Assignment 1

Hongyu He (2632195) & Bruno Hoevelaken (2645065)

Group CS 6

Theoretical Exercises

Exercise 1.1

1. (Section 1.2 E10) This sampling method appears to be flawed because the survey is only available online using a voluntary response sampling method (internet polls), which means the people who respond are most likely to be the ones who use digital devices frequently. 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.
3. (Section 1.2 E26) Same as "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." The result, therefore, is most likely to be incorrect.

Exercise 1.2

1. (Section 1.3 E22) Depths (km) of earthquakes are at the ratio level of measurement.
2. (Section 1.3 E32) The numbers of starting lineup are at the nominal level of measurement. Thus, there is no reason to calculate the mean of them.

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, 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.

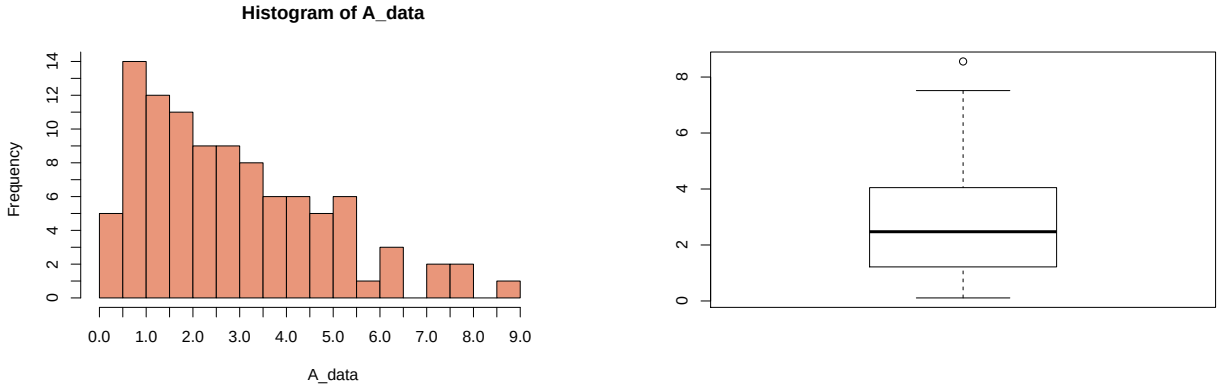
- first off, the y-axis does not start from 0;
- the step size of the y-axis is too large, which makes hard reading;
- the bars are not in the same width.

b.

1. Since the analysed data are at the nominal level of measurement, I 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, I think both histogram and boxplot are illegible for this case.

Exercise 1.5

a.



b.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1087  1.2170  2.4740  2.8300  4.0360  8.5560
```

```
## [1] "Variance:"
```

```
## [1] 3.662593
```

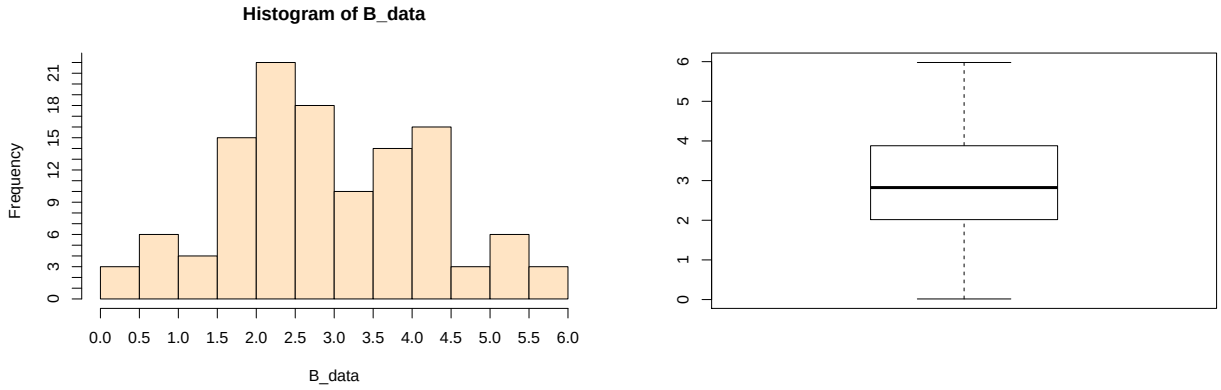
```
## [1] "Standard deviaton:"
```

