Statistical Methods Fall 2021

Assignment 1: Exploring and summarising data

Deadline: see Canvas

Topics of this assignment

The exercises below concern the topics covered in Lecture 1 and at the beginning of Lecture 2: data and summarising data. Before getting started with the assignment, study these topics. Numbers of exercises in the book refer to the 12th edition (New Pearson International Edition).

How to do the exercises? (Also take a look at Assignment_Example.pdf on Canvas!)

- Solve the exercises as efficiently as possible. Some exercises or their sub-questions of exercises do not require you to use R, while some others do. Write your report in English. To hand in: create a single PDF file of your work including your name and group number and upload it on Canvas.
- Data files and/or local R-functions needed for the assignment are available on Canvas.
- The text of the report should not exceed 4 pages, this is excluding figures and the appendix with R-code.
- It is important to make clear in your answers how you have solved the questions: do not only give answers and results, but also motivate your answers. Put the **relevant, executable, and copiable** R-code (without the prompt sign ">") ***in an appendix***. Do not copy R-code in the answers themselves, and only include in the appendix the code that led to your answer. Do not put entire data sets in the appendix.
- Graphs should be made and viewed on screen first; put the final version in your report. Multiple graphs can be put into one figure using the command `par(mfrow=c(k,r))`, see `help(par)`. Make sure the dimensions of the graphs are adequate and that figures are concise: a single figure should not take up a whole page.
- In your report, round the results that you obtained from R to a suitable number of digits.

Not adhering to these rules may have as a consequence that some of your points will be deducted!

Theoretical exercises

Exercise 1.1 *In a)-c), name the chosen sampling method and determine whether the sampling method seems to be sound or is flawed. Always motivate your choice.*

a) In a survey on COVID-19 vaccinations, the Dutch Central Bureau of Statistics randomly selected and mailed 2052 teens (aged 12-17) about their vaccination status.

b) In another survey on COVID-19 vaccinations, the Dutch Central Bureau of Statistics randomly selected 20 secondary schools in The Netherlands and asked all of their students about their vaccination status.

Also, explain what is wrong in c).

c) In an online poll conducted by a big Dutch online newspaper, 5012 Internet users chose to respond, and 76% of them stated that they were fully vaccinated against COVID-19.

Exercise 1.2 *Determine which of the four levels of measurement (nominal, ordinal, interval, ratio) is most appropriate. Also, if there is anything wrong with the given summary statistics, explain what is wrong.*

a) A survey about the public transport system should measure the agreement to the statement "I enjoy taking off-peak Intercity trains." The survey used a 5-point Likert scale with the following options and numbers of replies: "Strongly disagree" (11 times), "Disagree" (7), "Neither agree nor disagree" (13), "Agree" (20), "Strongly agree" (37). The statisticians who conducted the study reported that the mean was slightly below "Agree".

b) A Dutch bank analyzed the balance of 1000 randomly selected bank accounts of their customers. The first 5 balances (in Euros) were: 1,167.51, 2,614.12, 4,698.06, -152.63, 307.95. The average (mean) balance of all 1000 accounts was 1,714.42 Euros with a standard deviation of 625,63 Euros.

Exercise 1.3 *In a), determine whether the given description corresponds to an observational study or an experiment. Give a brief explanation of your choice.*

a) In a clinical trial of the cholesterol drug Lipitor, 190 subjects were given 21-mg doses of the drug, and 3.8% of them experienced nausea.

