Assignment 1

Hongyu He (2632195) & Bruno Hoevelaken (2645065)

Group CS 6

& Theoretical Exercises

Exercise 1.1

1. (Section 1.2 E10) The sampling method used in this scenario is **flawed**. *The book clearly states* (p7) **that internet polls**, in which people online can decide whether to respond, **are by their very nature seriously flawed**, as they result in **voluntary response samples**. This also means **nonresponses** might have occuered. This would result in **missing data** regarding individuals who were unwilling to participate in the online poll. Due to the missing data, we are unable extrapolate any meaningful conclusions out of this sample, regarding the wider populous. For a voluntary response sample, we can draw valid conclusions only about the specific group of people who chose to participate. From a statistical viewpoint, such a method is fundamentally flawed and should not be used for making general statements about a large population.
2. (Section 1.2 E12) In this case, the 1018 samples used by Gallup pollsters are *simple random samples* so the sampling method appears to be **sound**, assuming the subjects picked up and (willingly) participated in the survey.
3. (Section 1.2 E26) We believe that the sampling method is **flawed**. The exercise presents a similar scenario to that of "Section 1.1 E10", the online poll uses *voluntary response samples* (or *self-selected samples*). Specifically, the *Internet poll* method is used in this survey. According to the book: *"By their very nature, all are seriously flawed because we should not make conclusions about a population on the basis of such a biased sample."*. Another issue is the fact that it is unclear what the other 59% of users said. This exercise presents a scenario which suffers from the *same issue as exercise 10*, as it relies on an **internet poll**. Another issue is the fact that it is unclear what the other 59% of users said.

Exercise 1.2

1. (Section 1.3 E22) Depths (km) of earthquakes are at the **ratio** level of measurement because the depths (km) of earthquakes measured relative to the earth's crust(s), which means there is a meaningful **natural zero starting point**.
2. (Section 1.3 E32) The lineup numbers exist as distinguishing features, each corresponding to a player. This means that they fall under the **nominal** level of measurement. Thus, calculating their mean is pointless.

Exercise 1.3

1. (Section 1.4 E6) This should be an **experiment** since the researchers first applied some treatment to the subjects and then observed the effects on the **experimental units**.
2. (Section 1.4 E12) The researcher used **systematic sampling** because every 5 meters *(interval)*, a sample was selected.
3. (Section 1.4 E18) In this case, the sampling method used by ABC News should be **cluster sampling** where the voters were (implicitly) divide into clusters (stations) and then those stations were randomly selected. Finally, all members (voters) of the selected clusters are chosen.

Exercise 1.4

a.

- first off, the y-axis does not start from 0;
- the step size of the y-axis is too large, which makes hard reading;
- the bars are not in the same width.

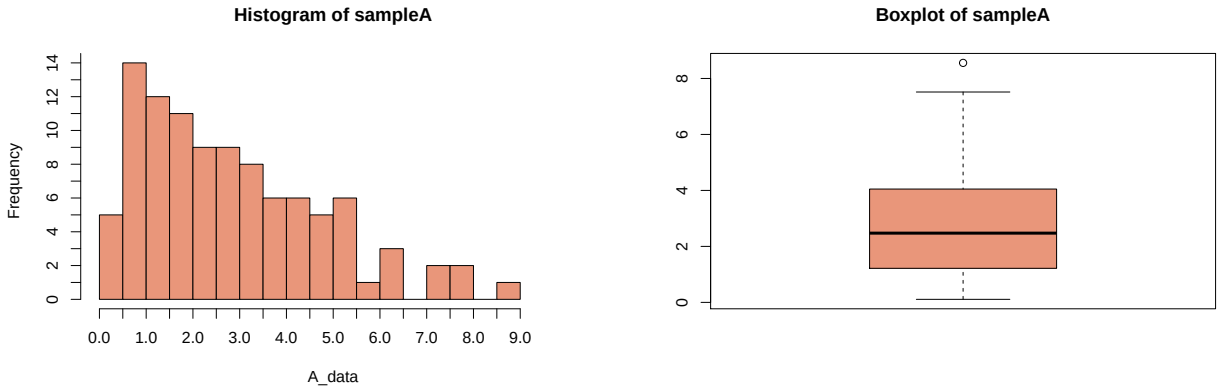
b.

1. Since the analysed data are at the *nominal level of measurement*, we think both **Pareto chart** and **pie chart** are good choices for illustrating the distribution of the food preference. In addition, the *relative frequencies* are needed to be calculated first for making a pie chart.
2. Due to the measured data are at the *ratio level of measurement*, we think both **histogram** and **boxplot** are illegible for showing the distribution of the numbers of previously done votes.

