IST772– Problem Set 1

Sathish Kumar Rajendiran

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)

```
cat("\n ***************************************************************************")
```

```
## 
##  ***************************************************************************
```

- Mean: To find the mean or average of all the values in a data set, simply add up all the numbers and divide by total number of values from the data set.
- Median: The median or the halfway value of a data set is the middle data value when the data are put into ascending order. If the sample size n is odd, then the median is the middle value. If the sample size n is even, then the median is the mean of the middle data values.
- Mode: The mode or the modal value of the data is the data value that occurs the greatest frequency.
- Variance: variance is approximately the mean of the squared diviations or simply standard deviation squared.
- Standard deviation: Standard deviation can be interpreted as the typical differnce between a data value and the sample mean for a given data set or simply the square root of the variance.
- Histogram: Graphical representation of the numerical data distribution (spread) using bars with varying heights. Each bar groups number into ranges. Taller bars show that more data falls in that range.
- Normal distribution: Distribution of a continuous data with symmetric tails and gradual curves toward a peak in the middle is called normal distribution or simply "the bell curve".
- Poisson distribution: When the distribution of discrete data represented by different shapes at differnt levels of lamdba ("mean" value) is called Poisson distribution. In this all the observations are positive.

```
cat("\n ***************************************************************************")
```

```
## 
##  ***************************************************************************
```

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()
```

Data sets in package 'datasets':

AirPassengers Monthly Airline Passenger Numbers 1949-1960 BJsales Sales Data with Leading Indicator BJsales.lead (BJsales) Sales Data with Leading Indicator BOD Biochemical Oxygen Demand CO2 Carbon Dioxide Uptake in Grass Plants ChickWeight Weight versus age of chicks on different diets DNase Elisa assay of DNase EuStockMarkets Daily Closing Prices of Major European Stock Indices, 1991-1998 Formaldehyde Determination of Formaldehyde HairEyeColor Hair and Eye Color of Statistics Students Harman23.cor Harman Example 2.3 Harman74.cor Harman Example 7.4 Indometh Pharmacokinetics of Indomethacin InsectSprays Effectiveness of Insect Sprays JohnsonJohnson Quarterly Earnings per Johnson & Johnson Share LakeHuron Level of Lake Huron 1875-1972 LifeCycleSavings Intercountry Life-Cycle Savings Data Loblolly Growth of Loblolly pine trees Nile Flow of the River Nile Orange Growth of Orange Trees OrchardSprays Potency of Orchard Sprays PlantGrowth Results from an Experiment on Plant Growth Puromycin Reaction Velocity of an Enzymatic Reaction Seatbelts Road Casualties in Great Britain 1969-84 Theoph Pharmacokinetics of Theophylline Titanic Survival of passengers on the Titanic ToothGrowth The Effect of Vitamin C on Tooth Growth in Guinea Pigs UCBAdmissions Student Admissions at UC Berkeley UKDriverDeaths Road Casualties in Great Britain 1969-84 UKgas UK Quarterly Gas Consumption USAccDeaths Accidental Deaths in the US 1973-1978 USArrests Violent Crime Rates by US State USJudgeRatings Lawyers' Ratings of State Judges in the US Superior Court USPersonalExpenditure Personal Expenditure Data UScitiesD Distances Between European Cities and Between US Cities VADeaths Death Rates in Virginia (1940) WWWusage Internet Usage per Minute WorldPhones The World's Telephones ability.cov Ability and Intelligence Tests airmiles Passenger Miles on Commercial US Airlines, 1937-1960 airquality New York Air Quality Measurements anscombe Anscombe's Quartet of 'Identical' Simple Linear Regressions attenu The Joyner-Boore Attenuation Data attitude The Chatterjee-Price Attitude Data austres Quarterly Time Series of the Number of Australian Residents beaver1 (beavers) Body Temperature Series of Two Beavers beaver2 (beavers) Body Temperature Series of Two Beavers cars Speed and Stopping Distances of Cars chickwts Chicken Weights by Feed Type co2 Mauna Loa Atmospheric CO2 Concentration crimtab Student's 3000 Criminals Data discoveries Yearly Numbers of Important Discoveries esoph Smoking, Alcohol and (O)esophageal Cancer euro Conversion Rates of Euro Currencies euro.cross (euro) Conversion Rates of Euro Currencies eurodist Distances Between European Cities and Between US Cities faithful Old Faithful Geyser Data fdeaths (UKLungDeaths) Monthly Deaths from Lung Diseases in the UK freeny Freeny's Revenue Data freeny.x (freeny) Freeny's Revenue Data freeny.y (freeny) Freeny's Revenue Data infert Infertility after Spontaneous and Induced Abortion iris Edgar Anderson's Iris Data iris3 Edgar Anderson's Iris Data islands Areas of the World's Major Landmasses ldeaths (UKLungDeaths) Monthly Deaths from Lung Diseases in the UK lh Luteinizing Hormone in Blood Samples longley Longley's Economic Regression Data lynx Annual Canadian Lynx trappings 1821-1934 mdeaths (UKLungDeaths) Monthly Deaths from Lung Diseases in the UK morley Michelson Speed of Light Data mtcars Motor Trend Car Road Tests nhtemp Average Yearly Temperatures in New Haven nottem Average Monthly Temperatures at Nottingham, 1920-1939 npk Classical N, P, K Factorial Experiment occupationalStatus Occupational Status of Fathers and their Sons precip Annual Precipitation in US Cities presidents Quarterly Approval Ratings of US Presidents pressure Vapor Pressure of Mercury as a Function of Temperature quakes Locations of Earthquakes off Fiji randu Random Numbers from Congruential Generator RANDU rivers Lengths of Major North American Rivers rock Measurements on Petroleum Rock Samples sleep Student's Sleep Data stack.loss (stackloss) Brownlee's Stack Loss Plant Data stack.x (stackloss) Brownlee's Stack Loss Plant Data stackloss Brownlee's Stack Loss Plant Data state.abb (state) US State Facts and Figures state.area (state) US State Facts and Figures state.center (state) US State Facts and Figures state.division (state) US State Facts and Figures state.name (state) US State Facts and Figures state.region (state) US State Facts and Figures state.x77 (state) US State Facts and Figures sunspot.month Monthly Sunspot Data, from 1749 to "Present" sunspot.year Yearly

Sunspot Data, 1700-1988 sunspots Monthly Sunspot Numbers, 1749-1983 swiss Swiss Fertility and Socioeconomic Indicators (1888) Data treering Yearly Treering Data, -6000-1979 trees Diameter, Height and Volume for Black Cherry Trees uspop Populations Recorded by the US Census volcano Topographic Information on Auckland's Maunga Whau Volcano warpbreaks The Number of Breaks in Yarn during Weaving women Average Heights and Weights for American Women

Data sets in package 'ggplot2':

diamonds Prices of over 50,000 round cut diamonds economics US economic time series economics_long US economic time series faithfuld 2d density estimate of Old Faithful data luv_colours 'colors()' in Luv space midwest Midwest demographics mpg Fuel economy data from 1999 to 2008 for 38 popular models of cars msleep An updated and expanded version of the mammals sleep dataset presidential Terms of 11 presidents from Eisenhower to Obama seals Vector field of seal movements txhousing Housing sales in TX

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)