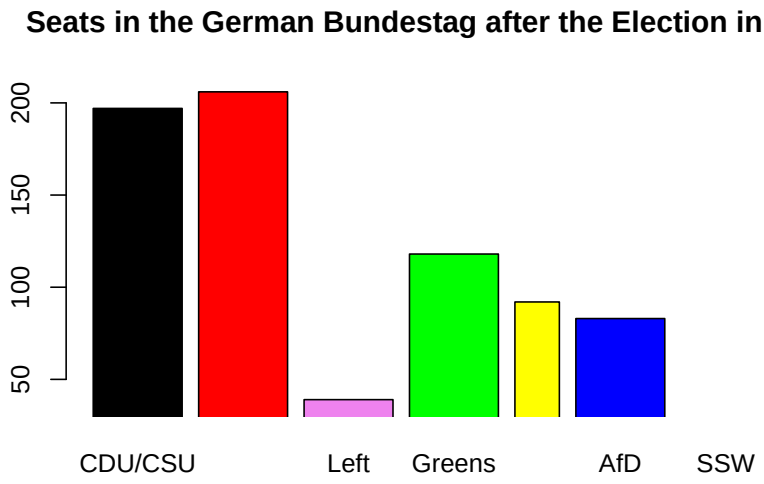
In b) and c), identify which of these types of sampling is used: random, systematic, convenience, stratified, or cluster.

b) When collecting data from different sample locations in a lake, a researcher uses the "line transect method" by stretching a rope across the lake and collecting samples at every interval of 10 meters.

c) On the day of the last presidental election, a television channel organized an exit poll in which specific polling stations were randomly selected and all voters were surveyed as they left the premises.

Exercise 1.4

a) The graph below shows the seat distribution in the 20-th German Bundestag (after the election in September 2021). The seat distribution is as follows: CDU/CSU: 197, SPD: 206, Left: 39, Greens: 118, FDP: 92, AfD: 83, SSW: 1. What is wrong with the presentation?



b) Human resources departments of several IT companies were surveyed about areas in which job applicants make mistakes (multiple choices possible). The areas found in the survey were: interview, résumé, cover letter, reference checks, interview follow-up. Which of the following graphs would be best for describing the mistakes: histogram; bar chart; Pareto chart; pie chart?

R-exercises

Hints concerning R:

- For the exercises below you can use, for instance, the *R*-functions `hist`, `boxplot`, `mean`, `median`, `sd`, `min`, `max`, and `summary`. If necessary, experiment with the different options these functions have.
- The *R*-function `quantile(x,`α) gives the α-quantile of the values in the vector `x`. For example, `quantile(x,0.25)` gives the first quartile of x. Instead of one single value, also a vector $(\alpha_1, \alpha_2, \ldots, \alpha_k)$ can be inserted for the parameter α in `quantile`. Check which output this function gives when the parameter α is not specified.

Exercise 1.5 *Always describe the most important findings in numerical and graphical summaries.*

a) Make a suitable histogram and boxplot for the data in the file `sampleA`.

b) Give one or more suitable numerical summaries for the location and the spread of the distribution of these data.

c) Based on your summaries in parts a) and b), briefly answer for this data set as many of the basic questions (location, spread/variation, range, extremes, accumulations, symmetry,...) about the data distribution as possible.

d) Perform parts a), b) and c) for the data in the file `sampleB`.

e) Based on all results of parts a)–d), do you think that the two data sets originate from the same population distribution? Why or why not?

Exercise 1.6 In the file `mileage` you can find data about fuel usage of cars. The first two components in the list give the fuel usage in miles per gallon and the number of cylinders of cars of type 1. The third and fourth component give the same quantities for cars of type 2. Look at the data first.
Make appropriate graphical and numerical summaries of those cars of type 1 which have 4 cylinders. Repeat the same for the cars of type 2 which have 4 cylinders.
Comment on these summaries. Is one type more fuel efficient than the other?
Is it without any risk to directly compare the data of both types of cars when including also the cars with more cylinders?

