**Graphical and Numerical Summaries of Data**

Today we 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, and made more powerful by accompanying numerical summaries.

**1. Summaries of Quantitative Data**

**1 a) First Example**: Our class survey asks you to estimate the population of Canada (in millions). A random selection of students (group 1) were first told that the population of the U.S. is about 320 million; the other students (group 2) were told that the population of Australia is about 23 million. The resulting data—your estimates of Canada’s population in millions—from the first twenty responses are:

Group 1: 100, 75, 300, 150, 40, 250, 50, 24, 50, 140, 35, 150

Group 2: 40, 15, 15, 35, 15, 20, 16, 17

Question/Task: Determine the median for each of the two groups by hand, which will be easier if you ask R to put these data in order.

>grp1=c(100, 75, 300, 150, 40, 250, 50, 24, 50, 140, 35, 150)

>grp2=c(40, 15, 15, 35, 15, 20, 16, 17)

>sort(grp1)

>sort(grp2)

|  |  |
| --- | --- |
| Median of group 1 | Median of 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.

Question/Task: Determine the five number summary of these data, by hand, and complete this chart.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Min | Q1 | Median | Q3 | Max |
| Group 1 |  |  |  |  |  |
| Group 2 |  |  |  |  |  |

“Anchoring” occurs when a given value is introduced in advance of someone being asked to estimate a quantity. The anchoring bias can be significant even in settings where the initial value has nothing to do with the following estimate. Does it appear that mentioning the US population or the Australian population in advance of asking about the population of Canada has had an anchoring effect on the responses?

Now let’s have the software do these calculations for you. Enter these two commands in R.

>summary(grp1)

>summary(grp2)

In addition to the mean of the data, this will yield the five number summary. You should 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.

Report the values from the software here:

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

Many algorithms exist for the calculation of percentiles, and different software packages give different answers. (We talked about this a little bit in class on Friday.) 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. While percentiles and quartiles are different, the underlying ambiguity is the same.

**1 b) Second Example:** I have placed two datasets on our Moodle page, including a new file format so that 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. You can import an .R data file using the “load” command. First, download the data to your laptop/desktop. You can then hunt for the file among your directories using

>load(file.choose())

The alternate method is to change the working directory of R to the location where you downloaded the data (File<Change dir…<). Then use

>load("summaries\_ex2.R")

Whichever method you choose, 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.

>ls()

Comma Separated Values (.csv) is one of the most common file formats for data sets. To import the .csv version of the data you can use the “read.csv” command using either …

>ex2=read.csv(file.choose())

or

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

depending on whether you like to track your working director or hunt for files.

The data are from a survey of Penn State students 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. If we want to know how many observations are in the dataset, we can use any one of:

>dim(ex2)

which this returns 2 numbers: the number of rows first, and then number of columns. You can also use

>nrow(ex2)

which gives the number of rows, or

>length(ex2$Phone)

As before, we can use the command “summary(ex2$Phone)” to get the mean and the 5-number summary for the variable describing the number of minutes the individuals talk on a cell phone each week. You can also create a histogram and boxplot of these data using “hist(ex2$Phone)” and “boxplot(ex2$Phone)”.

Question/Task: Use the tools that R makes available to you to determine:

1. N = number of students in the dataset
2. The mean time of the reported cell phone weekly usage
3. The median time reported
4. The interval that describes the lowermost 1/2 of the students’ phone time amounts?
5. The interval that describes the lowermost 1/4 of the students’ phone time amounts?
6. The interval that describes the uppermost 1/4 of the students’ phone time amounts?
7. The interval that describes the middle 1/2 of the students’ phone time amounts?

Finally, use the shape of the data to explain the difference between the mean and median.

Suppose we want to compare how male and female students respond to the question on phone usage. We might want to separately obtain the number of males and females and get their respective means. This can be done with the following code:

>install.packages('psych')

>library(psych)

>with(ex2,describeBy(Phone,Sex))

or

>describeBy(ex2$Phone, ex2$Sex)

The response from R will contain quite a bit of summary information. Find the following statistics among those items.

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

It is also helpful to produce a graphical summary comparison. To create a side-by-side boxplot of phone time by sex, use:

>boxplot(ex2$Phone~ex2$Sex))

Question/Task: Do the summary statistics and box-and-whisker diagrams indicate that the reported phone usage is different for women and men?

**1 c) Third Example:** The third data set we will explore comes from the class survey. The file ClassHeight.csv on our Moodle page contains responses from the questions about sex, class year, and height. Import these data into R.

