ST102/ST109 Exercise 1

Practicalities: Each week you should spend a reasonable amount of time on each set of exercises. Do not worry if you have been unable to complete all questions, as long as you have done your best. Exercise 1 does **not** need to be submitted. It covers aspects of descriptive statistics, which many of you may have already covered in previous studies. Solutions to this problem set will be covered in your first class, in the week commencing 2 October 2023.

In Questions 4 and 5 you are asked to use R to calculate some descriptive statistics. You do not need to print out any R output, however you are encouraged to experiment with R as some familiarity with R is highly desirable on a CV for anyone considering a career in finance! 😉

You should also be working through the 'R for Statistics' self-study course developed by LSE's Digital Skills Lab, accessible via the ST102 Moodle page.

R is a free software environment for statistical computing and graphics, downloadable from:

`https://www.r-project.org/`

with helpful reference manuals accessible at:

`https://cran.r-project.org/manuals.html`.

1. (a) Uncle Sam is interested in the weather in London, so he looks up some statistics on temperatures. He learns that the average of the 365 daily maximum temperatures in London in 2020 was 14.0 degrees Celsius (°C), and the standard deviation was 7.7 °C.

   Uncle Sam is American, so he is more familiar with temperatures expressed in degrees Fahrenheit (°F). What would you guess are the average and standard deviation of these same 365 temperatures, expressed in °F?
   Note that the transformation formula between the two scales is $\text{F} = (9/5)\text{C} + 32$.

   (b) Let $X_1, X_2, \ldots, X_n$ be a random sample of n observations of a variable X. Define a new variable Y obtained from X as $Y_i = aX_i + b$, for $i = 1, 2, \ldots, n$, where a and b are constants, i.e. known numbers which are the same for all observations $i = 1, 2, \ldots, n$. Let $\bar{X}$, s_X^2 and s_X denote the sample mean, variance and standard deviation of X, respectively. Similarly, let $\bar{Y}$, s_Y^2 and s_Y denote the sample mean, variance and standard deviation of the transformed variable Y, respectively. Show that:

   i. $\bar{Y} = a\bar{X} + b$

   ii. $s_Y^2 = a^2 s_X^2$

   iii. $s_Y = |a| s_X$.

   (c) Use (b) to answer (a). Were your guesses in (a) correct?

2. Show that:

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2.$$

3. A group of 44 students was asked to guess, to the nearest metre, the width of the lecture hall in which they were sitting. The true width of the hall was 13.1 metres. The students' guesses are shown in Table 1 below.

   (a) Create, by hand, a frequency table of the values of the guesses. Also, include relative frequencies (percentages) and cumulative percentages in the table.

   (b) Calculate, by hand, the sample mean, median and standard deviation of the guesses. Remember that you can use the table you produced in (a) to make the calculations easier.

Table 1: The data for Question 3.

22	15	25	15	10	17	18	12	16	15	15	15	9	40	10
14	11	11	15	16	38	14	13	17	17	13	20	10	14	35
17	15	13	12	8	11	18	10	10	11	10	27	16	15	

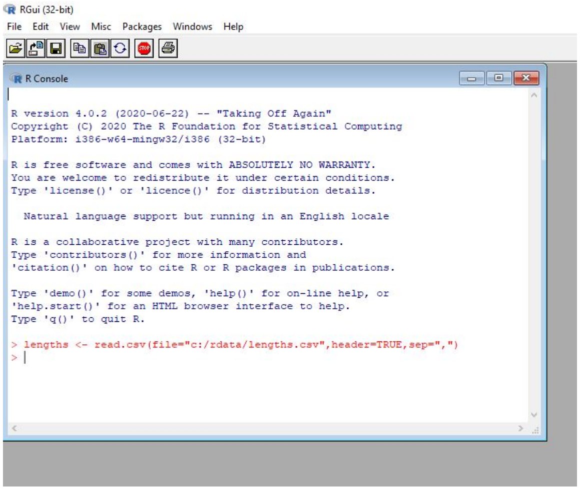
Source: Hand et al. (1994). *A Handbook of Small Data Sets.* The data were collected by Professor T. Lewis.

4. The file `lengths.csv` (on Moodle) contains the same data as in Table 1. Use R to calculate the same statistics you calculated by hand in Question 3 (b).

