Assignment 1

by

Ronnie Nhalevilo (2730319), Yonatan Getachew (2660122) and Dimitar Bachvarov (2728704)
- group 87

1.1)

A) simple random sampling, because for every within the sample size (2052), every teen was randomly mailed by the Dutch Bureau of statistics, therefore every student had an equal opportunity of being mailed .
It is sound, as by randomly selecting or mailing the teens it lessens the chances of favoring any of the teens chances of being selected (mailed).

B) Cluster sampling _ as the teens are split into groups or clusters (the schools), and then within those 20 clusters which were randomly selected , all the teens were asked about their vaccination status.
It is sound as a large enough Sample size(20 schools) was used therefore making the risk of only receiving data from one type of student extremely small.

C) It is flawed because it is basically a voluntary sampling, as the data collected is only from the Internet users that chose to respond, thus increasing the risk of it being biased

1.2)

A) ordinal measurement
With ordinal measurement level a mean cannot be calculated, only the mode.

B) Ratio as there is a natural zero for the balance on the bank account, signifying bankruptcy, and a negative value would signify the debt of

the bank account. There seems to be no issue with the statistical summary, as a mean and standard deviation are suitable methods of finding the average bank balance and the variance of the balances from the mean.

1.3)

A)The following study is experimental, as the study subjects were induced with a cholesterol drug therefore some type of treatment was applied to the subjects thus they were modified.

B) It is systematic_ because a starting location is chosen (a lake), where the researcher uses a systemic method (line transect method) to collect random samples of data every 10 meters. Through this systemic method other researchers will be able to collect similar samples of data that the same lake.

C) Cluster sampling , because the television channel organized a poll where the voters are divided into clusters (the specific polling stations) and these stations are randomly selected , and all of the voters within those polling stations are selected and surveyed.

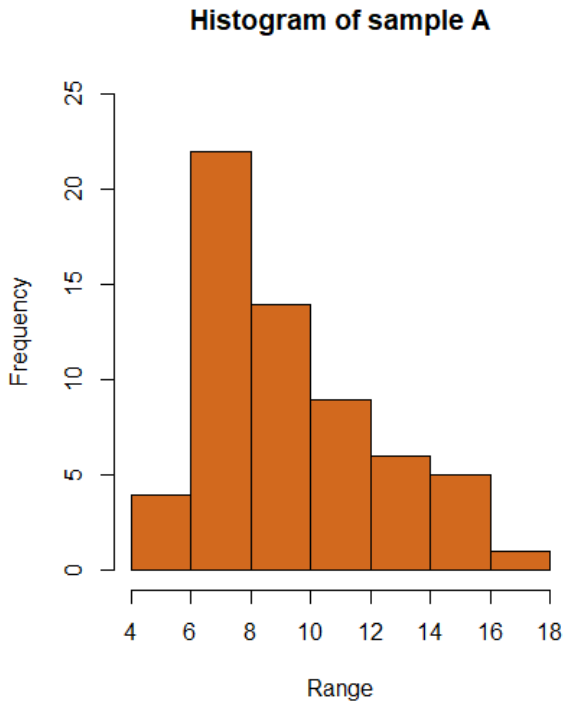
1.4)

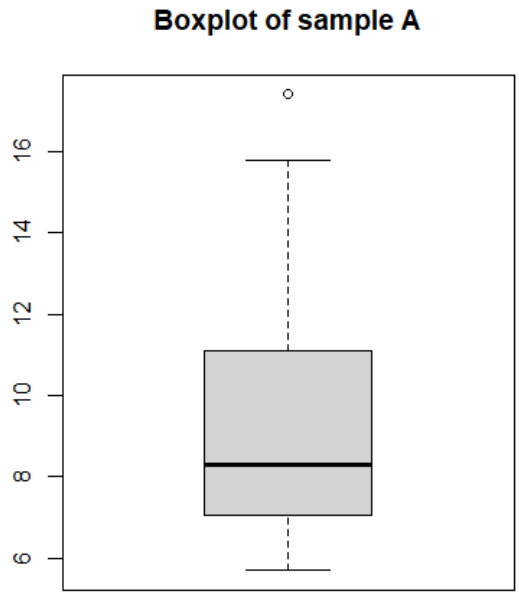
A) The following graph presentation is missing the SPD label on the x-axis, also the FDP label, the the scale on the y-axis should be extended as the maximum value for the SDP (206) exceeds the scale given, also the width for the bar of each value isn't uniformal .

