1. (10points) A questionnaire was sent to 34,000 randomly chosen individuals living in USA asking about their current and past smoking habits. For those who answered the questionnaire, they are tracked in the next 10 year for the death rate. Their data was pulled from the health record database so that if they died, the death is recorded with the diagnosis of the cause of the death. The death rates due to heart disease among the smokers and nonsmokers are then compared to see if smoking causes heart disease.

What is the targeted population? What is the sampling population? What is the sample? What are the parameters we are interested in? What are the statistics? Is this an observational study or controlled experiment? Is it possible to draw conclusion from this study that smoking causes heart disease? Why and why not?

Answer: Targeted Population: Individuals living in USA

Sampling Population: 34000 randomly chosen individuals.

Sample: Individuals that responded to the questionnaire

Parameters: Proportion of all smokers who died due to heart disease

Statistics: Proportion of all smokers who died due to heart disease of the individuals who responded to the questionnaire.

Observational or Control: This is an observational study because we are assigning people to groups.

Conclusion: We cannot draw a conclusion that smoking causes heart disease because this is not a controlled experiment.

2. (10pt) Each month, University of Michigan conducts phone interviews (starting 2014, the interviews are for cell phones only) for at least 500 households in continental United States (i.e., excluding Alaska and Hawaii). Based on the answers to a 50 questions questionnaire, a consumer sentiment score is calculated for each household. The mean score is published in the index of consumer sentiment by University of Michigan.

What is the targeted population? What is the sampling population? What is the sample? What is parameter we are interested in? What is the statistics?

Answer: Target Population: Households in continental USA

Sample Population: People with cell phones in continental USA

Sample: The (minimum) five hundred households that are interviewed each month

Parameters: Consumer sentiment score from each interviewed household

Statistics: Mean of the Consumer sentiment score

3. (10pt) Fancy College Mathematics Department Chair saw the starting salary data for their PhD for the last two years, and announced with a lot fanfare that average starting salary has increased 50% last year for their PhD graduates. Linda is suspicious of this claim, and surveyed her classmates. She said that the starting salary for the average Math PhD in the program are essentially the same for the last two years.

Which of these two is using the mean and which is using the median? Can both statements be true? Does the data indicates that the career prospect of Math PhD in this program has brighten recently?

Answer: In first case, when department chair saw increase in 50% salary for their all PhD graduates, it would be "Mean" definitely because median is the single person's salary, not all PhD graduates.

In second case, when Linda said that starting salary for average Math PhD is same, she is talking about "Median" because it is about single person.

Both statements can be true simultaneously.

From the data, it can be concluded that career prospect of Math PhD has brightened because most of the graduates are getting increased salary. It cannot be decided based on middle person's salary. It'll be dependent on a group's performance.

4. (5pt) A researcher believes that a college education is a waste of time. To prove this, she collected data on the cumulated wealth of every member of the Harvard Class of 1977 (and those dropouts supposed to graduate in 1977) in 25 years after graduation (wealth gained in the period of year 1977-2002). The average cumulated wealth of the dropouts is much higher than that of the graduates. She concludes from this data that finishing the college education is not as important as being admitted to the college. What do you think? (Bonus point: can you guess why the average wealth of the dropouts is higher than the graduates?)

Answer: I think that the conclusion is incorrect because

1. Harvard Class of 1977 is not a good representative of all college students.

2. data is highly skewed because number of dropout students is a lot less than number of graduates. To get a better result, we must normalize the data first.

The average wealth of the dropouts is higher than the graduates because dropouts from prestigious colleges such as Harvard are mostly entrepreneurial and have great vision, which makes them the target of investment and allows them to build successful businesses and are often offered opportunities to make more money.

