Lab 1: Descriptive Statistics

SOLUTIONS

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?

   People in the US in 1973.

   The idea is to get some idea about arrests rates for violent crimes by people in the US.

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.

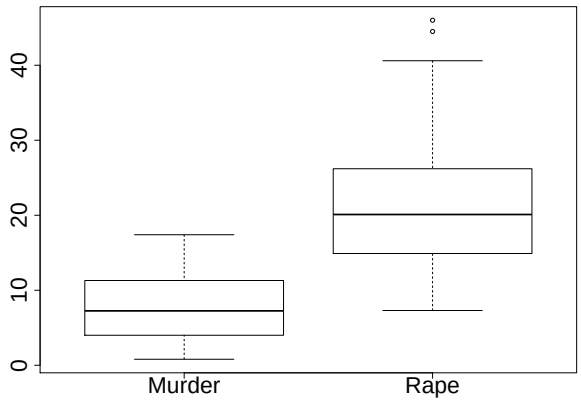
   ```
   > summary(Murder)
      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     0.800   4.075   7.250   7.788  11.250  17.400
   > summary(Rape)
      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      7.30   15.08   20.10   21.23   26.17   46.00
   ```

   I'd say the mean and median are reasonably close and looking at the boxplots in the next question, the data seem to be reasonably symmetric. On the other hand, R thinks that there are two outliers for state arrest rates as indicated by the circles. Alas, Alaska had the second highest arrest rate, and Nevada the highest.

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.

   (Boxplot on next page.)

   There are more arrests per 100,000 state residents for rapes than for murders. This is probably not too surprising since (one would at least hope) that fewer murders are committed per capita than rapes.



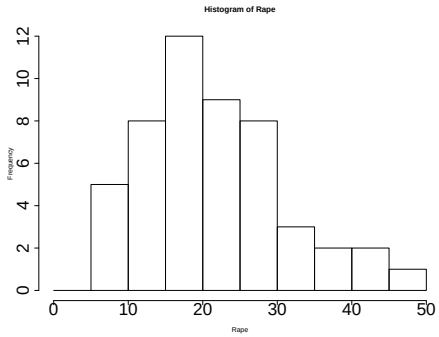
	Murder	Rape
Alaska	10.0	44.5

The per capita rape arrest rate is high for Alaska compared to other states. The per capita murder arrest rate is the 15th largest. That's pretty high in truth, though it's still below Q3.

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?

   This has been discussed in part above. Many answers are possible here, based on your personal viewpoint. My speculation would be that the long winters, high alcohol abuse rates, high gun ownership rates, urban and rural population demographics, etc. all might contribute to these high rates. Of course, more data would need to be collected to support any of these hypotheses.

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 rape arrests per 100,000 residents per year? 12. Write the `R` command you used to plot this histogram in the space below.



hist(Rape,seq(0,50,5),cex.axis=2.5)

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

   You can. There is a bit of a 'skew to the right,' making the mean larger than the median. See the answer to question 2 above.

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.

   The median will stay the same, but the mean will go down since a relatively large value was removed.

```
> summary(incomes)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  17790   39350   51830   59730   78300  240000
> summary(d)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  17790   39220   51620   57910   77530  115100
```