```r
cat("\n ****************************************************************************")
```

```
##
##  ****************************************************************************
```

```r
#Biochemical Oxygen Demand
# ?BOD   #view definition of the BOD dataset; This data set has two numeric variables Time and demand a
myBOD <- BOD

cat("\n Structure of the Biochemical Oxygen Demand data set is:\n")
```

```
##
##  Structure of the Biochemical Oxygen Demand data set is:
```

```r
str(myBOD) # analyze the structure of the data set.
```

```
## 'data.frame':    6 obs. of  2 variables:
##  $ Time  : num  1 2 3 4 5 7
##  $ demand: num  8.3 10.3 19 16 15.6 19.8
##  - attr(*, "reference")= chr "A1.4, p. 270"
```

```r
# View(myBOD)  # view the data set

cat("\n Summary of the Biochemical Oxygen Demand data set is:\n")
```

```
##
##  Summary of the Biochemical Oxygen Demand data set is:
```

```r
summary(myBOD)  # Summarize the variables from the dataset
```

```
##       Time          demand
##  Min.   :1.000   Min.   : 8.30
##  1st Qu.:2.250   1st Qu.:11.62
##  Median :3.500   Median :15.80
##  Mean   :3.667   Mean   :14.83
##  3rd Qu.:4.750   3rd Qu.:18.25
##  Max.   :7.000   Max.   :19.80
```

```r
  # Time             demand
  # Mean   :3.667   Mean   :14.83

#calculate Mean and Median of "Time" of the variable
mean_time <- mean(myBOD$Time)
median_time <- median(myBOD$Time)
cat("\n mean of the Time variable is:",mean_time)
```