====== Following are problems to do in R ==========

5. (10pt) Collect your response time data for the mini project. (See lab1 instruction). Then produce the following:

a) A scatterplot of your response times versus the observation order;

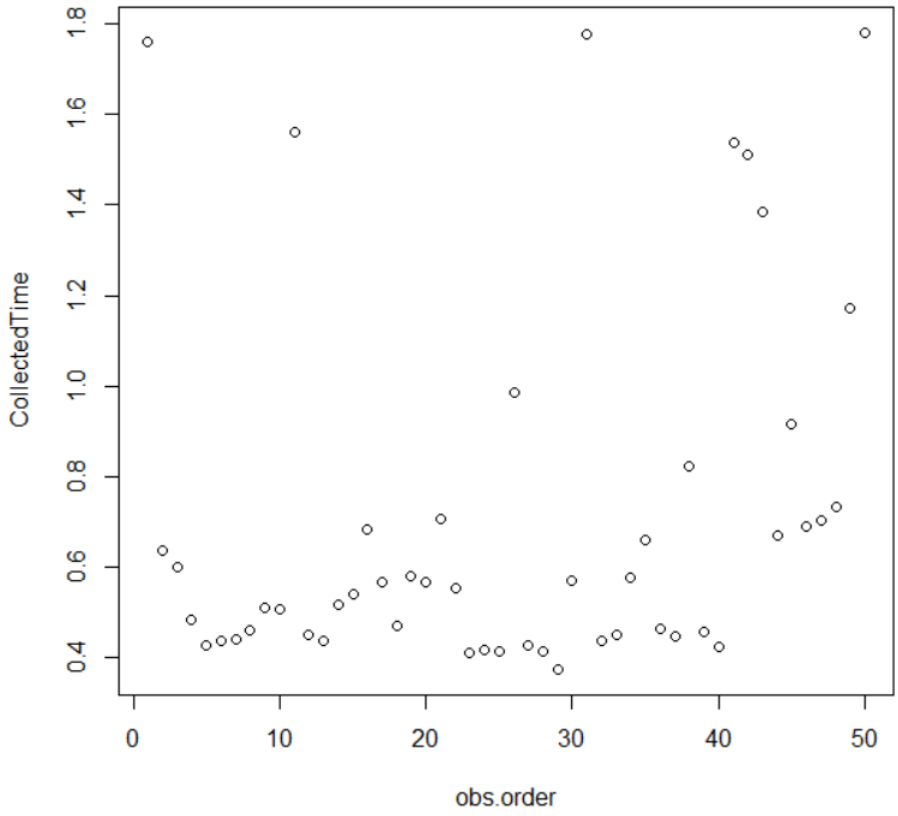
b) A stratified scatterplot of your response times in three phrases each of length 10;

c) A printout of the program you used in producing the above two plots.

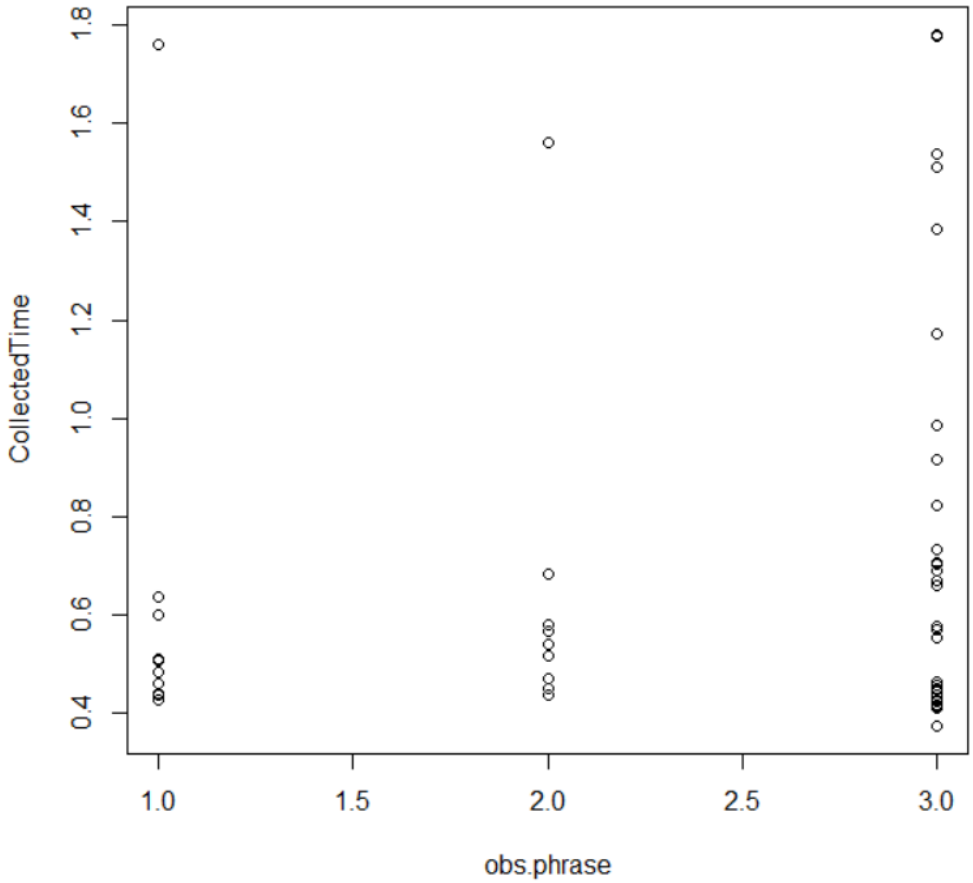
d) From your plots, do the response times seem to come from a stationary process?

