Worksheet 4

Author:

Discussants:

==========

The following worksheet will give you practice examining relationships between quantitative data and reviewing the material we have covered in class so far. Please answers the following questions and submit a compiled pdf with your answers to Moodle by 11:59pm on Sunday October 7nd.

A list of functions we have used in class can be found on a shared class google document here

Part 1: A few Lock 5 exercises looking at scatter plots and correlation

Lock5 exercise 2.158 (1st edition): Please describe whether you expect a positive or negative assocation between these two quantitative variables: The *distance driven* since the last fill-up of the gas tank, and the *amount of gas* left in the tank

Answer:

Lock5 exercise 2.174: Ages of Husbands and Wives: Suppose we record the husband's age and the wife's age for many randomly selected heterosexual married couples. Please answer the following questions:

a) What would it mean about ages of couples if these two variables had a negative relationship?

b) What would it mean about ages of couples if these two varaibles had a positive relationship?

c) Which do you think is more likely, a negative or a positive relationship?

d) Do you expect a strong or weak relationship in teh data? Why?

e) Would a strong correlation imply tehre is an association between husband age and wife age?

Answer:

Lock5 exercise 2.164 and 2.166: The table below shows the relationship between two quantiative variables X and Y (you might want to knit the document to see the table better). Please use R to create a scatter plot and calculate the correlation between these variables. Hint: the function to create vectors, c(), might be useful here. Be sure to label your axes as well! *Bonus points:* try to calculate the correlation without using the cor() function.

X	15	20	25	30	35	40	45	50
Y	532	466	478	320	303	349	275	221

** Lock5 exercise 2.196 (1st edition) The Honeybee Waggle Dance**: When honeybee scouts find a food source or a nice site for a new home, they communicate the location to the rest of the swarm by doing a "waggle dance". They point in the direction of the site and dance longer for sites farther away. The rest of the bees use the duration of the dance to predict distance to the site. The code below loads a data frame called HoneybeeWaggle that contains the distance (in meters), and the duration of the dance (in second) for seven honeybee scouts. Using this data:

a) Create a scatter plot of the data. Also describe whether there appears to be a linear trend in the data. If so, is it positive or negative?

b) Use the cor() function to find the correlation between the two variables. Hint, you can extract the vector of distances from the data frame using the syntax: HoneybeeWaggle$Distance

```
# download the data and load it into R
library(Lock5Data)
data(HoneybeeWaggle)
```

Answers: a) b)

Part 2: A review of R data analysis methods covered so far

Below you will review R functions we have used in class to analyze data. A list of functions we have used can be found on a shared class google document here.

Exercise 2a: To review the R code we have covered in class so far we will explore information about salaries and endowments at liberal arts colleges. The code below loads a data frame called *college_salaries* that has information from the 2016/2017 academic year. There is also code that creates a data frame called *assistant_salaries* extracts information related to only professors at the the academic rank of assisant professor. To get started, for both these data frames, please report how many cases and variables are in each data frame and what the cases correspond to.

```
# load the college salaries data frame
load("/home/shared/intro_stats/cs206_data/college_salaries.Rda")


# create the assistant_salaries data frame
assistant_salaries <- subset(college_salaries, rank == "Assistant")
```

Answer:

Exercise 2b - analyzing categorical data: Using the *college_salaries* data frame, print out a table showing how many professors are at the different ranks and report the rank that has the most professors in it. Also create a bar plot and a pie chart of the results. Which of these plots do you think is most informative here?

Answer:

Exercise 2c -quantitative data: Now let's explore the salaries of assistant professors using the *assistant_salaries* data frame. Using the salaries_tot variable (which has the salaries regardless of gender) compute the mean, standard deviation and five number summary, and also plot a histogram and a boxplot of the

data. Then describe what the shape of the data looks like. Also below is code that extracts the salaries of Hampshire professors. Describe how well Hampshire professors are doing relative to professors at other colleges?

```
# get the mean, standard deviation and five number summary,
# and also plot a histogram and boxplot




# extracting data for Hampshire College
Hampshire_data <- subset(assistant_salaries, school == "Hampshire College")
```

Answer:

Exercise 2d - comparing salaries across genders: One question of interest is whether women assistant professors make less than men. Try to create a few plots and calculate a few that show as clearly as possible whether this is the case for this data using the *assistant_salaries* data frame. What would you conclude based on this data?

Answer:

Exercise 2e - relationships between quantitative data: Do college that pay men more also pay women more? Create a plot and calculate a statistic to examine this using the *assistant_salaries* data frame. Also do this for the relationship between the endowment size (or log endowment size) and total salary. How strong of an association is there?