```
##
##  mean of the Time variable is: 3.666667
```

```r
cat("\n median of the Time variable is:",median_time)
```

```
##
##  median of the Time variable is: 3.5
```

```r
#calculate Mean and Median of "demand" of the variable
mean_demand <- mean(myBOD$demand)
median_demand <- median(myBOD$demand)

cat("\n mean of the demand variable is:",mean_demand)
```

```
##
##  mean of the demand variable is: 14.83333
```

```r
cat("\n median of the demand variable is:",median_demand,"\n")
```

```
##
##  median of the demand variable is: 15.8
```

```r
cat("\n ****************************************************************************")
```

```
##
##  ****************************************************************************
```

```r
cat("\n Mean:\n")
```

```
##
##  Mean:
```

```
cat("    Mean is the average value from the observations.It is calculated by sum of all
    observation values and divide by the total number of observations.
 Practical Meaning:
    In this dataset there are two variables (Time and demand) having 6 observations of each.")
```

```
##     Mean is the average value from the observations.It is calculated by sum of all
##     observation values and divide by the total number of observations.
##  Practical Meaning:
##     In this dataset there are two variables (Time and demand) having 6 observations of each.
```

```
cat("\n Time:")
```

```
##
##  Time:
```

```
time <- myBOD$Time
S_Time <- sum(time)
n_Time <- length(time)
avg_time <- S_Time/n_Time

cat("\n    Sum of the Time variable is:",S_Time)
```

```
##
##     Sum of the Time variable is: 22
```

```
cat("\n    Count of the Time variable is:",n_Time)
```

```
##
##     Count of the Time variable is: 6
```

```
cat("\n    Mean of the Time variable is:",avg_time)
```

```
##
##     Mean of the Time variable is: 3.666667
```

```
cat("\n Demand:")
```

```
##
##  Demand:
```

```
demand <- myBOD$demand
S_demand <- sum(demand)
n_demand <- length(demand)
avg_demand <- S_demand/n_demand
cat("\n    Sum of the Demand variable is:",S_demand)
```

```
##
##     Sum of the Demand variable is: 89
```

```r
cat("\n    Count of the Demand variable is:",n_demand)
```

```
##
##     Count of the Demand variable is: 6
```

```r
cat("\n    Mean of the Demand variable is:",avg_demand)
```

```
##
##     Mean of the Demand variable is: 14.83333
```

```r
cat("\n ****************************************************************************")
```

```
##
##  ****************************************************************************
```

```r
cat("\n Median:\n")
```

```
##
##  Median:
```

```r
cat("    The median or the halfway value of a data set is the middle data value when the data are put i
    If the sample size n is odd, then the median is the middle value.
    If the sample size n is even, then the median is the mean of the middle data values.
 Practical Meaning:
    In this dataset there are two variables (Time and demand) having 6 (even number of sample size) obse
    taking the mean of middle data values. In this case 3 and 4 position values")
```

```
##     The median or the halfway value of a data set is the middle data value when the data are put int
##     If the sample size n is odd, then the median is the middle value.
##     If the sample size n is even, then the median is the mean of the middle data values.
##  Practical Meaning:
##     In this dataset there are two variables (Time and demand) having 6 (even number of sample size)
##     taking the mean of middle data values. In this case 3 and 4 position values
```

```r
cat("\n Time:")
```

```
##
##  Time:
```

```r
sorted_time <- sort(time,decreasing = FALSE)
sum_middle_values_time <- sum(sorted_time[3:4])
n1 <- length(sorted_time[3:4])
cat("\n    sorted observations of time variable values:",sorted_time)
```

```
##
##     sorted observations of time variable values: 1 2 3 4 5 7
```

```r
cat("\n    sum of middle 2 values of sorted time variable:",sum_middle_values_time)
```

```
##
##     sum of middle 2 values of sorted time variable: 7
```

```r
cat("\n    median of time variable:",sum_middle_values_time/n1)
```

```
##
##     median of time variable: 3.5
```

```r
cat("\n Demand:")
```

```
##
##  Demand:
```

```r
sorted_demand <- sort(demand,decreasing = FALSE)
sum_middle_values_demand <- sum(sorted_demand[3:4])
n2 <- length(sorted_demand[3:4])
cat("\n    sorted observations of demand variable values:",sorted_demand)
```

```
##
##     sorted observations of demand variable values: 8.3 10.3 15.6 16 19 19.8
```

```r
cat("\n    sum of middle 2 values of sorted demand variable:",sum_middle_values_demand)
```

```
##
##     sum of middle 2 values of sorted demand variable: 31.6
```

```r
cat("\n    median of demand variable:",sum_middle_values_demand/n2)
```

```
##
##     median of demand variable: 15.8
```

```r
cat("\n ****************************************************************************")
```

```
##
##  ****************************************************************************
```

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)

```r
library("ggplot2")
library("hrbrthemes")
```

```
## NOTE: Either Arial Narrow or Roboto Condensed fonts are required to use these themes.