Solution:

a) Response Time vs Observation Order



b) Stratified Scatterplot in 3 phrases.



c) Code Snippet used to create plots.

```
CollectedTime <- scan(file = "C:/Users/abhil/OneDrive/Desktop
                      /MSAM-Northeastern/MATH7343/Homeworks/1/CollectedResponseTimes.txt")
time.data <- data.frame(CollectedTime)
time.data$obs.order <- seq(length(time.data$CollectedTime))
time.data$obs.phrase <- ifelse(time.data$obs.order <= 10, 1, ifelse(time.data$obs.order <= 20, 2, 3))

#scatter plot of ResponseTime vs Observation Order

plot(CollectedTime ~ obs.order, data = time.data)

#Stratified scatter plot of 3 stages

plot(CollectedTime ~ obs.phrase, data = time.data)

#Summary

summary(CollectedTime)
summary(CollectedTime[1:10])
summary(CollectedTime[11:20])
summary(CollectedTime[21:30])
```

d) Summary

```
> summary(CollectedTime)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3750  0.4485  0.5605  0.7104  0.7047  1.7800
> summary(CollectedTime[1:10])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.4280  0.4445  0.4950  0.6261  0.5767  1.7600
> summary(CollectedTime[11:20])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.4360  0.4835  0.5540  0.6376  0.5763  1.5610
> summary(CollectedTime[21:30])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.375   0.414   0.422   0.527   0.565   0.987
```

The difference in descriptive statistics for all three phrases indicates that the response times are not coming from a stationary process.

6. (15pt) Analyze the salary data set, contained in the 'salary.txt' file which give the weekly salaries and gender for production workers in Anaheim, California. (This is not the 'fsalary.txt' file used in the example.)

(a) Produce a histogram of the salaries using R default setting.

(b) Produce a histogram use your own break points (at least 15 intervals).

(c) Produce a boxplot of the salaries.

(d) Produce side by side boxplots of the salaries in two gender groups.

(e) Produce summary statistics (as those in 'psych' package) of the salaries as one group, and also summary statistics within each gender.

(f) Submit the two histograms, two boxplots, and the summary statistics, as answer for each part (a-e). From the above plots, which histogram

(a) or (b) shows the feature in the data set better? Do the salaries for men and women seem to be the same?

