Lab 1: Descriptive Statistics

due date: Wednesday, September 4 in class

Introduction

Below are some data on violent crime in US States during the year 1973. This data set comes from R, a free statistics software package widely used by statisticians.

You are asked to compute some summary statistics, create boxplots and histograms, and answer a few questions. All of this work should be done with R or RStudio. (Part of your grade is simply to complete this lab using the software package R.) Information about the website where you can download R and some basic R commands have been included at the end of this handout to help. After starting RStudio and loading the data set USArrests, type help(USArrests) to get some information about this dataset.

For submission to your instructor, write your answers in the space provided and hand in graphs.

Data

```
Violent Crime Rates by US State

Description:

     This data set contains statistics, in arrests per 100,000
     residents for assault, murder, and rape in each of the 50 US
     states in 1973. Also given is the percent of the population living
     in urban areas.
```

Questions

1. What is the population under study?

   The population under study is the 100,000 residents in each state during 1973.

2. Compute the Min, Q1, Median, Q3, Max for the numerical variables Murder and Rape. Then compute the mean for these two variables. Is the mean bigger than, smaller than, or roughly equal to the median? What does your answer here tell you about the variables Murder and Rape.

$$Murder, \begin{bmatrix} Min. & 1stQu. & Median & 3rdQu. & Max. \\ 0.800 & 4.075 & 7.250 & 11.250 & 17.400 \end{bmatrix} \tag{1}$$

$$Rape, \begin{bmatrix} Min. & 1stQu. & Median & 3rdQu. & Max. \\ 7.30 & 415.07 & 20.10 & 26.18 & 46.00 \end{bmatrix} \tag{2}$$

$$Mean \begin{bmatrix} Murder & Rape \\ 7.788 & 21.23 \end{bmatrix} \tag{3}$$

   Comparing the mean of these two variable it becomes clear that in the U.S, rape occurs approximately 3 times as often. In comparison to the median for both variables the mean is a little larger but hardly so, this tells us that there are hardly no outliers if there are any at all.

3. Compute boxplots for the variables Murder and Rape. Summarize what you learn from viewing the boxplots for the two variables. Plots of these boxplots should be handed in with your lab.

   The boxplot for Murder is rather short, which means that there isn't a great variation from state to state in the number of incidents. However the boxplot for Rape is the opposite, it's a lot taller and even has outliers(Alaska, and Nevada). In terms of distribution both boxplots are similar, they are both relatively even below and above the median

4. Looking at the data for Alaska, how does it compare to other states? Do you think Alaska is typical? Can you think of any factors that might explain the data for Alaska in comparison with the other states?

$$Alaska \begin{bmatrix} Murder & Assault & UrbanPop & Rape \\ 10.0 & 263 & 48 & 44.45 \end{bmatrix} \tag{4}$$

   Alaska is in first quartile in UrbanPop but is in the third quartile in Murders, fourth quartile in Assault, and is second highest in rape. Definitely not typical, It would make sense that violent crime is committed at higher rates in more urban areas, yet the opposite is true for Alaska. It makes sense that the the urban population is so low, since Alaska even now, is somewhat underdeveloped and this data was taken back in 1973, only 14 years after it was made a state in 1959. It's possible that Alaska was seen as a sort of ""Wild West".

5. Make a frequency histogram for the numerical variable Rape. Let each bin be of size 5, and have the range on the x-axis be from 0 to 50. (This means your bins will be $[0, 5], [5, 10]$, etc.) How many states have between 15 and 20 rapes per 100,000 residents per year? Write the `R` command you used to plot this histogram in the space below.

   There are 12 states with between 15 and 20 rapes per 100,00 residents per year. The command used to build the histogram is,

$$hist(Rape, xlim = c(0, 50), breaks = 10); \tag{5}$$

6. Looking at the histogram for the variable Rape, can you tell if the mean is bigger than the median? Explain.

   Yes, because the histogram is skewed slightly to the right, which means that there are large data points that are driving the mean higher.

7. Now load the data table in the file `annual_income` into `R` and compute the mean and median, and plot a histogram. Choose settings for the histogram to display the data in a 'good' light, and include a graph of this histogram with your completed lab.

   Remove the largest test score and save it to a new variable called d. Type

   ```
   d = incomes[1:99]
   ```

   at the `R` prompt to do this. Explain the effect of removing the largest test score on the mean and the median.

   $$Incomes, \begin{bmatrix} Mean & Median \\ 59734 & 51826 \end{bmatrix} \tag{6}$$

   $$d, \begin{bmatrix} Mean & Median \\ 57913 & 51618 \end{bmatrix} \tag{7}$$

   When we remove the outlier the mean drops by 2000. However the median stays the same. This proves that skewed histograms will have a higher or lower mean depending on the direction of the skew.

R and RStudio are available over the internet. The URL for the Comprehensive R Archive Network is

http://www.r-project.org/

but a simple Google search will find you plenty of hits. You will need to choose a mirror and download a package appropriate for your operating system. If you are not familiar with downloading such a software package, see me during office hours for help.

All commands should be typed in R's console window.

Helpful commands from R:

```
> help(?????)            get help on command ?????

> data(USArrests)        loads data set USArrests
> USArrests              displays dataset
> attach(USArrests)      makes variables Murder, Assault,
                         UrbanPop, Rape available by name
> help(USArrests)        get info on dataset
> Murder                 display variable Murder
> summary(??)            display summary statistics for variable ??
> fivenum(??)            display Tukey's five number summary for variable ??
> head(USArrests)        displays the first part of the dataset USArrests
> boxplot(Murder,Rape)   creates boxplots for variables Murder, Rape
> boxplot(USArrests)     creates boxplots for all variables in dataset
> hist(Rape,seq(0,50,5)) creates a frequency histogram for Rape, with bins of
                         size 5 and a range from 0 to 50
```

```
> load("annual_income")          loads data from file
> ls()                           displays variable names
> incomes                        displays the variable ‘incomes’
> getwd()                        get the working directory
> setwd(’??’)                    set the working directory to ??
```