##       Please use hrbrthemes::import_roboto_condensed() to install Roboto Condensed and

##       if Arial Narrow is not on your system, please see https://bit.ly/arialnarrow
```

```r
cat("\n ****************************************************************************\n")
```

```
##
##  ****************************************************************************
```

```r
# Level of Lake Huron 1875-1972
# ?LakeHuron # Annual measurements of the level, in feet, of Lake Huron 1875-1972. A time series of len
myLakeHuron <- data.frame(LakeHuron)
str(myLakeHuron)
```

```
## 'data.frame':    98 obs. of  1 variable:
##  $ LakeHuron: Time-Series  from 1875 to 1972: 580 582 581 581 580 ...
```

```r
summary(myLakeHuron)
```

```
##    LakeHuron
##  Min.   :576.0
##  1st Qu.:578.1
##  Median :579.1
##  Mean   :579.0
##  3rd Qu.:579.9
##  Max.   :581.9
```

```r
# View(myLakeHuron)
# sd(myLakeHuron$LakeHuron)
# mean(myLakeHuron$LakeHuron)

#calculate Mean and Median of "Time" of the variable
mean_Huron <- mean(myLakeHuron$LakeHuron)
median_Huron <- median(myLakeHuron$LakeHuron)
sd_Huron <- sd(myLakeHuron$LakeHuron)
v_Huron <- var(myLakeHuron$LakeHuron)
cat("\n mean of the LakeHuron variable is:",mean_Huron)
```

```
##
##  mean of the LakeHuron variable is: 579.0041
```

```r
cat("\n median of the LakeHuron variable is:",median_Huron)
```

```
##
##  median of the LakeHuron variable is: 579.12
```

```
cat("\n Standard Deviation of the LakeHuron variable is:",sd_Huron)
```

```
##
##  Standard Deviation of the LakeHuron variable is: 1.318299
```

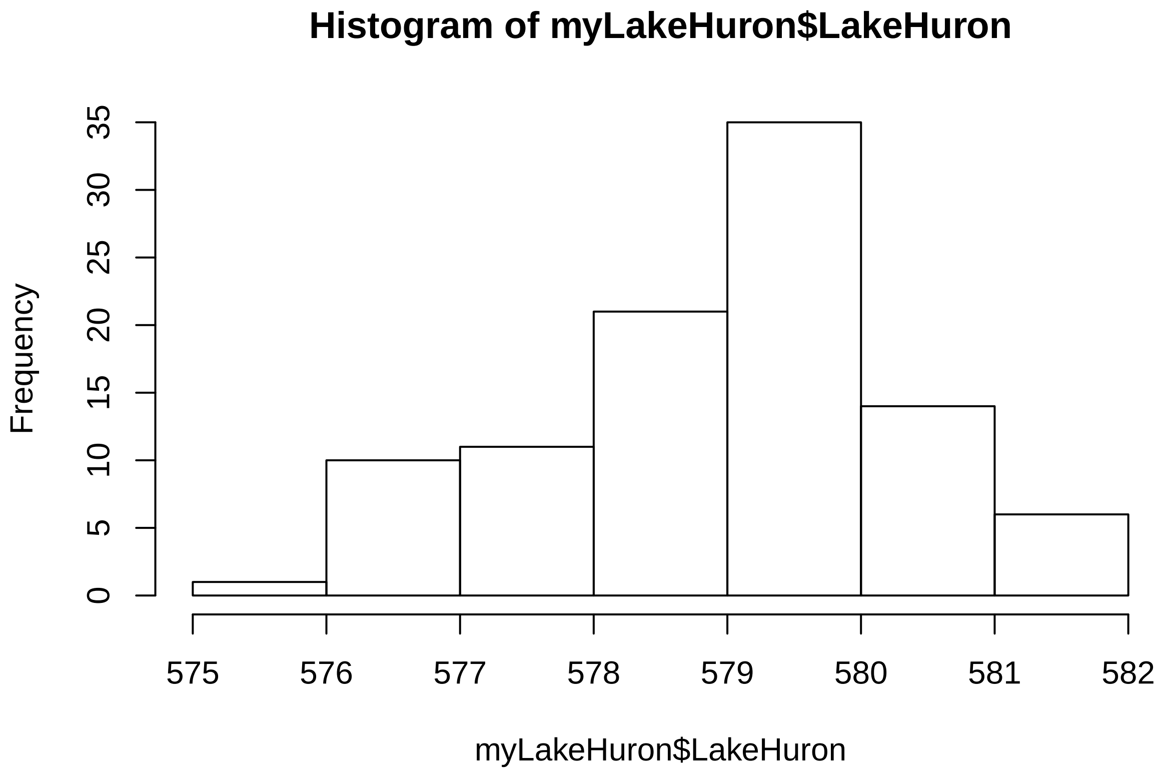
```
cat("\n Variance of the LakeHuron variable is:",v_Huron)
```