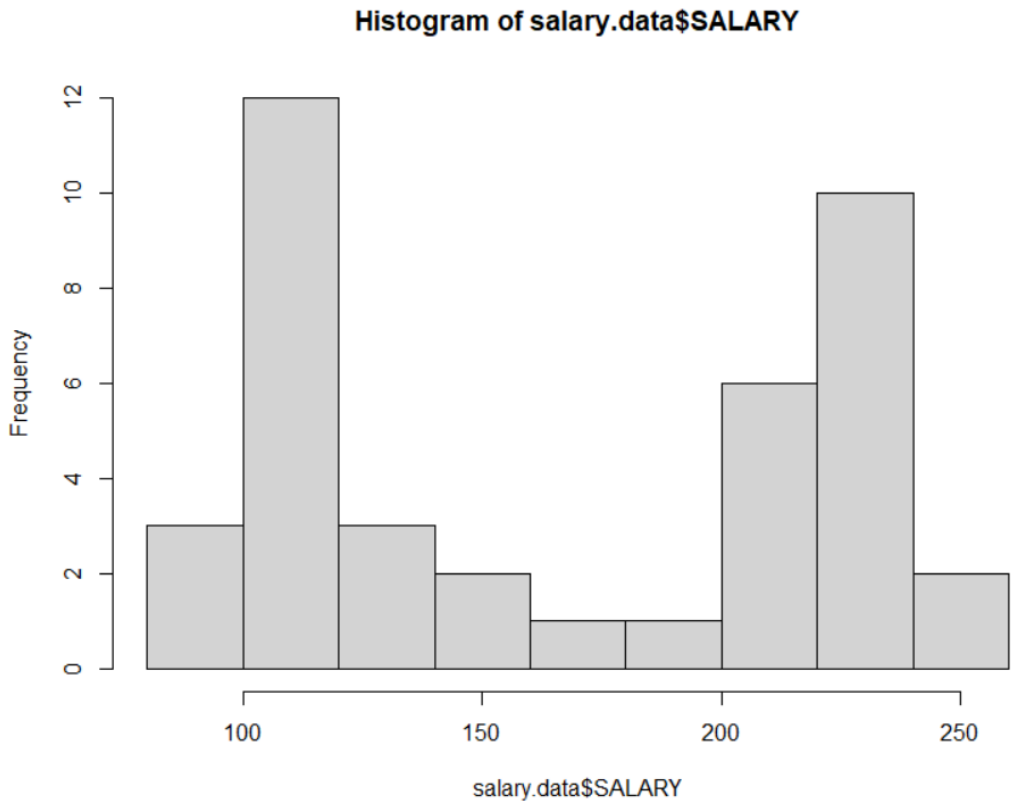
(g) Can you use one number to summarize the center of the distribution of the salaries? If yes, what is the number? If not, why?

(h) Are there any outliers in the data? If there are, can you identify them?

Solution:

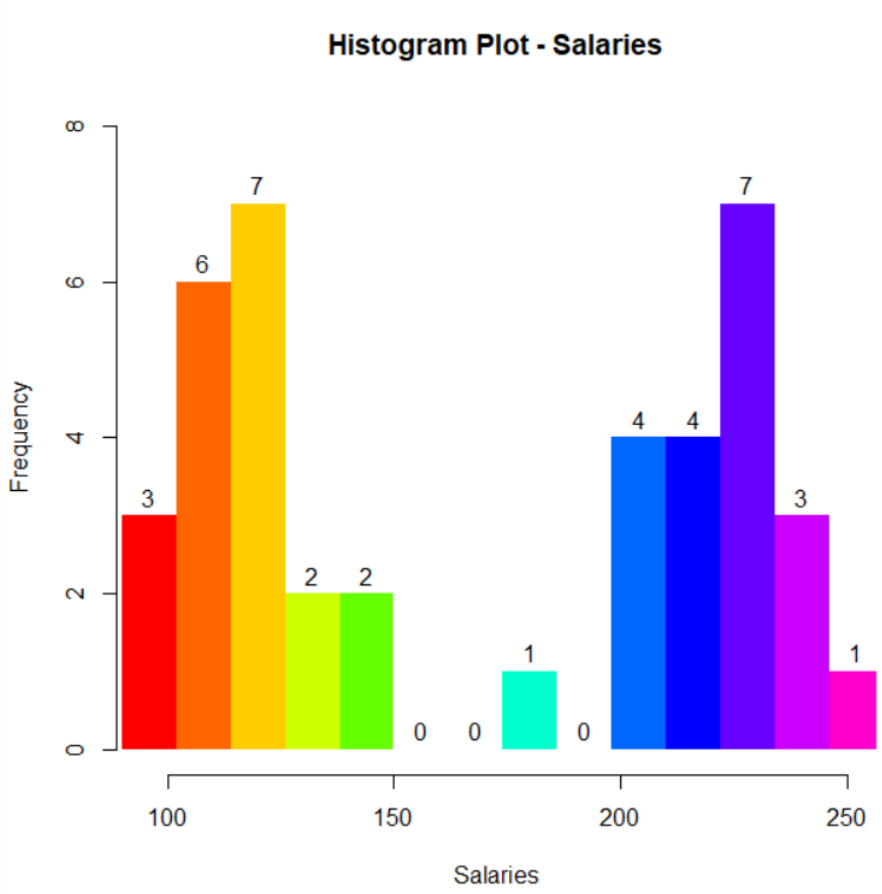
a) Histogram of salaries using default R settings

```
salary.data <- read.table(file = "C:/Users/abhil/OneDrive/Desktop
            /MSAM-Northeastern/MATH7343/Homeworks/1/salary.txt", header
      = TRUE)
hist(salary.data$SALARY)
```