Assignment 1
Statistical Methods 2021

Project Group 3:
Caspar M. S. Grevelhörster (2707848),
Zülfiye Tülin Manisa (2664758), and
Sid R. Singh (2708756)

Vrije Universiteit Amsterdam,
De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

Exercise 1

a) The method used here is stratified sampling, because the population is divided into a subgroup, ages 12-17, and then a simple random sample is drawn from this group. Depending on the way you would look at this it can be a sound or flawed method. In case the researchers desire to conclude something about the entire population, it would be flawed since only the ages 12-17 are represented. However, if the researchers desire to conclude something about the ages 12-17, it is a sound method. It is a category in which the subjects have the same characteristics and in addition to that everyone in the subgroup has a random and equal chance to be chosen.

b) The method used here is cluster sampling, the population is divided into clusters: secondary schools. Afterwards, 20 random schools were picked as clusters, of which all subjects are asked the same question. This method is biased and flawed, because it does not represent the whole population, nor the subgroup of students attending secondary school. The students within a secondary school may have very similar opinions due to confounding variables such as the city they are located in, their level of education or social status et cetera.

c) The method used here is voluntary response sampling, because the poll is published online and anyone who wants to, can answer the poll. It is a flawed and biased method of sampling, because the poll is directed towards a specific audience, the readers of the online newspaper. Furthermore, the group of unvaccinated people may be more reluctant to answer this poll in comparison to the group of vaccinated people and therefore it is not reliable.

Exercise 2

a) The most appropriate level of measurement for this survey is the ordinal measurement because there is a ranking and ordering in the measurements (strongly agree to strongly disagree), but the differences between each rank are not clearly specified or numeric. The summary statistics use the mean value, which is not sensible when dealing with ordinal measurements, because, as stated above, the differences between each rank are not clearly specified. Thus, calculating a mean does not make sense.

b) The most appropriate level of measurement for this analysis is the interval measurement. This is the case because there are meaningful differences between the values, but there is no natural starting point for the balance of an account (because the customers of the bank can have debt). The summary statistics in this analysis depicts the mean and standard deviation which makes sense because with the mean, the average amount of money of all the customers of the bank can be seen. The standard deviation shows the bank how spread out the balances of the customers are in relation to the mean. However, if the bank detects a lot of outliers in the data (people who are extremely rich or heavily in debt), it should consider using the median in the summary statistics, because this value, in contrast to the mean, is not affected by outliers.

Exercise 3

a) The given description corresponds to an experiment. In an observational study the environment of the subjects is not being manipulated. Since the subjects in this description were given a drug, their environment was manipulated which makes this trial an experiment.

b) Systematic sampling was used by this researcher, since they collected samples according to a fixed interval which makes it a systematic approach.

c) For this survey cluster sampling was used, because first the TV station randomly picked a few clusters (polling stations) and then surveyed all subjects in the respective clusters.

Exercise 4

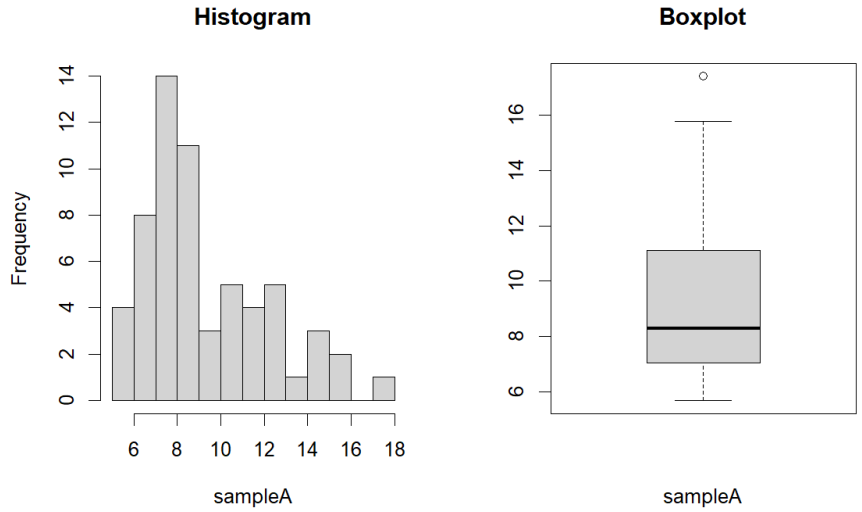
a) The scaling on the y-axis does not start from 0 which makes the differences between the numbers of seats between the parties seem larger than they are. For example, according to the data, the CDU has about 5 times more seats than the Left party but in the graph it seems like the difference is even more dramatic.

b) Since the data is not quantitative, we cannot use a histogram. A bar chart would be a sensible choice, but to visualize which mistakes are most common, it would make sense to use an ordered bar chart which is a Pareto chart. However, to see how often each mistake is mentioned in comparison to all the mistakes, the most suitable graph is the pie chart, since it shows proportions intuitively.