& R-exercises

Exercise 1.5

a.



b. Numerical summaries for *sampleA*

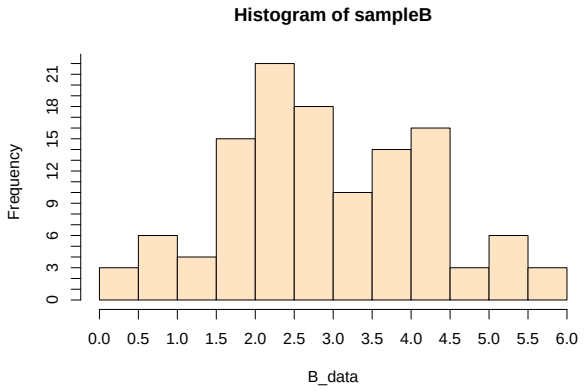
Min.	1st Quartile	Median	Mean	3rd Quartile	Max.	Variance	Standard deviaton
0.109	1.217	2.474	2.830	4.036	8.556	3.663	1.914

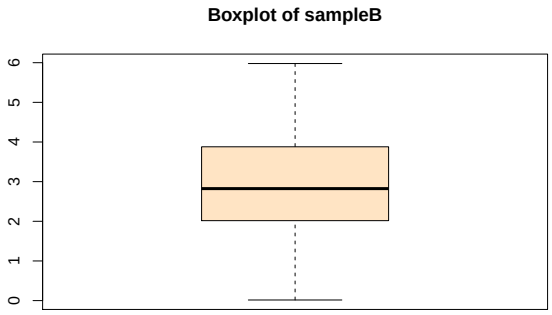
c.

- As we can see from the two diagrams, the data is **right-skewed**;
- Since the data is quite right-skewed, so the trend of the **cumulation** should be relatively low;
- Clearly, the data distributiion is **unimodal (mode/peak is around 0.8)** and not **symmetric**;
- Based on the **median** value, the **location** of the distribution is around 2;
- The **range** of the data is 8.556 - 0.109 = 8.447;

- There is one **extreme** value (**outlier**) which is 8.556;
- Comparing the (sample) mean and standard deviation, the **variation** of the distribution is quite big.

d. Plots and Numerical summaries for *sampleB*





Min.	1st Quartile	Median	Mean	3rd Quartile	Max.	Variance	Standard deviaton
0.015	2.017	2.824	2.929	3.874	5.979	1.705	1.306

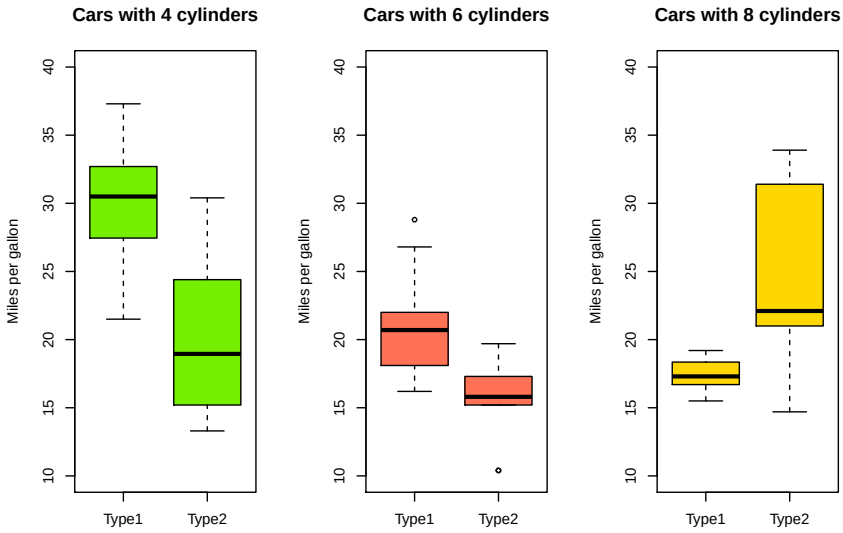
- As we can see from the two diagrams, the distribution is **roughly symmetric**;
- Compared to `A_data`, the **accumulation** of this data set is faster;
- The distribution is **unimodal** (mode/peak is around 2.2);
- Based on the **median** value, the location of the data is around 2.5;
- The **range** of the data is 5.963;
- There is no **extreme** value (**outlier**);
- Comparing the (sample) mean and standard deviation, the **variation** of the distribution is relatively small.

e.

According to the previous analyses on the two data sets, none of the characteristics (shape, mode, location, range, outliers, variation) of them are quite alike. Therefore, we strongly convinced that they are highly unlikely originated from the same population distribution.