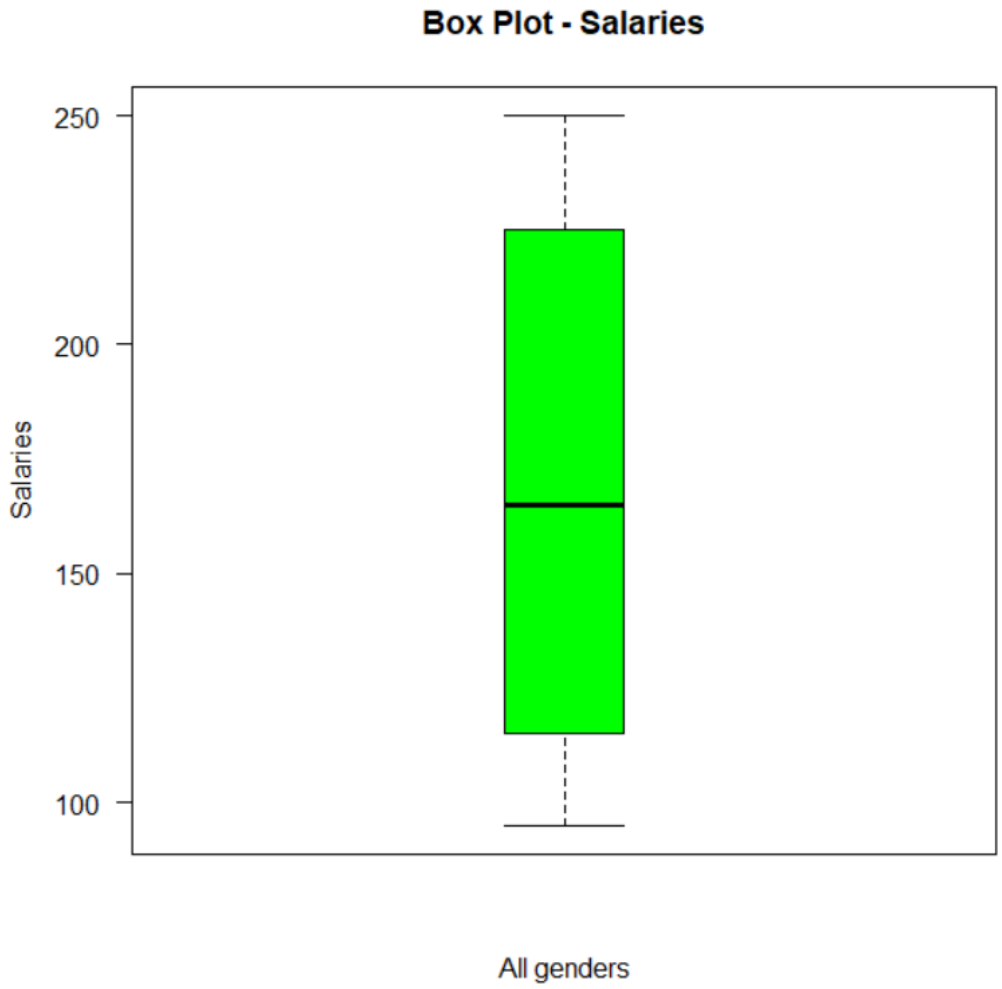
b) Histogram using my own break points

```
hist(salary.data$SALARY,
     breaks = 90+(0:14)*12,
     main = "Histogram Plot - Salaries",
     xlab = "Salaries",
     ylab = "Frequency",
     border = FALSE,
     labels = TRUE,
     xlim = c(min(salary.data$SALARY), max(salary.data$SALARY)),
     ylim = c(0, 8),
     col = rainbow(15))
```



