Enrico Dal Pos (2708533) & Osama Obaidat(2704856) & Justus Beck(2717047) - Group 64

Assignment 1

Exercise 1.1

a) Simple Random sample: sound, as it is very representative and unbiased towards the population

b) Cluster Sampling: flawed, as it leads to biased data as the mindset students/parents in each secondary school will be very similar

c) Voluntary response sample: flawed, subjects decide themselves to be included in the sample, which makes it biased. It is also not representative of the entire population because not everyone is reading this newspaper

Exercise 1.2

a) Ordinal. The summary is wrong since it is not possible to calculate the mean of qualitative data.

b) Interval, since people can be in debt. Nothing wrong with the summary as it's possible to take the mean and standard deviation to numerical data.

Exercise 1.3

a) Experiment: The drug that the subjects received modified their behaviour which causes some of them to experience nausea, therefore it is an experiment.

b) Systematic sampling, since the data collected is based on a fixed interval.

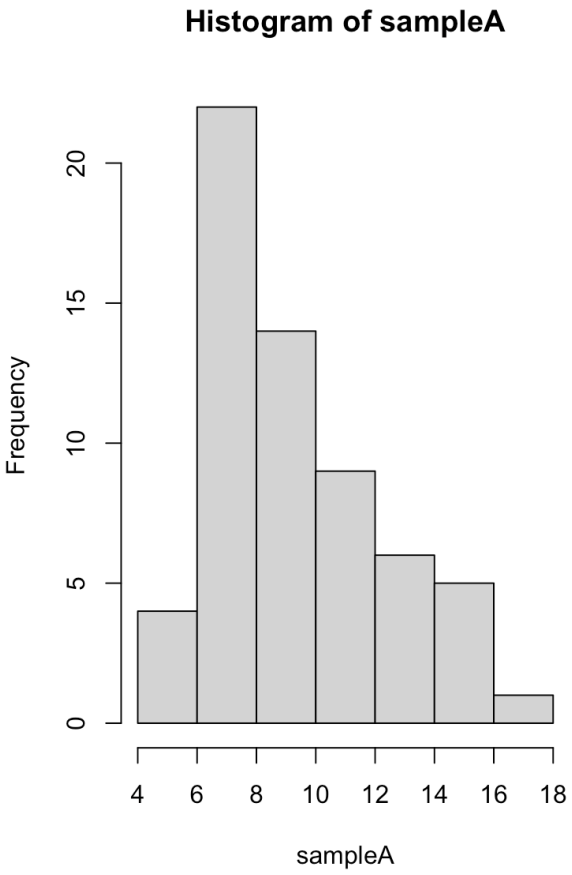
c) Cluster sampling, since random stations were chosen and all voters in them were surveyed

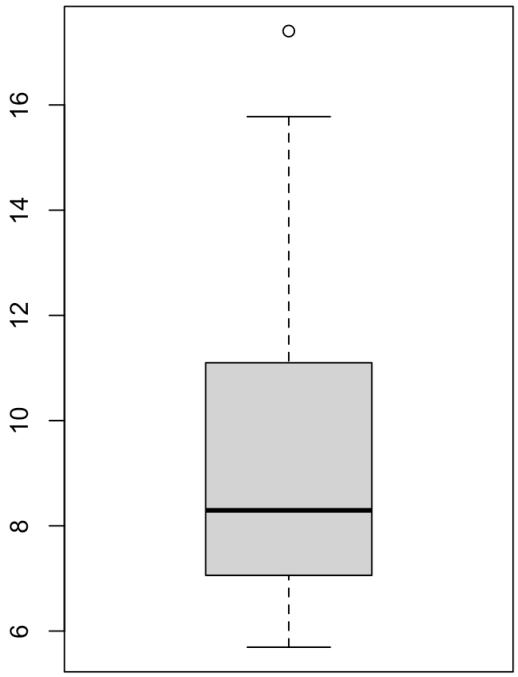
Exercise 1.4

a) There are some labels missing and the data would be better represented in a pie chart, because there are only a limited number of seats available and all of them need to be distributed over a fixed number of political parties.

b) Pareto chart. It can't be a pie chart, since it is not part of a whole, it can't be a histogram since it's qualitative, and between a bar chart and a Pareto chart, as it is easier to visualize in which areas more mistakes are made.