Exercise 5

a) The following figure shows a histogram as well as a boxplot that were both generated with R from the provided sampleA.txt-file. The utilized code can be found in Appendix 1.

Fig. 1: Histogram and Boxplot. sampleA.



b) Table 1 shows percentiles of the sampleA-dataset which can be used to understand both the location as well as the spread of the data. The table is about the central tendency; the standard deviation of $\sigma = 2.86$ shows that a significant proportion of the data is narrowly clustered around the mean of 9.29.

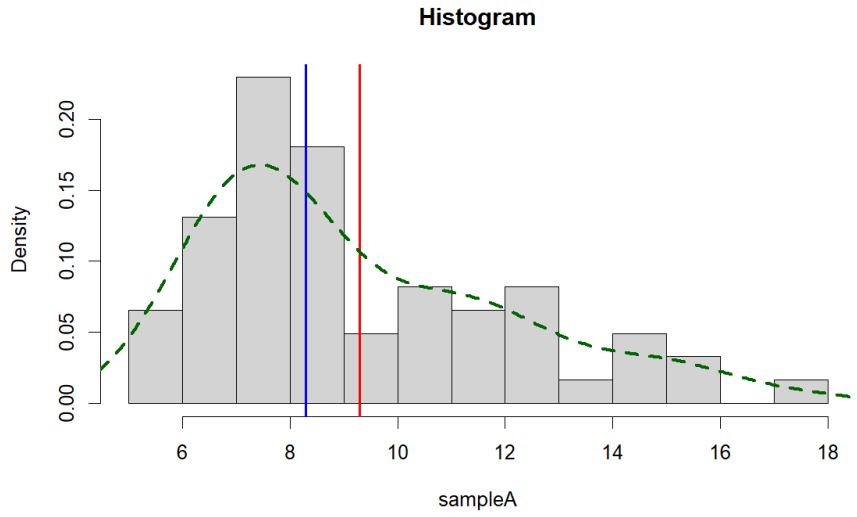
Table 1: Measures of Location. sampleA.

Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	σ
5.69	7.06	8.29	9.29	11.10	17.40	2.86

The table also shows the measures of spread; its values depict how varied (i.e. spread out) the data is. Half of the values lie within the upper and lower margins of 7.06 and 11.10. These values make up the respective lower and upper edge of the grey box in the boxplot of figure 1 as well as the interquartile range, meaning that 50% of the values are between 7.06 and 11.10.

c) To further visualize the data from the dataset, we created an additional graph (fig. 2) from figure 1 and table 1. The dataset is ASYMMETRICAL. This can be determined using the boxplot (fig. 1) by analyzing the distance of the two whiskers from the black bar representing the median. Since the two differ in length significantly, the sample data is asymmetrical. The asymmetry can also be seen in the approximated density histogram (fig. 2) since the shape of the smooth curve gives and idea of data distribution.
The LOCATION of the data of an approximately symmetrical dataset can be determined using the mean (red line in fig. 2). The latter is, however, easily influenced by outliers (one can be seen represented in the boxplox as the circle lying outside of the whiskers) which is not the case for the median (blue line in fig. 2) since this measure

Fig. 2: Approximated Density Histogram. sampleA.



is based on ranks. For the data in sampleA, the center location is therefore at approximately 8.29, represented by the median in table 1 or the blue line in figure 2. The center of the data is the location where the biggest ACCUMULATIONS of values lie. These accumulations are represented by the higher bars of the histogram.
A boxplot (fig. 1) gives information about the SPREAD (or dispersion) of data, which includes the interquartile range as well as the overall range. The interquartile range is represented through the height of the grey boxes in the boxplot and can be measured as already described in B). The RANGE is determined by subtracting the lowest value (Min. in table 1) from the largest (Max. in table 1) value in a dataset. Thus, the range is 17.40 - 5.69 = 11.71. The lowest and largest values (also: 0th and 100th percentile respectively) represent the EXTREMES.

d) Like in **a)**, the histogram and boxplot (fig. 3) where computed but this time from the provided sampleB.txt-file. Additionally, the measures of spread (quantiles, table 2) were computed again.

