we should be sure to label our axes:

barplot(t,ylab="Frequency")

barplot(pct,ylab="Percentage")

PASTE BARPLOTS HERE

We could also produce a pie chart. Check out the following:

pie(t) ##this is really basic

pie(t,labels=paste(paste(names(t),round(pct\*100)),"%",sep="")) ##a little bit nicer

library(plotrix)

pie3D(t[2:1],main="Pie Chart of Tattoo",labels=paste(paste(names(t)[2:1],round(pct[2:1]\*100)),"%",sep="")) ##fancy

PASTE PIE CHARTS HERE

**i.** Let’s explore the relationship between a student’s gender and with whom a student finds it easier to make friends by creating a two-way table:

t=with(ex3,table(Sex,Friends))

Ask R to print t and note that this produces the raw number of people that fall into each of the 6 categories. What we probably want is proportions of folks in each gender group who fall into the 3 different categories of the Friends variable. To do this use:

prop.table(t,1) ##The 1 requests the proportions across the first dimension, which are

##rows. Using a 2 would produce proportions of gender in a column.

What percent of the males said they find it easiest to make friends with the opposite sex?

|  |
| --- |
| Ans: |

What percent of the females said they find it easiest to make friends with the opposite sex?

|  |
| --- |
| Ans: |

Using the counts given in *t*, show how the percent of females who said “opposite sex” was calculated.

|  |
| --- |
| Ans: |

**j.** Refer to the output created in part **i.** and write two or three sentences that describe the differences between males and females in answering the question about with whom it’s easiest to be friends.

|  |
| --- |
| Ans: |

**z-scores and Empirical Rule**

**Example #3** Car and truck speeds at a particular location have approximately a bell-shaped distribution with mean = 65 mph and standard deviation = 5 mph.

*For parts a-c, uses the empirical rule to fill in the blanks in each part:*

**a.** About 68% of cars and trucks travel between \_\_\_\_\_\_\_ and \_\_\_\_\_\_\_ at this location.

**b.** About 95% of cars and trucks travel between \_\_\_\_\_\_ and \_\_\_\_\_\_\_\_ at this location.

**c.** About 99.7% of cars and trucks travel between \_\_\_\_\_\_ and \_\_\_\_\_\_\_\_ at this location.

**d.** A z-score is a measure of how many standard deviations a value is from the mean. Later in the course, we will learn that a z-score is an important measure of the size of a value.

The formula for a z-score is z = .

Determine a z-score for a vehicle speed of 72 mph.

Determine a z-score for a vehicle speed of 65 mph.

Determine a z-score for a vehicle speed of 60 mph. (Note: A value below the mean has a negative z-score.)

**e.** Complete the following two sentences.

The z-scores for about 68% of the vehicle speeds will be between \_\_\_\_\_\_\_ and \_\_\_\_\_\_\_ .

The z-scores for about 95% of the vehicle speeds will be between \_\_\_\_\_\_\_ and \_\_\_\_\_\_\_ .

**Example #4** In the RStudio website folder for today, you’ll find a dataset called summaries\_ex4.R, which you can load again using

load(file.choose())

and then navigating to the .R datafile on the server. The data now exists in R, and you can see it in the Environment tab of the upper-right hand corner pane in RStudio.

The data are from students in a statistics class at UC Davis. The variables are ***Sleep*** = hours of sleep the previous night, ***momheight*** = student’s guess at their mother’s height, and ***exercise*** = student’s self-reported hours of exercise in a typical week.

**a.** Draw a histogram of the ***Sleep*** variable (**with(ex2,hist(Sleep))**). Characterize the shape of the histogram (bell-shaped, symmetric but not bell-shaped, skewed, etc.)

|  |
| --- |
| Ans: |

**b.** Now suppose we want some basic summary information about the Sleep variable. Get the mean and standard deviation, and report them below (Reminder: functions mean() and sd() will produce these values).

|  |
| --- |
| Ans: |

Mean hours of sleep =

|  |
| --- |
| Ans: |

standard deviation =

**c.** Assuming that the empirical rule applies, calculate an interval that should include about 95% of the data values for the hours of sleep variable.

|  |
| --- |
| Ans: |

**d.** Draw a box plot of ***momheight*** using **with(ex2,boxplot(momheight))**. What noteworthy feature(s) of the data is indicated by the plot?

PASTE BOXPLOT HERE

|  |
| --- |
| Ans: |

**e.** Using this boxplot, what are the values of the quartiles and IQR?

|  |
| --- |
| Ans: Q1 = Q3 = IQR = |

**f.** A data value is marked as an outlier in a boxplot either if it is larger than Q3+(1.5×IQR) or smaller than Q1 −(1.5×IQR). For ***momheight***, calculate the two boundaries for marking outliers

|  |
| --- |
| Ans: Q3 + (1.5×IQR) = Q1 − (1.5×IQR) = |

**g.** Refer to part f. Would a **momheight** = 57 inches be marked as an outlier? Why or why not?

|  |
| --- |
| Ans: |

**h.** Suppose that we were to delete the two outliers from the ***momheight*** data. For each of the following statistics, briefly explain whether you think that the value of the statistic would change or not.

