Worksheet 2

Author:

Discussants:

==========

In the following worksheet you will explore more categorical and quantitative data using data. Please submit a compiled pdf with your answers to the exercises and Lock 5 questions to Moodle *by 11:59pm on Sunday September 23rd*, and be sure to answer all the questions that are posed in the different exercise steps!

Some useful functions for completing this worksheet are: dim(), length(), table(), prop.table(), barplot(), hist(), mean(), median(), max(), fivenum(), and boxplot(). You might also find the following symbols useful: μ, $\bar{x}$, π, $\hat{p}$

R exercises 1: The code below loads from 5 different brands (Toyota Corollas, Subaru Imprezas, Honda Civics, Hyundai Elantras and the Mazda 3s) in a data frame called car_data. The data comes from Edmunds.com which is a website that helps people shop for cars, and this data was used in the 2015 5 Colleges DataFest.

Start by using the dim() function on the data frame to report how many cases are in the data frame, and report what the variables are in the data frame and whether they are categorical or quantitative. Also report what the population for this data is. Hint: to view the data frame you can load the data on the console and then use the View() on the console as well. Remember that you can't use the View() function in an R Markdown document and also that your R Markdown document does not have access to the objects in your R environment but can only access objects that are created inside the R Markdown document itself.

```
# do no change the lines below
load("/home/shared/intro_stats/cs206_data/cars_small.Rda")
```

Answer:

Step 1a: In the first set of exercises we will get just a little more practice exploring categorical data. Let's start by extracting a vector that has the names of the car brand for each car sold and assigning this data to an object called brands (hint: use the $ symbol). The use the table() function on the brands vector to create a count of the number of different cars sold for each brand and assign the results to an object called bran_counts. Report which car brand sold the most cars.

Answer:

Step 1b: Use the prop.table() function on the brand_counts to get the relative frequencies of car brands sold. What proportion of cars listed were from the brand that sold the most cars? Which brand has the second most cars sold?

Answer:

Step 1c: Use the barplot() function on brand_counts to create a bar chart of the brand counts. Suppose we had a new data set from 2016, do you think the proportion of cars sold by the top selling brand would change? Why or why not?

Answers:

Step 1d: Use the pie() function on brand_counts to create a pie chart of the brand counts. Which plot do you think is the most informative, the bar chart or the pie chart? Explain your reason for preferring one type of plot over the other.

Answers:

R exercise 2: In thes next exercises we will examine quantiative data by examining the prices of Subaru Imprezas. The code below creates a vector called subaru_prices contains the prices for the Subaru Impreza's that were sold. Use the length() function to verify that the number of subaru's in this vector matches the number that you found when you ran the table() function above. Note: the length() function tells you how many elements are in a *vector* while the dim() function tells you how many rows and columns are in a *data frame*, and since we are finding the number of elements in a vector we are using the length() function here.

```
# do not change the lines below
subaru_prices <- subset(car_data$price, car_data$brand == "Subaru")
```

Step 2a: Use the mean() and median() functions on the subaru_prices. Is the mean or the median higher? Does this indicate that the data is left or right skewed?

Answers:

Step 2b: Use the hist(my_vector, n) function to create a histogram of the prices for the Subarus. Try using different values for the n argument to create different number of bins in the histogram and be sure to label your axes appropriately. Then describe the shape of the data, and whether there seem to be any noticable outliers.

Answers:

Bonus (optional) question: The full Edumunds transaction data set can be loaded by uncommenting the code below. Load the data set and see if you can create plots showing interesting trends in this data.

```
# load('/home/shared/data/edmunds/edmunds_transaction_data2.Rda')
# # load the data prices <- transaction.data2$price_bought #
# get a vector that has the prices paid for different cars
```

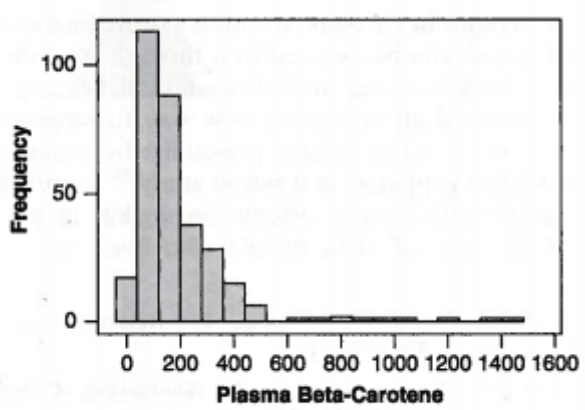

Figure 2.15 ***Concentration of beta-carotene in the blood***

Figure 1: concentration of beta-carotene in blood

Please also complete the following Lock5 exercises

Lock5 exercise 2.8: Of all 1,547,990 students who took the SATs (Scholastic Aptitude Test), 1,114,273 were from a public high school. Use these numbers to calculate the proportion of students who were from public school and use the correct symbol to denote this proportion (hint: some possible symbols to use are: μ, $\bar{x}$, π and $\hat{p}$).

Answers:

Lock5 exercise 2.44: A set of lucky numbers are: 41, 53, 38, 32, 115, 47, and 5. For these lucky numbers find: a) the mean $\bar{x}$, b) the median m, and c) indicate whether there appear to be any outliers and if so, what they are.

Answers:

(a)

(b)

(c)

Lock5 exercise 2.58: The plasma beta-carotene level (concentration of beta-carotene in the blood), in ng/ml was measured for a sample of n = 315 individuals. A histogram of this sample is shown below (in figure 2.15). Please a) Describe the shape of this distribution and if there are any obvious outliers. b) Estimate the median of this sample, c) estimate the mean of this sample.

Answers:

(a)

(b)

(c)

Reflection

How are you feeling so far with R and the concepts? Does the class feel like it is going too fast or too slow? Could the jokes in class be better? Please briefly reflect below on how you feel and if you have any remaining questions - and if so, bring them up at the start of next class!