Exercise 1.6

(1)



- Numerical summary of cars wtih 4 cylinders:

#	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
Type1	21.50	27.45	30.50	30.02	32.70	37.30
Type2	21.40	22.80	26.00	26.66	30.40	33.90

- Numerical summary of cars wtih 6 cylinders:

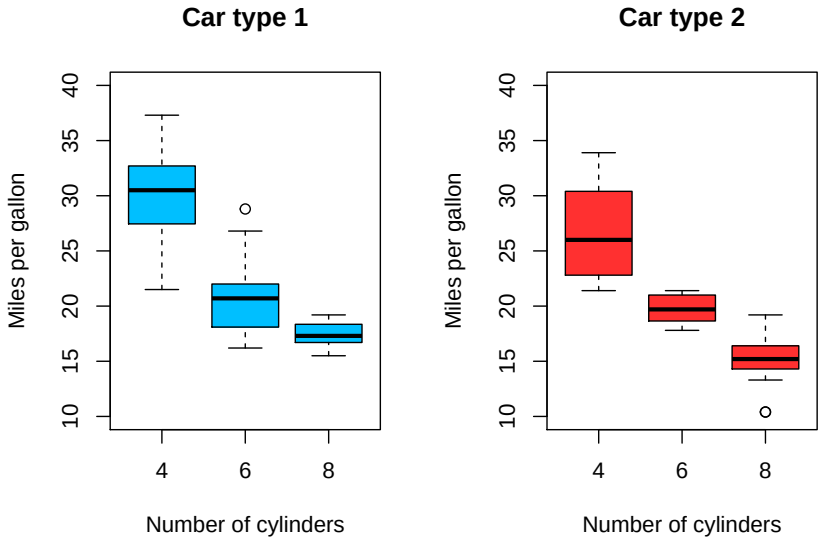
#	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
Type1	15.50	17.75	22.00	24.25	29.65	37.30
Type2	17.80	18.65	19.70	19.74	21.00	21.40

- Numerical summary of cars wtih 8 cylinders:

#	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
Type1	15.50	16.80	17.30	17.42	18.27	19.20
Type2	10.40	14.40	15.20	15.10	16.25	19.20

The 3 plots above compare the fuel usage of car type 1 & 2 under the same number of cylinders. As they illustrate, when the numbers of cylinders are 4 and 6, the type 2 cars are generally more fuel-efficient than type 1 cars, whilst for 8-cylinder cars, type 1 is more fuel-efficient than type 2.

(2)



No, we think there indeed a risk to directly compare the data of both types of cars when including also the cars with more cylinders, in the sense that we are comparing two data sets with **heterogeneous variations** of characteristics. As we can see from the two boxplots above, for both type 1 & 2, the cars of the same type with more cylinders are always more fuel-efficient than the ones with fewer cylinders.

Appendix

```
# Exercise 1.5
# a.
A_data = scan("sampleA.txt")

hist(A_data, breaks = 20, xlim = range(0, 10), axes = FALSE, col = "darksalmon", main = "Histogram of sa
xlabel <- seq(0, 9, by = .5)
axis(1, at = xlabel)
ylabel <- seq(0, 15, by = 1)
axis(2, at = ylabel)
boxplot(A_data, outlier.tagging = TRUE, col = "darksalmon", main = "Boxplot of sampleA")

# b.
summary(A_data)
var(A_data)
sd(A_data)

# d.
B_data = scan("sampleB.txt")

hist(B_data, breaks = 20, axes = FALSE, col = "bisque", main = "Histogram of sampleB")
xlabel <- seq(0, 6, by = .5)
axis(1, at = xlabel)
ylabel <- seq(0, 25, by = 1)
```

```r
axis(2, at = ylabel)
boxplot(B_data, outlier.tagging = TRUE, col = "bisque", main = "Boxplot of sampleB")

summary(B_data)
var(B_data)
sd(B_data)

# Exercise 1.6
# (1)
par(mfrow=c(1,3))
mileage = source("mileage.txt")

y_label = "Miles per gallon"
boxplot(mpg1[cyl1==4], mpg2[cyl1==4], ylim = range(10:40), names = c("Type1", "Type2"), ylab = y_label,
boxplot(mpg1[cyl1==6], mpg2[cyl1==6], ylim = range(10:40), names = c("Type1", "Type2"), ylab = y_label,
boxplot(mpg1[cyl1==8], mpg2[cyl1==8], ylim = range(10:40), names = c("Type1", "Type2"), ylab = y_label,

summary(mpg1[cyl1==4])
summary(mpg2[cyl2==4])
summary(mpg1[cyl2==6])
summary(mpg2[cyl2==6])
summary(mpg1[cyl1==8])
summary(mpg2[cyl2==8])

# (2)
par(mfrow=c(1,2))
x_names <- seq(4, 8, by = 2)
x_label = "Number of cylinders"

boxplot(mpg1[cyl1==4], mpg1[cyl1==6], mpg1[cyl1==8], ylim = range(10:40), names = x_names, xlab = x_lab
boxplot(mpg2[cyl2==4], mpg2[cyl2==6], mpg2[cyl2==8], ylim = range(10:40), names = x_names, xlab = x_lab
```