Standard deviation

|  |
| --- |
| Ans: |

Range

|  |
| --- |
| Ans: |

Median

|  |
| --- |
| Ans: |

**j.** Produce a histogram of the ***exercise*** variable. Characterize the shape of the histogram and explain whether the empirical rule would apply to this variable.

PASTE HISTOGRAM HERE

|  |
| --- |
| Ans: |

**Regression, Correlation and Scatterplots**

**Example #5** In the RStudio website folder for today, you’ll find a dataset called introReg\_ex5.R, which you can load again using

load(file.choose())

and then navigating to the .R datafile on the server. The data now exists in R, and you can see it in the Environment tab of the upper-right hand corner pane in RStudio.

For a statistics class project at a large northeastern university (Penn State), a student examined the relationship between *x =*  body weight (in pounds) and *y =*  time to chug a 12-ounce beverage (in seconds). The student collected data from 13 individuals.

**a.** Produce a scatterplot of the measurements. The y-variable is “chug time” and the x-variable is weight. In R, we can use:

plot(ex5) ##This works here as a special case because there are only 2 variables

##that happen to be in order: x first, y second.

with(ex5,plot(x=Weight,y=ChugTime)) ## This is more explicit!

with(ex5,plot(x=Weight,y=ChugTime, main='Relationship between Weight and Time to Chug a 12oz. Beverage')) ## How about a nice descriptive title…

with(ex5,plot(x=Weight,y=ChugTime, main='Relationship between Weight and \nTime to Chug a 12oz. Beverage')) ## Wanna split that title into 2 lines??

- Describe the main features of the graph. Specifically, is there a negative or a positive association? Does the pattern look to be linear or curved? Are there any outliers? If there is an outlier, describe where it’s located on the graph.

|  |
| --- |
| Ans: |

**b.** In general, outliers should not be thrown out unless there’s a good reason, but there are several reasons why it may be legitimate to conduct an analysis without them. In this case, let’s ignore the data point for the heaviest person and then determine a regression line for the remainder of the data.

ex5\_tmp=subset(ex5,ex5$Weight<max(ex5$Weight))

with(ex5\_tmp,summary(lm(ChugTime~Weight)))

Write the estimated regression equation (look in the Estimates column of the Coefficients table for the intercept and slope, in that order).

|  |
| --- |
| Ans: |

**c.** Write a sentence that interprets what this slope says about the relationship between chug time and body weight.

|  |
| --- |
| Ans: |

**d.** Use the equation found in part b to estimate the chug time for somebody who weighs 160 pounds.

You might try to write an R function here. For example:

myreg=function(x){

return(m\*x+b) ##**you fill in the slope (m) and the intercept (b)**

}

myreg(160) ##this calculates the prediction

|  |
| --- |
| Ans: |

**e.** A prediction error, also called a residual, is calculated as “actual y-value – predicted y-value.” Suppose that a person (not in the dataset) who weighs 160 pounds can do a chug time of 6.5 seconds. What is the value of the prediction error for this person? Note: You got the predicted chug time in part d.

6.5-myreg(160) ##boom.

|  |
| --- |
| Ans: |

**f.** What is the value of the prediction error (residual) for a person who weights 200 pounds and can do a chug time of 5.2 seconds?

#See what I did in (e).

|  |
| --- |
| Ans: |

**g.** What is the predicted chug time for a person who weighs 300 pounds? What is (obviously) invalid about this prediction? Note: This part is about the problem caused by “extrapolation,” which is predicting too far beyond the observed range of the data.

|  |
| --- |
| Ans: |

**Example #6** In the RStudio website folder for today, you’ll find a dataset called introReg\_ex6.R, which you can load again using

load(file.choose())

and then navigating to the .R datafile on the server. The data now exists in R, and you can see it in the Environment tab of the upper-right hand corner pane in RStudio.

The data are latitude and temperature data for 20 U.S. cities. ***Latitude*** is the geographic latitude of the city, ***JanTemp*** is the mean January temperature, ***AprTemp*** is the mean April temperature, and ***AugTemp*** is the mean August temperature.