```
##
##  Variance of the LakeHuron variable is: 1.737911
```

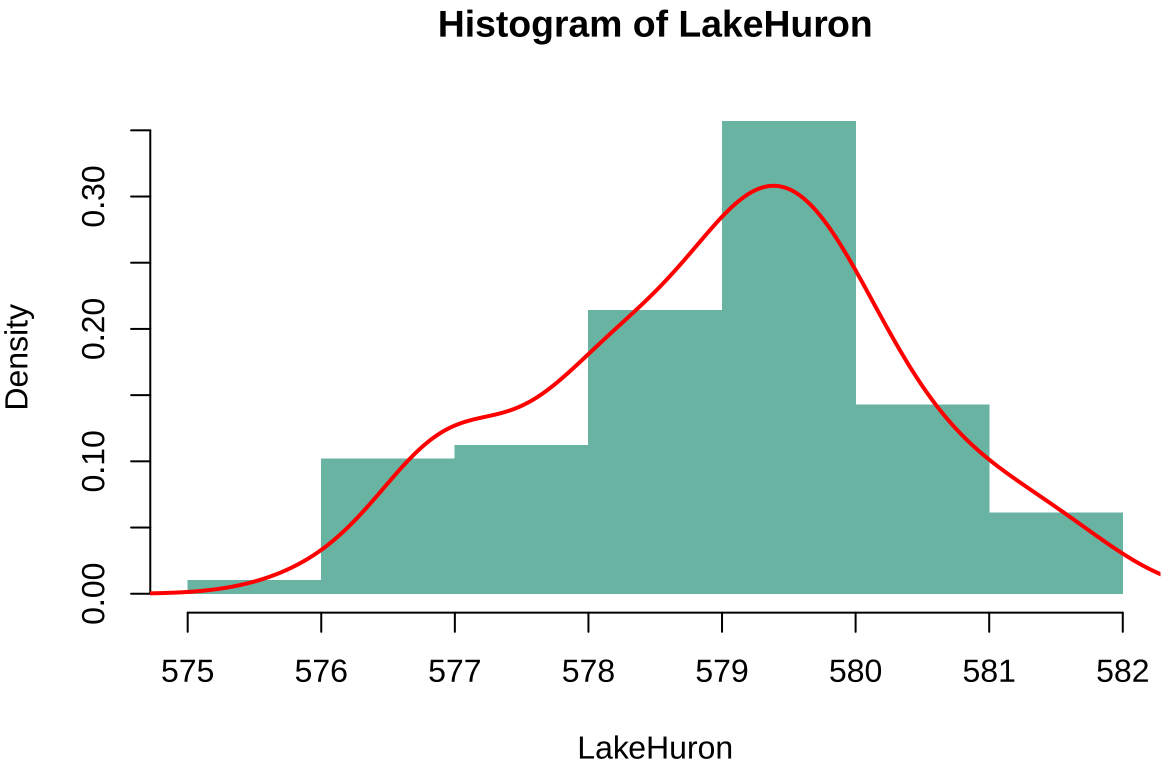
```
cat("\n ****************************************************************************\n")
```

```
##
##  ****************************************************************************
```

```
#createa a histogram
hist(myLakeHuron$LakeHuron)
```



```
# create a histogram
# hist(myLakeHuron$LakeHuron)
hist(myLakeHuron$LakeHuron,freq=FALSE,col="#69b3a2",lty="blank",include.lowest = TRUE,main = paste("His
lines(density(myLakeHuron$LakeHuron),col="red", lwd=2)
```



- Shape of the Histogram: This histogram represents a bell shaped curve with one peak value and two low tails on each side.
- Ditribution Type: This dataset has continuous data with decimal values; having almost 98 observations, the values are ditributed normally across with symmetric tails and gradual curve towards the peak.Hence, it fits normal distribution pattern closely.
- Conclusion: This histogram fits normal distribution closely;becuase, with 98 observations (continuous data), mean and median are almost same 579.0 and 579.1 with standard deviation as 1.and variance 1.7. In addition, the density curve - represents a bell shaped curve almost fitting the values under the curve.So, both the measures of central tendency and dispersion suggests a normal ditribution.

```
cat("\n ****************************************************************************")
```

```
##
##  ****************************************************************************
```