The height data involves two columns, one for feet and the other for inches. In order for us to understand the distribution of heights among these observations, we need to convert these two columns into a single column containing each individual’s height in inches. A person who is 5’8” is 68 inches tall because . To apply this operation across the “Feet” and “Inches” columns, you can use

>Height=12\*ClassHeight$Feet + ClassHeight$Inches

We then want to bind this column of heights to our original matrix of values, which we can do using a column binding command

>CH=cbind(ClassHeight,Height)

The matrix CH now contains all of the information that is in ClassHeight, with an additional column containing the reported heights of students in inches. You can learn the mean and five number summary for the heights with

>summary(CH$Height)

You should find that the mean and median heights are not that dissimilar. R will also compute the standard deviation for the heights, with

>sd(CS2$Height)

Question/Task: Look at the height data using the “boxplot” and “hist” functions in R.

1. Is the distribution of heights approximately symmetric or highly skewed?
2. Use the shape of the data to explain why the mean and median heights are fairly close to each other.

**2. Summaries of Categorical Data**

In this final section we continue to use the CH dataset you created, in order to illustrate how R can help visualize the distribution of categorical data. The variables “Sex” and “Class” in this dataset are both categorical variables. As “Class” contains more options, it is a more interesting variable to use in exploring the use of R.

The “table” function in R creates tabular results of a categorical variable. As examples,

>table(CH$Sex)

tells you how many females and how many males completed the survey. Similarly,

>table(CH$Class)

tells you how many respondents reported being in which class.

You can combine the table function with “pie” to produce a pie chart.

>pie(table(CH$Class))

You can also combine the table function with “barplot” to create a bar chart of the variable Class.

>barplot(table(CH$Class))

There are many ways in which you can spruce up this basic bar graph. You can introduce labels on the axes (“ylab” and “xlab”); a title to the bar graph (“main”); and introduce colors (“col”). As one example, type and enter

>barplot(table(CH$Class), main=“Enrollment in Math 186 by Class Year”, xlab=“Class Year”, ylab=“Number of Students”, col=c("maroon1","maroon2","maroon3","maroon4"))

You might also want to be able to visualize the students in each class year, divided by sex. To do this, first create a table with the relevant counts.

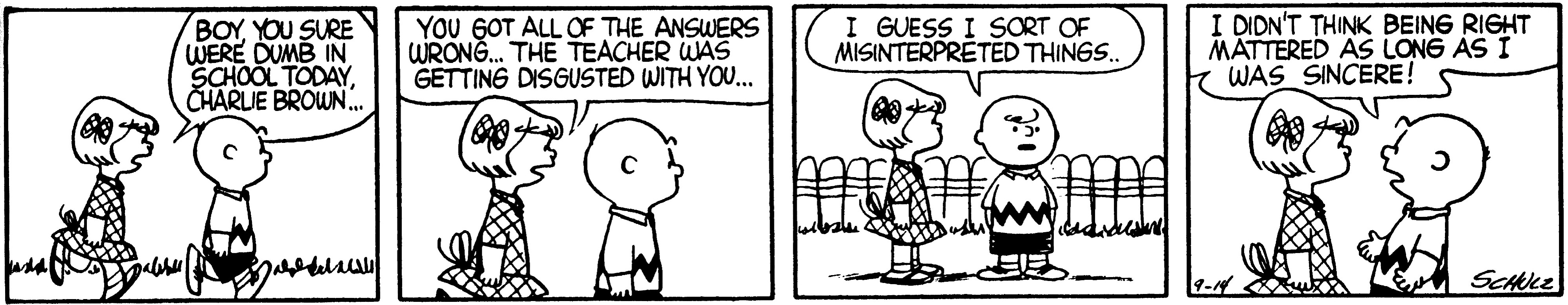
>Both=table(CH$Sex,CH$Class)

And then create the bar graph

>barplot(Both, beside=TRUE)

Question/Task: Figure out how to color the bars corresponding to female students one color, while coloring the bars corresponding to male students with another. A few of the allowed color names in R are: blue, cyan, green, orange, orchid, and red.

Note for future reference: In R a categorical variable is called a *factor*.

****

**Assignment!** You need to write-up two analyses for this week’s laboratory.

1. (5 pts) Use the data in ClassHeight, and your extension of it CH, to determine if there is an appreciable difference in the heights of female and male students in Math 186. Then write a paragraph that argues for your conclusion. Your argument should reference five number summaries and means. It should also include box-and-whisker diagrams or histograms (but not both!).

2. (5 pts) There is one more dataset on Moodle that I want you to look at. Use the five number summary, means and standard deviations, and histograms to compare the two groups in Lab2. Then write a brilliant paragraph explaining why summary statistics are not sufficient for the examination of a distribution. If you are having trouble distinguishing the difference between the two distributions, you may have to use the “breaks” option within the histogram function.

Your solutions should be typed, and you should hand me a single document that contains your answers to both questions. Please don’t forget to put your name on it!

**This assignment is due in lab on Monday, 12 September. Don’t email me your solutions.**