**a.** Create a scatterplot as before to examine the connection between ***AugTemp*** (*y*-variable**)** and ***latitude*** (*x*-variable). Once you have made the plot, add the regression line.

with(ex6,plot(latitude,AugTemp))

abline(with(ex6,lm(AugTemp~latitude)))

Based on the resulting plot, answer the following questions:

Does it look like a straight line is a suitable description of the data, or do the data look to be curved?

|  |
| --- |
| Ans: |

Is the correlation between the two variables positive or is it negative?

|  |
| --- |
| Ans: |

Are there any outliers? If so, which city (or cities) is an outlier? Explain why.

|  |
| --- |
| Ans: |

**b.** What is the equation of the regression line? Write the equation that relates August temperature to latitude.

summary(with(ex6,lm(AugTemp~latitude)))

|  |
| --- |
| Ans: |

**c.** Write a complete sentence that describes how much mean August temperature changes per each one degree increase in latitude.

|  |
| --- |
| Ans: |

**d.** In the output that you generated for part (b), you obtained a value for “Multiple R-squared:”. What is that number?

|  |
| --- |
| Ans: |

**e.** The R2 value is the squared correlation value, written as a percent. It is interpreted as the fraction of the observed variation in y-values that can be explained by the x-variable. Apply that interpretation to this situation by writing a sentence that interprets the R2 value for the relationship between August temperature and latitude. The sentence structure might be something like “ \_\_\_\_ percent of the observed variation in \_\_\_\_ is explained by \_\_\_\_\_\_.”

|  |
| --- |
| Ans: |

**f.** Use R to find the correlation between ***AugTemp*** and ***latitude***.

with(ex6,cor(AugTemp,latitude))

with(ex6,cor(latitude,AugTemp)) ##Are these different??

Give the numerical value of the correlation.

|  |
| --- |
| Ans: |

**g.** Square the value of the correlation that you found in part f.

with(ex6,cor(AugTemp,latitude))^2

Then compare the squared value to the R2 value that you reported in part d. How do they compare?

|  |
| --- |
| Ans: |

**h.** The equator has latitude = 0. Based on the regression equation found in part b, determine the predicted mean August temperature at the equator.

|  |
| --- |
| Ans: |

**i.** Refer to your answer to part h. Explain why this is probably a bad estimate of the mean August temperature at the equator. Hint: What is the interval of observed latitudes in the dataset?

|  |
| --- |
| Ans: |

**j.** Create a scatterplot to examine the connection between ***AugTemp*** (*y*-variable**)** and ***AprTemp*** (*x*-variable), and add the regression line as before.

with(ex6,plot(AprTemp,AugTemp))

abline(with(ex6,lm(AugTemp~AprTemp)))

Based on the resulting plot, it would be reasonable to look at the regression equation in:

summary(with(ex6,lm(AugTemp~latitude)))

Using this information, answer these questions:

**k.** Write a complete sentence that describes how much mean August temperature changes per each one degree increase in April Temperature.

|  |
| --- |
| Ans: |

**l.** Find the correlation between ***AugTemp*** and ***AprTemp*** (see part f)***.*** What is the correlation value?

|  |
| --- |
| Ans: |

**m.** Which variable, latitude or April temperature, is a stronger predictor of August temperatures? Explain why you think this.

|  |
| --- |
| Ans: |

**Example #7** In the RStudio website folder for today, you’ll find a dataset called introReg\_ex7.R, which you can load again using

load(file.choose())

and then navigating to the .R datafile on the server. The data now exists in R, and you can see it in the Environment tab of the upper-right hand corner pane in RStudio.

The data were collected in Stat 100 at Penn State and were used in some examples in the text. This activity shows how a third variable (confounding variable) can affect a correlation between two other variables.

**a.** Find the correlation between ***Fastest*** and ***RtSpan***. ***Fastest =*** self-reported fastest speed student has ever driven a car and ***RtSpan*** = student’s stretched right hand span (centimeters). What is the correlation value?

with(ex7,cor(Fastest,RtSpan)) ##What’s going on?? NA??

##We have some missing data, so we have to tell R to only use **com**plete cases

with(ex7,cor(Fastest,RtSpan,use="com")) ##Better.

|  |
| --- |
| Ans: |

Explain why the correlation value indicates that there is moderate positive association between fastest speed ever driven and stretched right hand span.

|  |
| --- |
| Ans: |

**b.** Now let’s visualize this relationship. Make the scatterplot using:

with(ex7,plot(RtSpan,Fastest)) ##Does being taller make you drive faster??

PASTE SCATTERPLOT HERE

But we know that hand span is probably just a proxy for general body size, and that men tend to have larger bodies than women. So let’s explore the role of gender here. Create this plot and comment on the main feature you notice:

with(ex7,plot(RtSpan,Fastest,col=Sex)) ##Men will be colored red, women will be black

PASTE SCATTERPLOT HERE

|  |
| --- |
| Ans: Main feature = |

**c.** In this part, let’s separately estimate this correlation for males and females.

Ex7M=subset(ex7,ex7$Sex=='M')

Ex7F=subset(ex7,ex7$Sex=='F')

with(ex7M,cor(RtSpan,Fastest,use="com"))

with(ex7F,cor(RtSpan,Fastest,use="com"))

|  |
| --- |
| Ans: |

Correlation for males only =

|  |
| --- |
| Ans: |

Correlation for females only =

Explain why these two correlation values indicate that, for each sex, there is a very weak relationship between fastest ever driven and right hand span.

|  |
| --- |
| Ans: Main feature = |

Explain how we can see that sex is a confounding variable in this problem.

|  |
| --- |
| Ans: |