**Graphical and Numerical Summaries of Data**

The purpose of this 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**: Seventeen students were asked to estimate the population of Canada (in millions). Ten students (group 1) were first told that the population of the U.S. was about 290 million at that time. The other seven students (group 2) are told that the population of Australia was about 18 million at that time. The idea is that the information given about either the U.S. or Australia population will influence student guesses. The resulting data (estimates of Canada’s population in millions) are:

Group 1: 135, 70, 190, 20, 30, 150, 35, 100, 120, 200

Group 2: 35, 100, 50, 20, 5, 45, 55

I want you to determine the median for each of the two groups by hand. Recall that we always must put the data in order first. You can use R to help with this:

grp1=c(135, 70, 190, 20, 30, 150, 35, 100, 120, 200)

grp2=c(35, 100, 50, 20, 5, 45, 55)

sort(grp1)

sort(grp2)

|  |  |
| --- | --- |
| Median for group 1 | Median for group 2 |
|  |  |

The lower quartile (or first quartile, or 25th percentile, or Q1) for a sample is defined in our textbook to be the median of the data values that are smaller than the median. The upper quartile (or third quartile, or 75th percentile, or Q3) is defined to be the median of the data values that are larger than the median. So, for example, let’s find Q1 for Group 2. When you ran the four lines of code above, you found the sorted list of estimates in the second group to be 5 20 35 45 50 55 100. Thus, the median is 45. So, according to our definitions, Q1 is the median of 5 20 35, which is obviously 20. Find the remaining values to fill in the chart below:

|  |  |  |
| --- | --- | --- |
|  | Q1 | Q3 |
| Group 1 |  |  |
| Group 2 | **20** |  |

Now let’s have the software do these calculations for you. In R, simply enter:

summary(grp1)

summary(grp2)

In addition to the mean of the data, this will yield the so-called 5-number summary, which consists of the minimum value, Q1, the median, Q3, and the maximum value. I hope you find that the medians you calculated by hand are the same as those reported by the software, but you may find different values for the first and third quartiles. Many algorithms exist for the calculation of percentiles, and different software packages give different answers. If you want to learn more about this, tell R to show you the help files for its quantile function by typing:

?quantile

In the resulting window, scroll down to see the formulas for the (nine!!) different types of percentiles it will calculate.

Report the values from the software here:

|  |  |  |
| --- | --- | --- |
|  | Q1 | Q3 |
| Group 1 |  |  |
| Group 2 |  |  |

**Example #2:** At the course website, access the folder for today’s lab. I have placed two copies of the data in this folder, so we can get more practice with various ways of getting data into R. One of the datasets is called summaries\_ex2.R, and the other has a .csv extension in place of the .R extension. On a Mac, you can double-click on this .R file directly. If R is not open, it will launch it and produce a line of code, and if it is already open, look in the console and see that it produced a line of code for you with the load command. The load command is the other way to bring a .R data file into R. On a Windows machine, you must use the load command. First, download the data (say to your Desktop), and then change the working directory of R to the location where you downloaded the data (File<Change dir…<). Then simply use

load("summaries\_ex2.R")

The data now exists in R, and to see what it is called, use the ls() command to get a list of all objects currently in R. For the .csv file, you can use:

ex2=read.csv("summaries\_ex2.csv")

assuming you saved the .csv file in the same location as the .R file.

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. You’ll have to download and install this package to gain access to the function dotPlot (remember R is case sensitive!). When you use the install.packages command below, R asks which server you want to use to grab the package – it doesn’t really matter which one you choose, but let’s scroll down to the USA (PA 1) option, which will grab it from Carnegie Mellon’s server. Use:

install.packages("BHH2") ##choose USA(PA 1) CRAN server!

library(BHH2)

with(ex2,dotPlot(Phone)) ##with works like attach, but it is temporary!

##could also use dotPlot(ex2$Phone), or just attach(ex2) and then call dotPlot()

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? \_\_\_\_\_\_\_\_\_

About what amount of time was the most commonly reported amount? \_\_\_\_\_\_\_\_\_\_\_

As before, we can use the command summary(ex2$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(ex2) ##this returns 2 numbers: the number of rows first, and then number of columns

nrow(ex2) ##this just gives the number of rows

length(ex2$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 = \_\_\_\_\_\_\_\_\_

The mean time reported = \_\_\_\_\_\_\_

The median time reported = \_\_\_\_\_\_\_

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

a. In the context of this situation, write a sentence that interprets the median.

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

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

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

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

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

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:

install.packages('psych')

library(psych)

with(ex2,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(ex2,boxplot(Phone~Sex))

Comment on the salient features of this plot. What can you see? What did you learn?

**Example #3:** At the course website, access the folder for today’s lab, where again I have placed a dataset called summaries\_ex3.R, which, on a Mac, you can double-click directly to load it into R (with the load command produced in the console). On a Windows machine, use the load command again (if the data are saved in the same location). The data are estimates of the percentage of adults who smoke in each state of the U.S. (and also the District of Columbia).

Inspect the data to find the percentage of adults who smoke in Pennsylvania. What is the value?

Create a dotplot of the smoking percentages (with(ex3,dotPlot(smoke\_pct))). Use it to answer the following questions:

About where do most states fall, in terms of the percentage that smokes?

About what is lowest percentage in the dataset?

About what is the highest percentage in the dataset?

Identify the states with the highest and lowest percentages and give the percentage smoking in each state. You could manually inspect this particular dataset because it is small, but there are several alternatives that are easier in general for larger datasets. For example, you could do:

n=nrow(ex3)

ex3\_sort=ex3[order(ex3$smoke\_pct),] ##this sorts the entire data frame by the values of

##smoke\_pct

ex3\_sort[c(1,n),] ##this asks for the first and last (e.g. smallest and largest) entries in

##sorted data frame

Some other fun things you can do to peruse the dataset:

ex3\_sort[c(1:5,(n-4):n),] ##asks for 5 smallest and 5 largest

ex3\_sort[ex3\_sort[,2]<20,] ##will show all states with smoke\_pct < 20

Now, determine the 5-number summary for the state smoking percents.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Minimum | Q1 | Median | Q3 | Maximum |
|  |  |  |  |  |

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

a. What is an interval that describes the uppermost 1/2 of the states’ smoking percentages?

b. What is an interval that describes the uppermost 1/4 of the states’ smoking percentages?

c. What is an interval that describes the lowermost 1/4 of the states’ smoking percentages?

**Example #4**: At the course website, access the folder for today’s lab, where again I have placed a dataset called summaries\_ex4.R, which, on a Mac, you can double-click directly to load it into R (with the load command produced in the console). On a Windows machine, use the load command again (if the data are saved in the same location). 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 =

Quantitative =

**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(ex4, hist(CDs)) ##Those are some thick bars, maybe we need a finer grid

with(ex4, hist(CDs,breaks=seq(0,900,50))) ##better.

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

with(ex4,summary(CDs)) ##this gives everything but standard deviation

sd(ex4$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.

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

Calculate the range of the data.

Calculate the IQR (interquartile range) for the data.

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

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

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

with(ex4,boxplot(Height~Sex))

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

What is the median for females?

What is the median height for males?

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

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

.

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

**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(ex4,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?

**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")

dev.new()

barplot(pct,ylab="Percentage")

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

install.packages('plotrix')

library(plotrix)

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

**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(ex4,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?

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

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

**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.