```
## [1] 1.91379
```

c.

- As we can see from the two diagrams, the data is *right-skewed*;
- Since the data is quite right-skewed, the *accumulation* is relatively low;
- Clearly, this data is *unimodal* (mode/peak is around 0.8) and not *symmetric*;
- Based on the *median* value, the *location* of the data is around 2;
- The *range* of the data is 8.5560 - 0.1087 = 8.4472;
- There is one *extreme* value (*outlier*) which is 8.5560;
- Comparing the mean and standard deviation value, the *variation* of the data is quite big.

d.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01543 2.01700 2.82400 2.92900 3.87400 5.97900
```

```
## [1] 1.7051
```

```
## [1] 1.305795
```

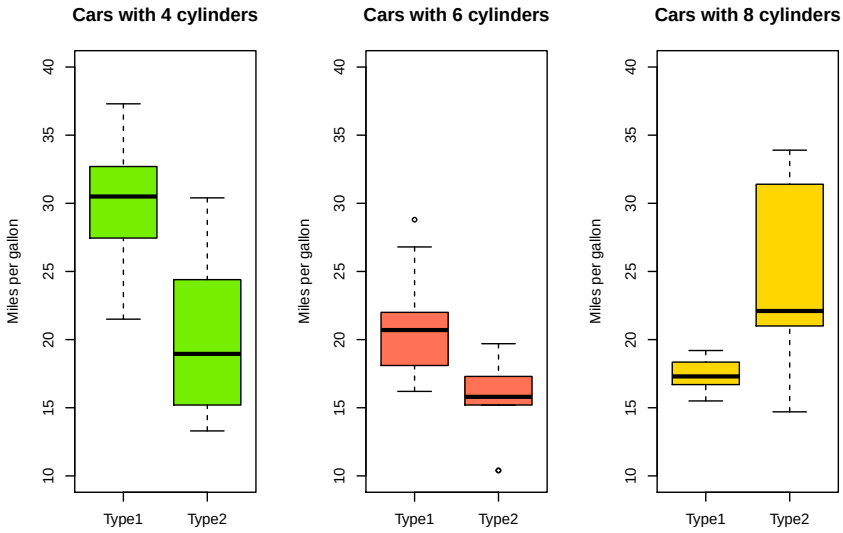
- As we can see from the two diagrams, the data is *roughly symmetric*;
- Compared to `A_data`, the *accumulation* of this data set is faster;
- This data is *unimodal* (mode/peak is around 2.2);
- Based on the *median* value, the location of the data is around 2.5;
- The *range* of the data is 5.96357;
- There is no *extreme* value (*outlier*);
- Comparing the mean and standard deviation value, the *variation* of the data is relatively small.

e.

According to the previous analyses on the two data sets, none of the characteristics (shape, accumulation, mode, location, range, outliers, variation) of them are alike. Therefore, I strongly convinced that they are not originated from the same population distribution.

Exercise 1.6

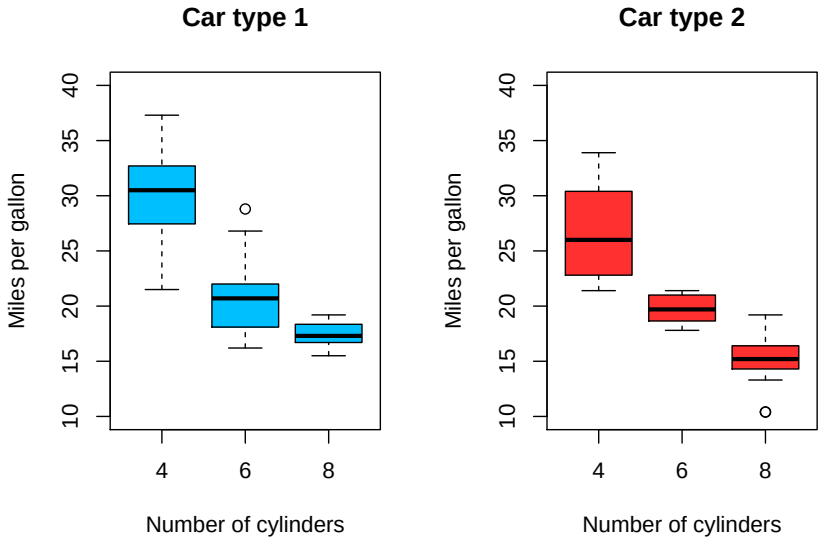
(1)




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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   21.40   22.80   26.00   26.66   30.40   33.90

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.50   17.75   22.00   24.25   29.65   37.30

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.80   18.65   19.70   19.74   21.00   21.40

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.50   16.80   17.30   17.42   18.27   19.20

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   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, I think there indeed a risk to directly compare the data of both types of cars when including also the cars with more cylinders. As we can see from the two boxplots above, for both type 1 & 2, the cars with more cylinders are always more fuel-efficient than the ones with fewer cylinders.