Exercise 1.5





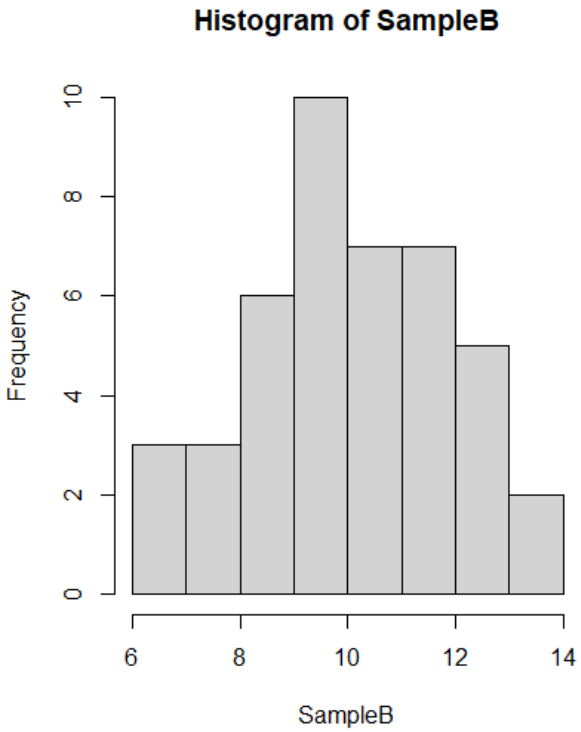
a)

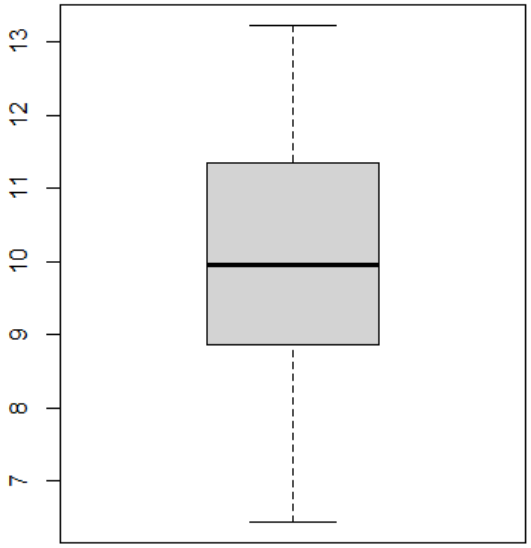
b) For the numerical location of the data we can use the median ≃ 8.29

For the numerical spread of the data we can use the standard deviation ≃ 2.86
The data was calculated using R.

c)

Location	The location is around 8
Spread/variation	Low spread
Shape	Unimodal
Range	10.2
Extremes	17.4
Accumulations	Around 6-10
Symmetry	Right-skewed





d)

