IST772 Problem Set 1

Abhijith Anil Vamadev

```
#The homework for week one is exercises 1, 3, and 4 on page 20.

#Attribution statement: `r (choose only one and complete as necessary)`
#1. I did this homework by myself, with help from the book and the professor.
```

Chapter 1, Exercise 1

Using the material from this chapter and possibly other information that you look up, write a brief definition of these terms in your own words: mean (aka average), median, mode, variance, standard deviation, histogram, normal distribution, and Poisson distribution. (1 point for each definition)

- Mean: Is the sum of all values divided by the count of the values. It is the 50th percentile, half the values are greater than mean and half the values are lesser than the mean.
- Median: Is the middle value of ordered values from smallest to the largest value.
- Mode: Is the most common occurring value from all the values in the data set. The value at the peak in a graph.
- Variance: Indicates how far a variable is from the mean.
- Standard deviation: Is the square root of the variance. Indicates the extent of variation within a group as a whole.
- Histogram: Counts the frequency of each value in the data set and indicates in the graph in the form of a bars.
- Normal distribution: Where the mean = median = mode.
- Poisson distribution: It is a discrete probability distribution that meansures the probability within a given time period.

Chapter 1, Exercise 3

Use the data() function to get a list of the data sets that are included with the basic installation of R: just type "data()" at the command line and press enter.

```
#data()
```

Choose a data set from the list that contains at least one numeric variable–for example, the Biochemical Oxygen Demand (BOD) data set. Use the summary() command to summarize the variables in the data set you selected–for example, summary(BOD). (1 pt) Write a brief description of the mean and median of each numeric variable in the data set. (1 pt for each value) Make sure you define what a "mean" and a *"median" are, that is, the technical definition and practical meaning of each of these quantities. (1 pt for each definition)*

```
data('women')
data <- women
summary(data)
```

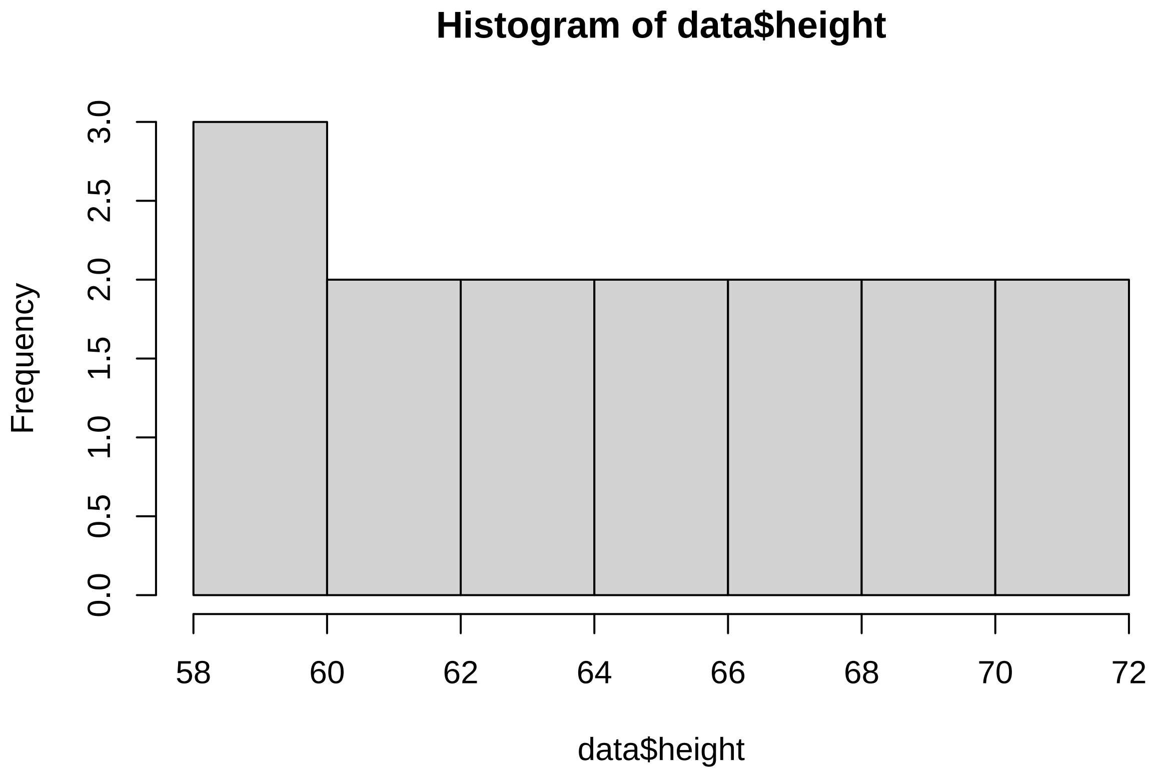
```
##      height         weight     
##  Min.   :58.0   Min.   :115.0  
##  1st Qu.:61.5   1st Qu.:124.5  
##  Median :65.0   Median :135.0  
##  Mean   :65.0   Mean   :136.7  
##  3rd Qu.:68.5   3rd Qu.:148.0  
##  Max.   :72.0   Max.   :164.0
```

#Mean - Is the sum of all the values in the column divided by the number of items in the column. On average the height of women from the data set is 65 and on average the weight of women from the data set is 136.7 #Median - The median would be ordering the values from the lowest to the largest and then taking the middle value, if total number of values in the data set are odd, then the middle number is taken. If the total number of values are even, then the middle pair is taken and their sum divided by 2. The median value for height is 65, the median value for weight is 135.0

Chapter 1, Exercise 4

As in the previous exercise, use the data() function to get a list of the data sets that are included with the basic installation of R. Choose a data set and pick out one variable, for example, the LakeHuron data set (levels of Lake Huron in the years 1875 through 1972). Use the hist() command to create a histogram of the variable–for example, hist(LakeHuron). (2 pts) Describe the shape of the histogram in words. (2 pts) Which of the distribution types do you think these data fit most closely (e.g., normal, Poisson). (2 pts) Speculate on why your selected data may fit that distribution. (2 pts)

```
hist(data$height)
```



#The histogram has a peak at 58 and for all the other values it is constant at 2. This distribution is close to uniform distribution as most of the values from 60 to 72 have a frequency of 2. I think the distribution is close ot uniform distribution as the frequency of most of the values are close to 2.