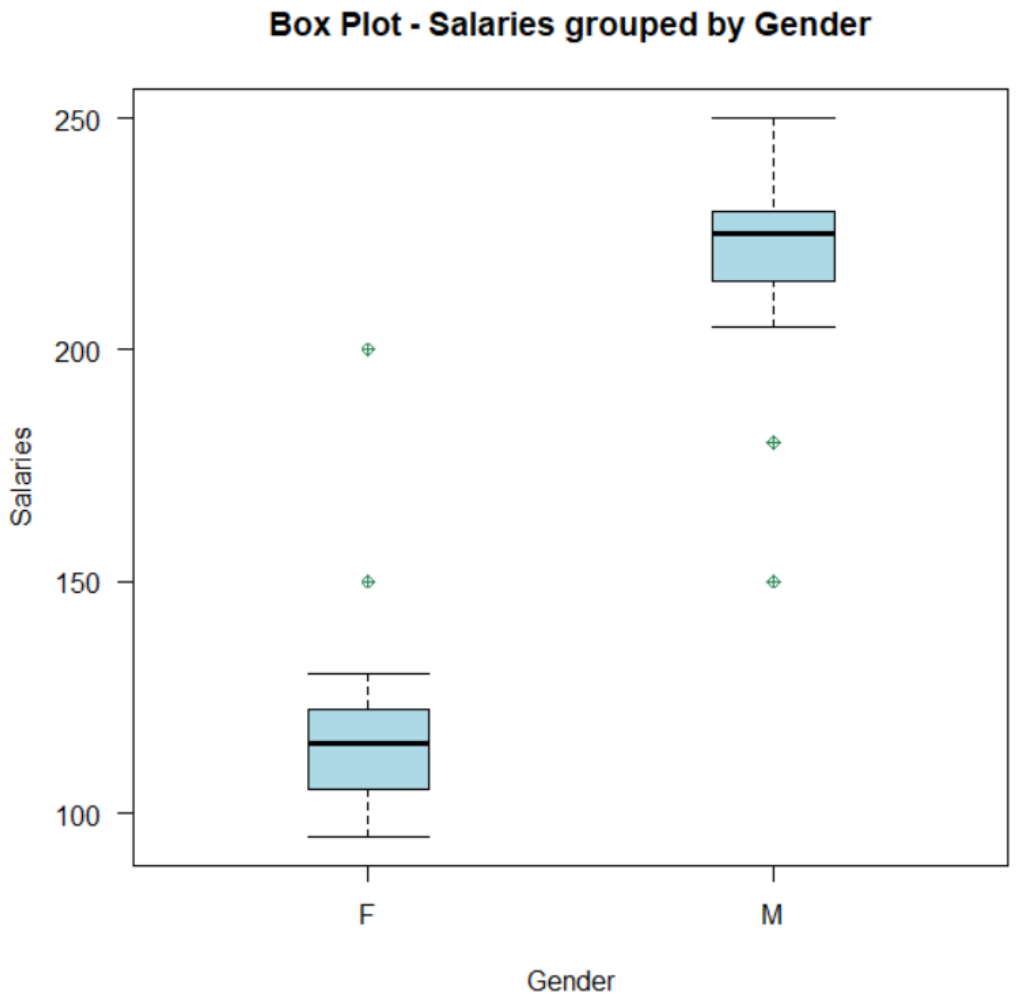
c) Boxplot of salaries

```
boxplot(salary.data$SALARY,
        main = "Box Plot - Salaries",
        xlab = "All genders",
        ylab = "Salaries",
        labels = TRUE,
        boxwex = 0.3,
        outline = TRUE,
        las = 1,
        notch = FALSE,
        staplewex = 1,
        col = "green")
```



