IST772 Problem Set 1

Shivani Sanjay Mahaddalkar

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

Attribution statement: 2. I did this homework with help from the book and the professor and these Internet sources: a. https://en.wikipedia.org/wiki/Poisson_distribution# (https://en.wikipedia.org/wiki/Poisson_distribution#):~:text=In%20probability%20theory%20and%20statistics,time%20or%20space%20if%20these

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: It is the sum of all values divided by the number of values
- Median: It is the center-most point of the data
- Mode: It is the most frequently occurring value
- Variance: It is the mean of sum of squared deviations from the mean
- Standard deviation: It is the square root of mean of sum of squared deviations from the mean
- Histogram: It is a plot that shows the frequency of values
- Normal distribution: Normal distribution is a bell shaped symmetric curve with a mean the same as the median and spans 99.7% of data in 3 standard deviations from the mean
- Poisson distribution: It is the distribution that expresses the probability of a given number of events occurring in a fixed interval. It is a discrete distribution.

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)

```
summary(Orange)
```

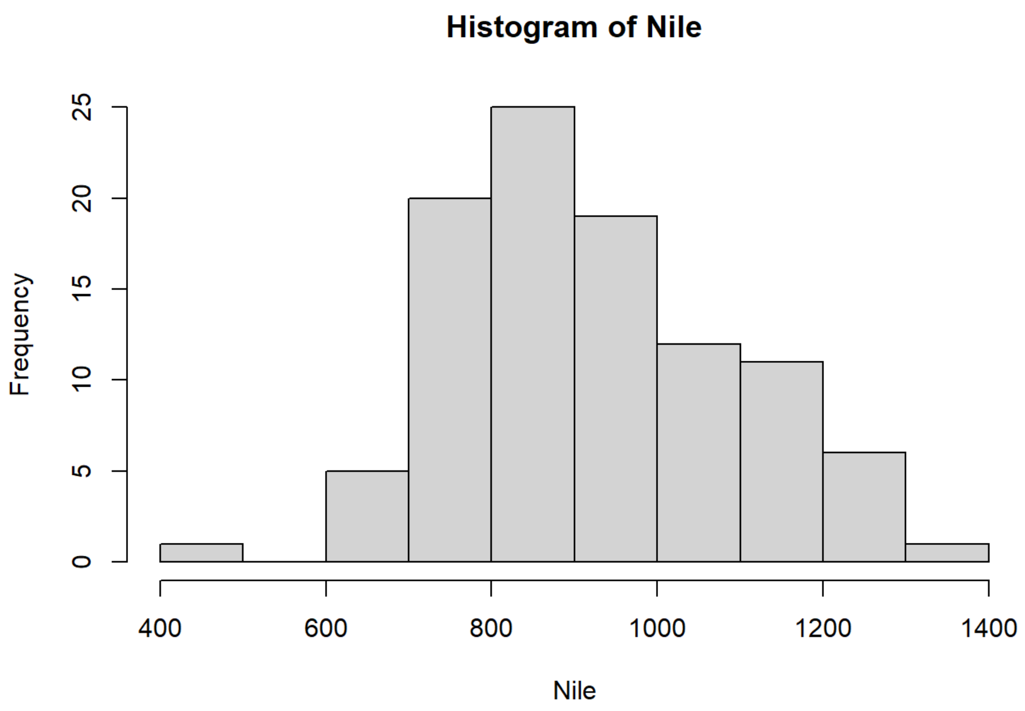
```
##  Tree       age         circumference  
##  3:7   Min.   : 118.0   Min.   : 30.0  
##  1:7   1st Qu.: 484.0   1st Qu.: 65.5  
##  5:7   Median :1004.0   Median :115.0  
##  2:7   Mean   : 922.1   Mean   :115.9  
##  4:7   3rd Qu.:1372.0   3rd Qu.:161.5  
##        Max.   :1582.0   Max.   :214.0
```

```
#Mean of age is 922.1 and circumference is 115.9. It is the sum of all values divided by the number of values.
#Median of age is 1004.0 and circumference is 114.0. It is the center-most point of the data.
#The mean age of the trees is 922 days and the the centre most point ie the median of age is 1004.0 days.
#The mean circumference of the trees is 115.9 mm and the median of the circumference is 114.0
```

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(Nile)
```



```
#The measurement of average flow of the Nile looks a lot like a bell shaped curve but skewed. It is close to a normal distribution but it is skewed. The data looks like it has the mean and the median around the same range ie between 800-900.
```