Fig. 3: Histogram and Boxplot. sampleB.

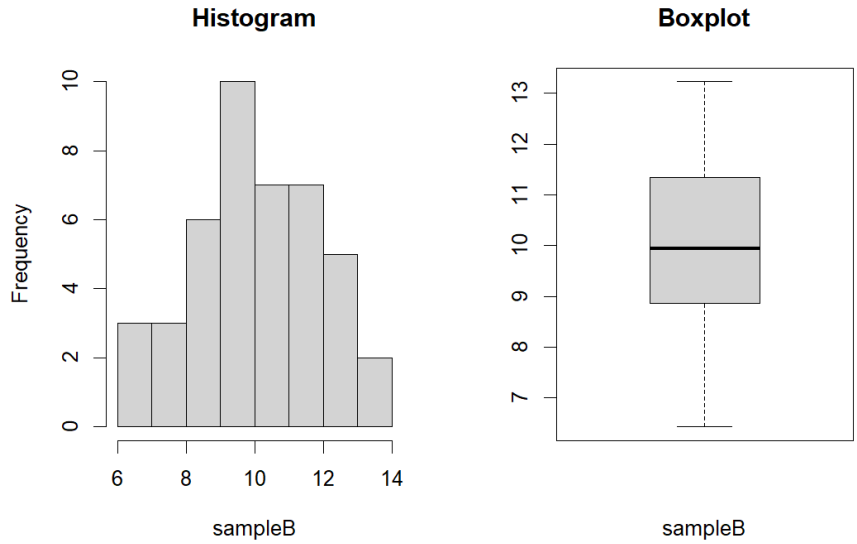


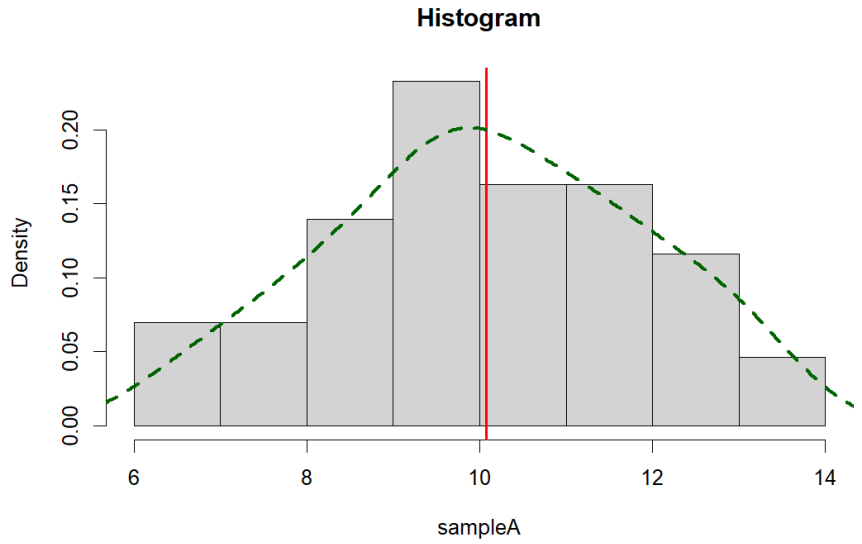
Table 2: Measures of Location. sampleB.

Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	σ
6.43	8.86	9.94	10.08	11.35	13.23	1.78

The standard deviation of $\sigma = 1.78$ (table 2) is lower than the one in table 1 which means that the data is more narrowly clustered than in sampleA. The quantiles show that half of the values lie within the upper and lower margins of 8.86 and 11.35. These values make up the respective lower and upper edge of the grey box in

the boxplot of figure 3 as well as the interquartile range. In contrast to sampleA, sampleB is SYMMETRICAL. This can be determined using the boxplot (fig. 3) by analyzing the distance of the two whiskers from the black bar representing the median. Since the two do not differ in length significantly, the sample data is nearly symmetrical. The symmetry can also be seen in the approximated density histogram (fig. 4) since the shape of the smooth curve gives and idea of data distribution.

Fig. 4: Approximated Density Histogram. sampleB.



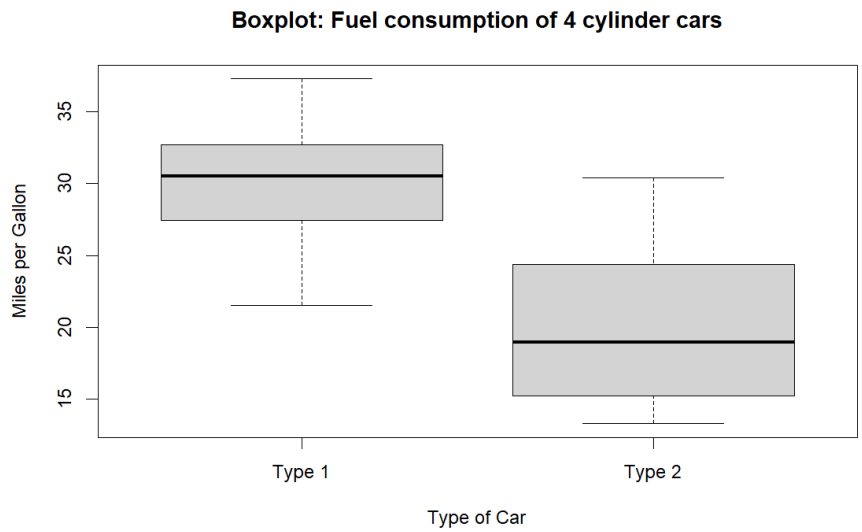
The LOCATION of the data of an approximately symmetrical dataset can be determined using the mean (red line in fig. 4). For the data in sampleB, the center location lies therefore at approximately 10.08, represented by the mean in table 2 or the red line in figure 4. In this region, the ACCUMULATIONS can also be seen by the higher bars of the histogram. The RANGE of sampleB is computed (see section **c)** for comparison) as 13.23 - 6.43 = 6.8 which is lower than for sampleA. Eventhough there exists a lower range, the SPREAD of the data is still similar (compare boxplots fig. 1 and 3). The EXTREMES of the dataset are represented by the minumim and maximum values found in table 2.

e) The two datasets likely do not originate from the same population distribution. There are several points that hint towards this: Differing symmetry - while sampleA is asymmetrical and right skewed, sampleB is symmetrically distributed. Lower range - the range of sampleA (11.71) is almost double of the sampleB-range (6.8). Different locations - the two samples are located around different values (sampleA: 8.29, sampleB: 10.08).

Exercise 6

In order to compare 4 cylinder cars of two different types, two boxplots were made depicting the data location of the car types regarding their miles-per-gallon-efficiency:

Fig. 5: Boxplot.



Additionally, a table was made providing a numerical overview regarding the two car types:

Table 3: Measures of Location. Comparing car type 1 to type 2

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	σ
Type 1	21.50	27.45	30.50	30.02	32.70	37.30	4.18
Type 2	13.30	15.85	18.95	20.19	23.68	30.40	5.24

The two boxplots in figure 5 show the efficiency of the two car types. It is visible that on average, cars of type 1 are more fuel efficient than cars of type 2 because they are centered around different medians (black bars in boxplots fig. 5 and mean in table 3) and their interquartile ranges (where 50% of their respective data is located) do not overlap. However, this does not mean that *all* type1-cars are more fuel efficient than type2s but it can be expected that when choosing a car of type 1, it is likely to get further with the same amount of fuel than one of type 2.