For the numerical location of the data we can use the median ≃ 9.94

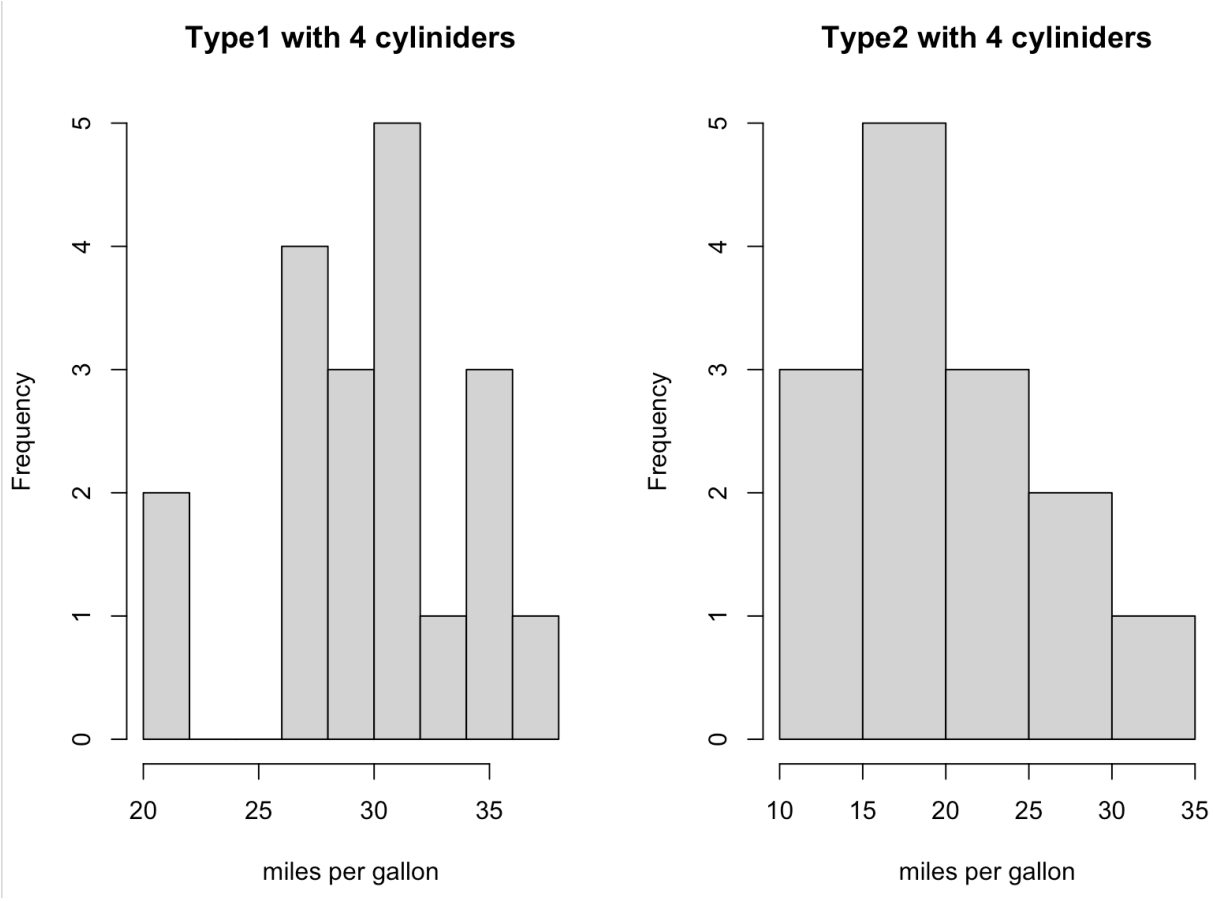
For the numerical spread of the data we can use the standard deviation ≃ 1.78
The data was calculated using R.

Location	The location is around 10
Spread/variation	Low spread
Shape	Unimodal
Range	6.8
Extremes	None
Accumulations	Around 9-12
Symmetry	Symmetrical

e) We think the data sets are not part of the same population since the symmetry and location are different.

Exercise 1.6

Graphical summary:



The location of the cars of type 1 is around 30, and the one of car 2 is around 17.
The shape of the first chart is unimodal, the shape of the 2nd chart is right-skewed and also unimodal.

Numerical summary:

Cars of Type 1

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
21.50	27.45	30.50	30.02	32.70	37.30

Standard deviation $\simeq$ 4.18

Cars of Type 2

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
13.30	15.85	18.95	20.19	23.68	30.40

Standard deviation $\simeq$ 5.24

The second type of car is more efficient than the first one since the mean is lower so for every mile less fuel is consumed.

When including the data for all cylinders for both cars we run into the risk that the two data samples have different cylinders so you can't measure which type is more fuel-efficient since more cylinders consume more fuel.
The mean result is biased since one of the data samples could have a bigger number of, for example, 6 cylinder cars, which would negatively influence the mean of that sample.

Appendix

Exercise 1.5

Relevant code:

```
sampleA = scan("sampleA.txt")
par(mfrow = c(1,2))
hist(sampleA)
boxplot(sampleA)


sampleB = scan("sampleB.txt")
par(mfrow = c(1,2))
hist(sampleB)
boxplot(sampleB)
```

Exercise 1.6

Relevant code:

```
source("mileage.txt")

mpg1_cyl4 = c()
mpg2_cyl4 = c()


for (i in 1: length(mileage$cyl1)) {
  if(mileage$cyl1[i] == 4) {
    mpg1_cyl4 <- append(mpg1_cyl4, mileage$mpg1[i])
  }
}

for (i in 1: length(mileage$cyl2)) {
  if(mileage$cyl1[i] == 4) {
    mpg2_cyl4 <- append(mpg2_cyl4, mileage$mpg2[i])
  }
}

par(mfrow = c(1,2))

hist(mpg1_cyl4, xlab = "miles per gallon", ylab =
"Frequency", main = "Type1 with 4 cyliniders")

hist(mpg2_cyl4, xlab = "miles per gallon", ylab =
"Frequency", main = "Type2 with 4 cyliniders")
```