Answer:

Part 3: Exploring data on your own

Please find an interesting data sets in the packages listed on this website. All these packages are already on asterius and can be loaded on asterius, so once you have found an interesting data set you can load it using the library() function and make the data visable using the data() function.

Below shows an example of loading the data from a data set that contains information about the body temperatures of beavers. Please delete the code below and replace it with code that loads data that you find interseting. Then create 2 plots and report 2 statistics that you find interesting and explain why you found them interesting. Also, if you'd prefer to use your own data, you can upload it to asterius using the upload button under the file table in the lower right section of R Studio, and then you can load the data using the read.table() function (see google for more details).

```
# delete the code below and replace it with some data you
# find interesting from the data on the website:
# https://vincentarelbundock.github.io/Rdatasets/datasets.html


# load the boot package that has information about beavers
library(boot)
```

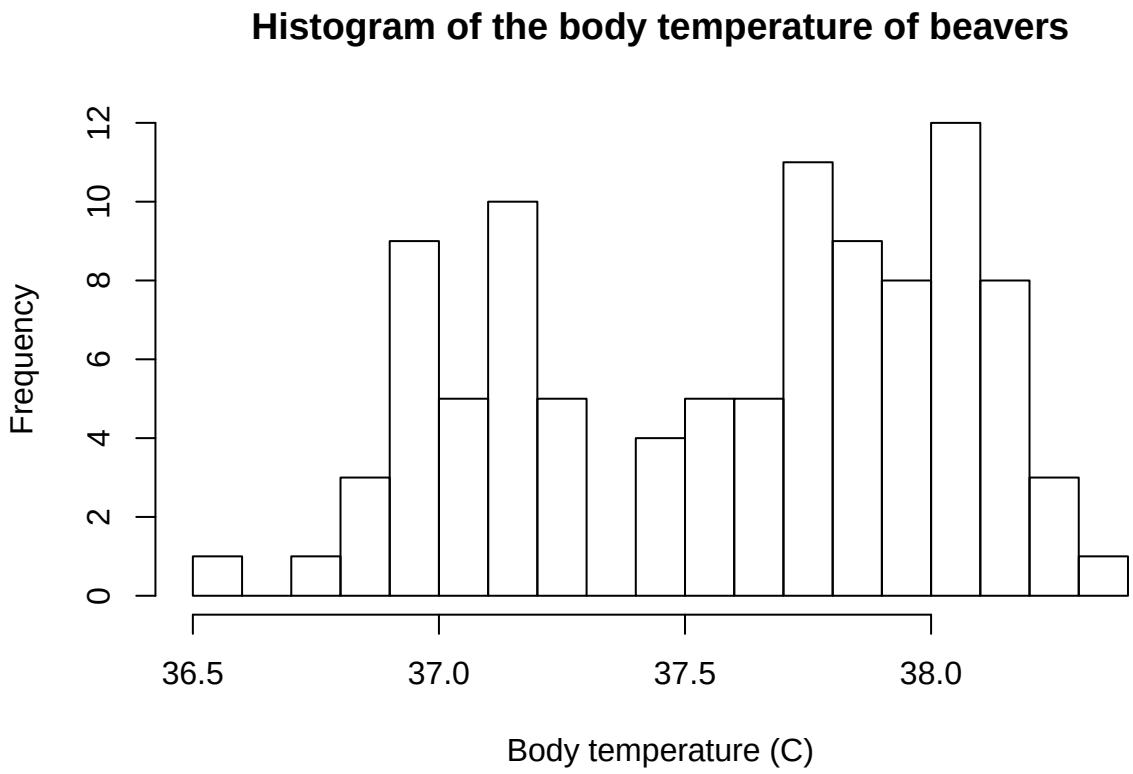
```r
# get the data
data(beaver)

# you can find information about the beaver data frame by
# running the ? function outside of R Markdown: ? beaver

beaver <- data.frame(beaver)  # careful, some of the data sets are not data frames

# you can use the View(beaver) to see the data but only
# outside of R Markdown

# plot a histogram of the beaver body temperatures
hist(beaver$temp, n = 20, xlab = "Body temperature (C)", main = "Histogram of the body temperature of b
```



Answer: What *interesting* things did you find?

Reflection

How did this worksheet go, and now that we are at the October break how is the class going more generally? If there are any concepts you are not clear about please try to review them over break because we will be moving on to statistical inference after the break which builds on what we have already covered. Also, if you want to give me some feedback through midterm self-evaluations on theHub that would be useful too.