It is not without risk to directly compare the data of both types of cars when including the cars with more cylinders. This is due to the fact that the number of cylinders can be seen as a confounding variable because this number strongly affects the dependent variable which is the fuel consumption in MPG. This means that it would no longer be possible to make a statement about which type of car (Type 1 or Type 2) is better when it comes to fuel efficiency, because differences in the MPG ratings of the two types could be for example due to one type having more 8 cylinder cars listed in the dataset than the other type.

Appendix

Appendix 1: Code of exercise 5 figure 1

```
par(mfrow = c(1, 2))
s = scan("sampleA.txt")
hist(s, breaks = 12, main = "Histogram", ylab = "Frequency",
    xlab = "sampleA")
boxplot(s, main = "Boxplot", xlab = "sampleA")
summary(s)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.693   7.059   8.294   9.291  11.099  17.404
```

```
quantile(s)
```

```
##        0%       25%       50%       75%      100%
##  5.693107  7.058678  8.294141 11.099486 17.404498
```

```
sd(s)
```

```
## [1] 2.862244
```

Appendix 2: Code of exercise 5 figure 2

```
par(mfrow = c(1, 1))
s = scan("sampleA.txt")
hist(s, breaks = 12, main = "Histogram", ylab = "Density",
    xlab = "sampleA", probability = TRUE)
abline(v = median(s), col = "blue", lwd = 2)
abline(v = mean(s), col = "red", lwd = 2)
lines(density(s), type = "l", lty = 2, lwd = 3, col = "darkgreen")
```

Appendix 3: Code of exercise 5 figure 3

```
par(mfrow = c(1, 2))
s = scan("sampleB.txt")
hist(s, breaks = "FD", main = "Histogram", ylab = "Frequency",
    xlab = "sampleB")
boxplot(s, main = "Boxplot", xlab = "sampleB")
summary(s)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.434   8.858   9.944  10.078  11.347  13.232
```

```
quantile(s)
```

```
##        0%       25%       50%       75%      100%
##  6.434261  8.858168  9.943619 11.347302 13.232243
```

```
sd(s)
```

```
## [1] 1.784021
```

Appendix 4: Code of exercise 5 figure 4

```
par(mfrow = c(1, 1))
s = scan("sampleB.txt")
hist(s, breaks = "FD", main = "Histogram", ylab = "Density",
    xlab = "sampleA", probability = TRUE)
abline(v = mean(s), col = "red", lwd = 2)
lines(density(s), type = "l", lty = 2, lwd = 3, col = "darkgreen")
```

Appendix 5: Code of exercise 6 figure 5

```
source("mileage.txt")
par(mfrow = c(1, 1))

cyl1 = mileage[["cyl1"]]
mil1 = mileage[["mpg1"]]
cyl2 = mileage[["cyl2"]]
mil2 = mileage[["mpg2"]]
i_cyl1 = which(cyl1 %in% c("4"))
i_cyl2 = which(cyl2 %in% c("4"))
# make both lists same lenghts
length(i_cyl1) = max(length(i_cyl1), length(i_cyl2))
length(i_cyl2) = max(length(i_cyl1), length(i_cyl2))
# combine to matrix
m = cbind(i_cyl1, i_cyl2)


for (row in 1:nrow(m)) {
    i_type1 = m[row, 1]
    i_type2 = m[row, 2]
    m[row, 1] = mil1[i_type1]
    m[row, 2] = mil2[i_type1]
}

colnames(m)[1] <- "Type 1"
colnames(m)[2] <- "Type 2"

boxplot(m, main = "Boxplot: Fuel consumption of 4 cylinder cars",
    ylab = "Miles per Gallon", xlab = "Type of Car")
summary(na.omit(m[, "Type 1"]))

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

sd(na.omit(m[, "Type 1"]))

## [1] 4.182447

summary(na.omit(m[, "Type 2"]))

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

sd(na.omit(m[, "Type 2"]))

## [1] 5.238446
```