B) Pareto chart would be the best way to represent the mistakes as it a

1.5)

A)

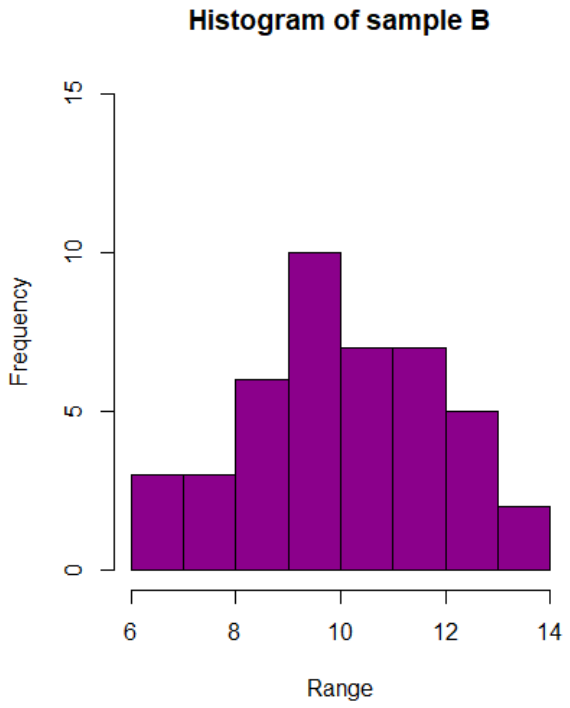


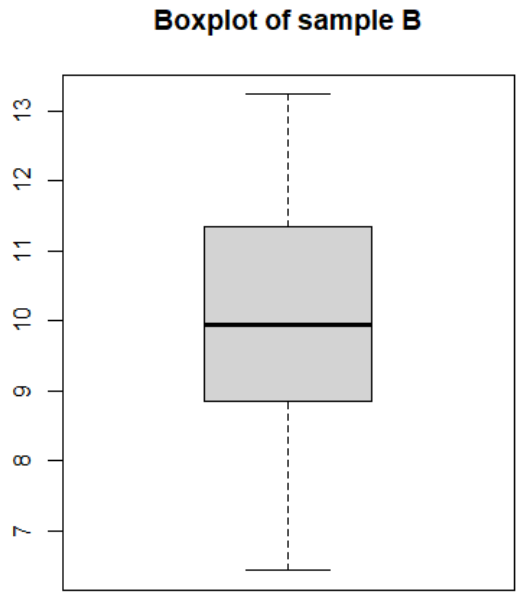


B) meanA = 9.29
varA = 8.19
sdA = 2.86

C) As shown in the histogram, sample A is asymmetric as it is skewed to the right, where the highest frequency of values range from 6-8. According to the boxplot, the dataset ranges from around 6 to around 16, with a median value of 8. This also shows that the data is skewed right, because if only the maximum and minimum values were used to calculate the median, the value returned would be 11. This is greater than the actual median of 8, which means that the data is positively skewed.
The mean of the sample A is 9.29 ± 2.86. However as seen in the boxplot, this might have been affected by the outlier as this mean is higher than the median average.

D)