d) Boxplot of salaries in two gender groups

```
boxplot(SALARY~GENDER, data=salary.data,
        main = "Box Plot - Salaries grouped by Gender",
        xlab = "Gender",
        ylab = "Salaries",
        labels = TRUE,
        boxwex = 0.3,
        outline = TRUE,
        outpch = 10,
        outcol = "seagreen",
        las = 1,
        notch = FALSE,
        staplewex = 1,
        col = "Light Blue")
```



e) Summary statistics

```
library(psych)
> describe(salary.data$SALARY)
   vars  n   mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 40 169.35 55.83    165  168.94 81.54  95 250   155 0.02    -1.83 8.83
```

Summary statistics by Gender

```
> describeBy(salary.data$SALARY, salary.data$GENDER)

 Descriptive statistics by group
group: F
   vars  n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 20 118.9 23.01    115  114.69 14.83  95 200   105 2.16     5.03 5.15
------------------------------------------------------------------------
group: M
   vars  n  mean    sd median trimmed  mad min max range  skew kurtosis   se
X1    1 20 219.8 22.57    225  223.19 7.41 150 250   100 -1.53     2.37 5.05
```

f) By comparing a) and b), it is safe to say that b) represents the data better in a histogram plot, since it gives the better spread of the data when intervals are close to each other. Higher number of intervals gives better distribution of the data but with greater complexity. Hence the number of intervals needs to be chosen wisely. In this case, 15 intervals give us better representation than the default 10 intervals.

From the box plot grouped by gender weekly salaries for the women is much lesser as compared to that of men.

g) The central tendencies for example mean, median produced by describe and describe by groups appear to be far apart from each other. This clearly implies that we cannot summarize the center of the distribution of the salaries.

h) From the above box plots grouped by gender, we could see that women have two outliers located above the maximum whisker which is Q3 + 1.5 times the inter quartile range. Their values are 150, 200. Similarly, men have two outliers located below the minimum whisker which is Q1 - 1.5 times the inter quartile range. Their values are 180, 150.

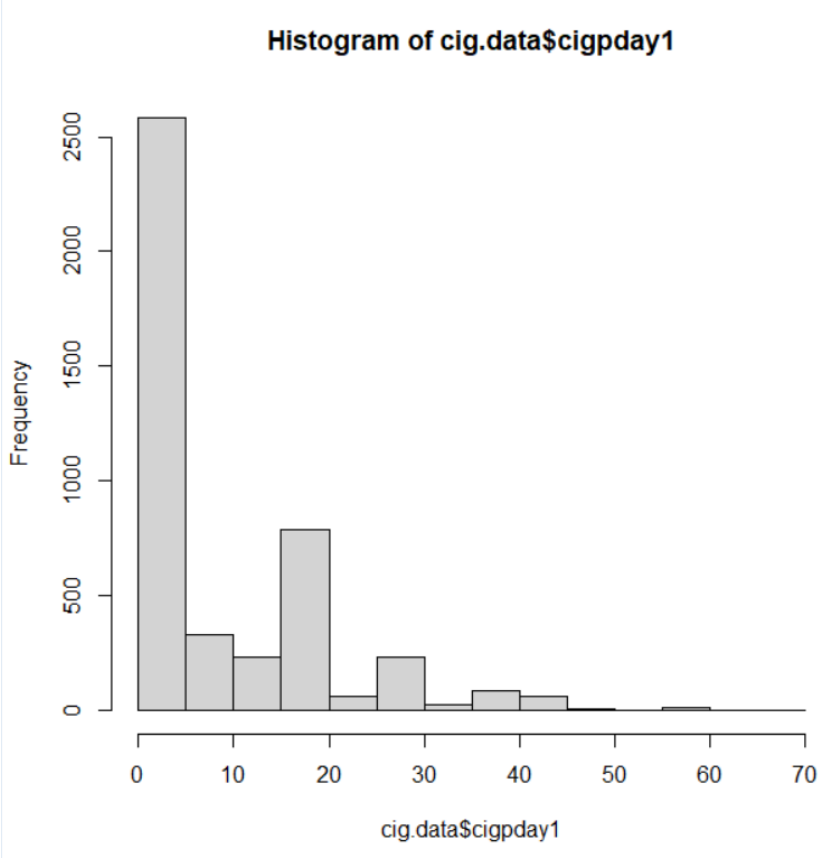
7. (5pt) Answer the following questions for the average number of cigarettes smoked per day for individuals enrolled in the Framingham Heart Study -- contained in file 'number_cigs.csv' under the variable name cigpday1.