   First, download the file `lengths.csv` from Moodle and save it to an appropriate folder. Here, I will assume the file has been saved in a folder '`rdata`' in the '`c`' drive, i.e. in `c:/rdata`. To read in the file to R, first open R (having installed it!) and use the command:

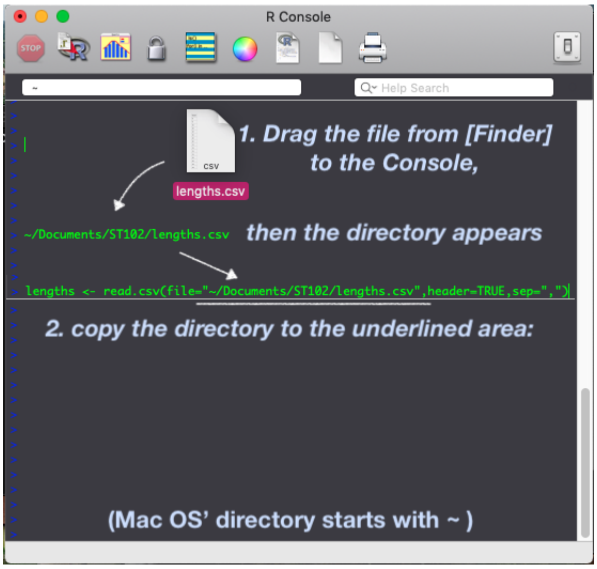
```
> lengths <- read.csv(file="c:/rdata/lengths.csv",header=TRUE,sep=",")
```

   If you save the file in a different location/folder, then replace `c:/rdata` as appropriate.



```
R version 4.0.2 (2020-06-22) -- "Taking Off Again"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> lengths <- read.csv(file="c:/rdata/lengths.csv",header=TRUE,sep=",")
> 
```

For Mac users, the following should work:



1. Drag the file from [Finder] to the Console,

```
> ~/Documents/ST102/lengths.csv
```

then the directory appears

```
> lengths <- read.csv(file="~/Documents/ST102/lengths.csv",header=TRUE,sep=",")
```

2. copy the directory to the underlined area:

(Mac OS' directory starts with ~)

To view the summary statistics use:

```
> summary(lengths$Guess)          > mean(lengths$Guess)
> median(lengths$Guess)           > sd(lengths$Guess)
```

5. The data file `gsseducation.csv` contains data for 1,445 respondents in the 1998 round of the U.S. General Social Survey (GSS). The GSS is a survey of the characteristics and attitudes of the general U.S. population. Here we use descriptive statistics from R to examine some variables on education. First, consider the following four variables:

- `EDUC`: the number of years of education that the survey respondent has completed
- `SPEDUC`: years of education completed by the respondent's spouse
- `PAEDUC`: years of education completed by the respondent's father
- `MAEDUC`: years of education completed by the respondent's mother.

(a) Draw a histogram of `EDUC`. What does it tell you?
Hint: Use the following:

```
> gss <- read.csv(file="c:/rdata/gsseducation.csv",header=TRUE,sep=",")
> hist(gss$EDUC,breaks=20)
```

(b) Calculate the sample mean, median, standard deviation, and first and third quartiles of the four variables. Also, draw side-by-side boxplots of the variables. Compare the results for the respondents and their spouses to the results for the fathers and mothers, i.e. effectively between two generations. What do you observe?
Hint: For example, for `EDUC`, use:

```
> summary(gss$EDUC)
> sd(gss$EDUC, na.rm=TRUE)
```

which removes the effect of missing values. (Replace `gss$EDUC` with `gss$SPEDUC` etc. to work with the other variables.)

For side-by-side boxplots, use:

```
> attach(gss)
> varnames <- c("EDUC","SPEDUC","PAEDUC","MAEDUC")
> boxplot(EDUC,SPEDUC,PAEDUC,MAEDUC,names=varnames)
```

(c) Draw a scatterplot with `EDUC` on the horizontal axis (x-axis) and `SPEDUC` on the vertical axis (y-axis). Here it is of interest to examine whether there seems to be any association between the two, i.e. whether, for example, people with high levels of education tend to have spouses who also have high levels of education. This would be evidence of *homogamy*, i.e. people tending to marry people who have similar characteristics to themselves. What do you conclude from the plot here?

Hint: Use `> plot(EDUC,SPEDUC)`