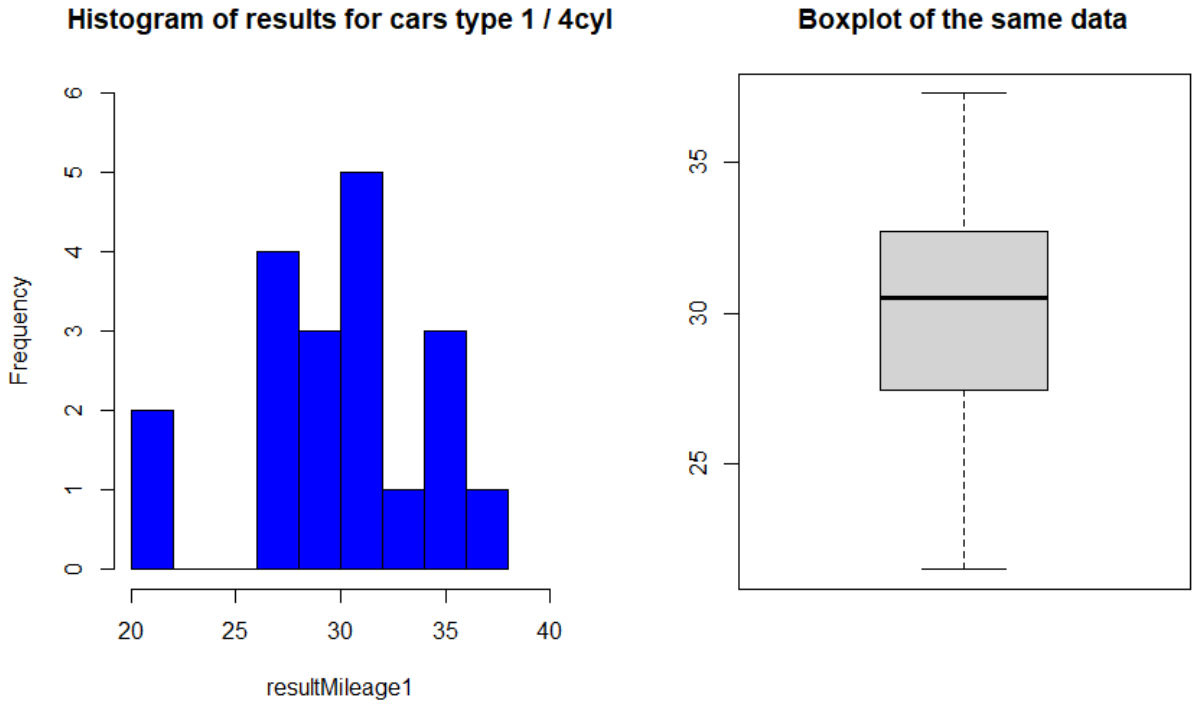
meanB = 10.08
varB = 3.18
stdB = 1.78

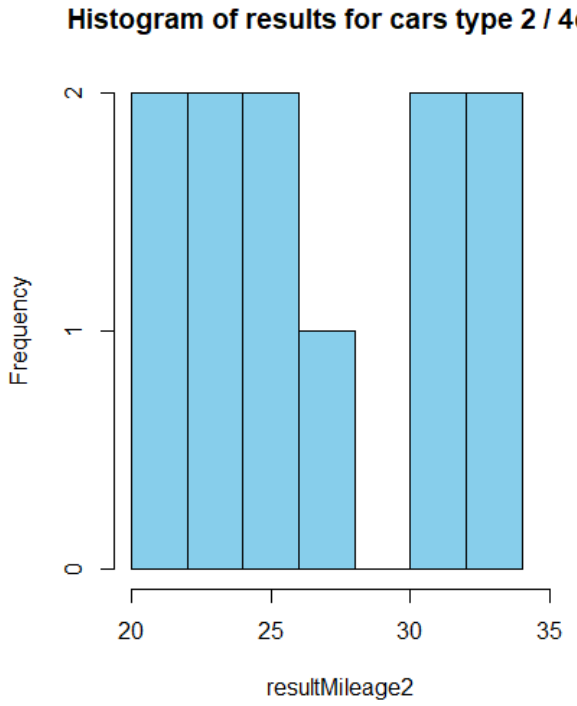
As shown in the histogram, sample B is shown to be symmetrical with no skew. The highest frequency of values range from 9-10.
According to the boxplot the data ranges approximately from 6 to 13, with a median value of 10. as when the median is calculated using just the max and min value, then the outcome would be 9.5. Though the expected median is less than the actual median, the significance of this is unknown, thus we can assume the data in sample B has no skew.
The mean of the sample B is 10.08 ± 1.78, which matches the median of the data, further proving centrality of the data.

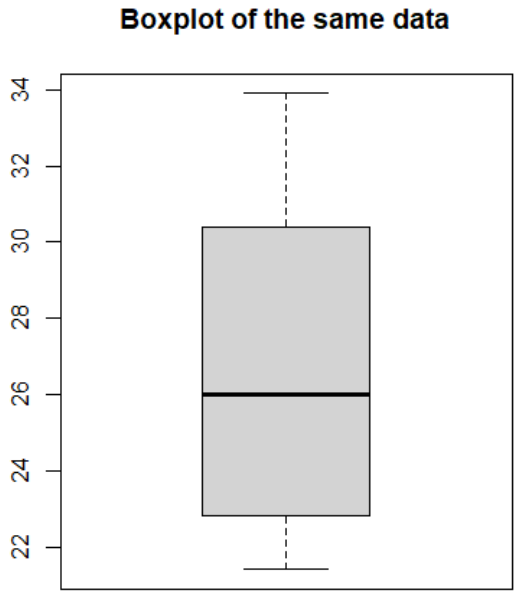
E) According to the results from the numeric summaries of samples A and B, it can be concluded that the two samples aren't derived from the same population.