(a) Produce a histogram and a boxplot.

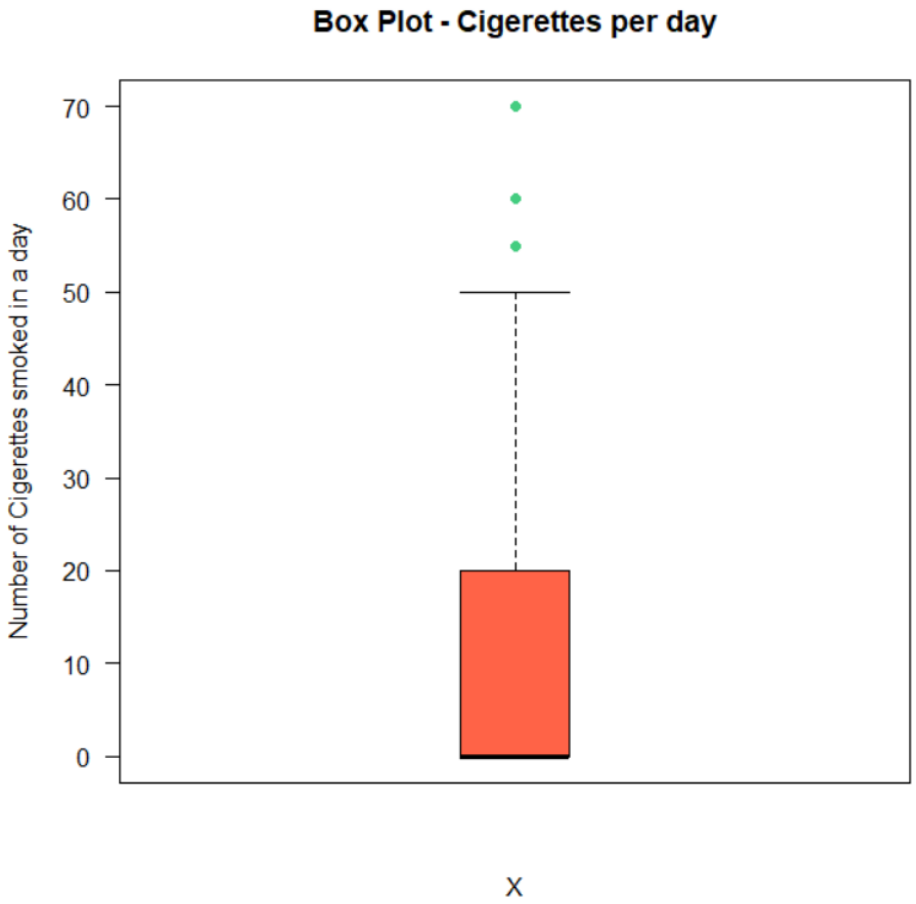
(b) Does the data appear to be skewed? If so, skewed to right or to the left?

Solution:

a) Histogram


Histogram of cig.data$cigpday1
Frequency
0
500
1000
1500
2000
2500
0
10
20
30
40
50
60
70
cig.data$cigpday1

Box Plot



b) The histogram plot clearly shows that the data is right skewed as the tail is elongated toward the right.