Firstly the sample A is positively skewed whereas B has no skew. The mean of sample A is higher than its median, whereas the mean and median of sample B are approximately the same giving strong evidence for its centrality.
Secondly, sample A has a higher variance/standard deviation than B. The data in B has less bias due to greater centrality and having a lower variance than sample group A. However, this data is inconclusive as there needs to be proof of significance by a measure of p-value for example to test how much of the results are affected by chance, where lower p-values may be able to prove that the skewness in sample group A is significant.

1.6)







From the summaries that we did in the code below and the graphs above we could say that cars of type 1 are more efficient as they use the same gallon to travel further distances.

No, because we are adding another variable to the mix (cylinders) which would mess up our calculations as the number of cylinders increases the fuel usage.

Appendix R:

```
setwd("C:/Users/dbach/OneDrive/Desktop/C++VU/Statistics/R statistics")

#Starting Sample A
A = scan(file = "sampleA.txt") #Scanning the file
par(mfrow=c(1,2))
hist(A, ylim = c(0,25), col="chocolate", main = "Histogram of sample A",
xlab = "Range") #Creating histogram
boxplot(A, main = "Boxplot of sample A")#Creating boxplot
```

```
#Calculating the suitable numerical summaries
meanA = mean(A)
varA = var(A)
sdA = sd(A)

round(A, 2)

min(A)
max(A)

par(mfrow=c(1,1))
slices <- c(quantile(A))
pie(slices, col=rainbow(length(slices)), edges = 10, main="Quantile
distribution of data in sampleA")

#Starting Sample B
B = scan(file = "sampleB.txt")
par(mfrow=c(1,2))
hist(B, ylim = c(0,15), col="darkmagenta", xlab = "Range", main =
"Histogram of sample B")
boxplot(B, main = "Boxplot of sample B")

meanB = mean(B)
varB = var(B)

round(B, 2)

min(B)
max(B)

par(mfrow=c(1,1))
slices <- c(quantile(B))
pie(slices, col=rainbow(length(slices)), edges = 10, main="Quantile
distribution of data in sampleB")
```

```
#Starting mileage
source(file = "mileage.txt")

cars = c(mileage[[1]]) # getting the first bunch of information(cars type 1)
cyl = c(mileage[[2]]) # getting the second bunch of information(cyl type 1)

#Function for extracting only the cars with 4 cyl
searchNumOfCil = function(cars, cyl, result){

  for(i in 1:length(cars)){
    if(cyl[i] == 4){
      result = append(result, cars[i], after = length(result))
    }
  }
  return(result)
}

result1 = c() #Creating empty vector
resultMileage1 = searchNumOfCil(cars, cyl, result1) # All the type 1 cars
with 4 cyl

par(mfrow=c(1,2))
hist(resultMileage1, ylim = c(0,6), xlim = c(20,40), col="blue", times = 1,
main="Histogram of results for cars type 1 / 4cyl")
boxplot(resultMileage1,  main = "Boxplot of the same data")

#Calculating the suitable numerical summaries
meanMileage1 = mean(resultMileage1)
minMileage1 = min(resultMileage1)
maxMileage1 = max(resultMileage1)
varMileage1 = var(resultMileage1)

cars2 = c(mileage[[3]])
cyl2 = c(mileage[[4]])
```

```
result2 = c() #Creating empty vector
resultMileage2 = searchNumOfCil(cars2, cyl2, result2) # All the type 2 cars with 4 cyl

par(mfrow=c(1,2))
hist(resultMileage2, ylim = c(0,2), xlim = c(20,35), col="skyblue", main="Histogram of results for cars type 2 / 4cyl")
boxplot(resultMileage2,  main = "Boxplot of the same data")

#Calculating the suitable numerical summaries
meanMileage2 = mean(resultMileage2)
minMileage2 = min(resultMileage2)
maxMileage2 = max(resultMileage2)
varMileage2 = var(resultMileage2)
```