# R-PROJECT 1 FOR LECTURE 1: R AS A CALCULATOR, AND AN ELEMENTARY PROBLEM IN BIG DATA ANALYTICS GRA4110

STEFFEN GRØNNEBERG

Please carefully read through all text.

**Please note:**

You are expected to base your work on `intro.R`, and fill in your answers in `1.R`, both files are uploaded to Itslearning.

**It is very important** that you write down all R-code needed to solve the following problems in a script (and I encourage you to base your work on `1.R`). On the home-exam, which will focus on R, I will expect this from you, and you ought to get into the habit of doing this right away.

If you do not finish this assignment set during class, **you have to finish it on your own**. Therefore try to work as efficient as you can.

Solutions will be posted at the end of the week. The delay comes from wishing to *strongly* encourage you to solve this problems *without looking at solutions*. You will learn much more from this.

## 1. R as a calculator

Read carefully through "intro.R" and use this code to solve the following problems.

(A) The line

$$Y = a + bX$$

which passes the points $(X_1, Y_1)$ and $(X_2, Y_2)$ has slope

$$b = \frac{Y_1 - Y_2}{X_1 - X_2}$$

and intercept

$$a = \frac{Y_2 X_1 - Y_1 X_2}{X_1 - X_2}.$$

Use R to find the slope and intercept when $X_1 = 3, X_2 = -4$ and $Y_1 = 2, Y_2 = 100$.

*Hint:* The answer is $a = 44$ and $b = -14$.

(B) Use R to find the slope and intercept when $X_1 = 0, X_2 = -11$ and $Y_1 = -2, Y_2 = -100$.

*Hint:* The answer is $a = -2$ and $b \approx 8.909091$.

(C) Simulate 500 standard normally distributed numbers and make a histogram of the simulated data. Use the `summary` command to identify the median of the simulated numbers.

(D) Simulate 20 standard normally distributed numbers, and place the result in `X`. Compute the average of `X`, and generate a new variable called `tildeX` whose $i$'th element fulfil

$$\tilde{X}_i = X_i - \bar{X}_n$$

where $X_i$ is the $i$'th element of `X`, where $\bar{X}_n$ is the average of the elements of `X`.

---

*Date*: August 29, 2022.

Verify that the average of `tildeX` equals 0 (up to numerical precision).

*Hint 1:* Recall the command `result <- X + 1` in `intro.R`. It performs an operation (adding one) on each element of `X`, and places the result in the variable `result`.

*Hint 2:* Recall the so-called scientific notation, so that e.g. $4.200000e-01 = 4.200000 \times 10^{-1} = 4.200000 \times \frac{1}{10^1} = 4.200000 \times \frac{1}{10} = 0.42$. If you have never seen the scientific notation before, read about it on Wikipedia.

(E) Simulate 20 standard normally distributed numbers, and place the result in `X`. Compute the average and standard deviation of `X`, and generate a new variable called `tildeX` (*overwriting* the variable from (E) that was also called `tildeX`) whose $i$'th element fulfil

$$\tilde{X}_i = \frac{X_i - \bar{X}_n}{s_X}$$

where $X_i$ is the $i$'th element of `X`, where $\bar{X}_n$ is the average of the elements of `X` and where $s_X$ is the standard deviation of `X`.

Verify that the average of `tildeX` equals 0 and that the standard deviation of `tildeX` equals 1 (up to numerical precision).

(F) Repeat Task E a few times, verifying that the conclusion is the same each time. What would it take to convince you that this conclusion is true *always, and for every sequence of numbers* $X_1, X_2, \ldots, X_n$? Could you even in principle confirm this by repeating Task E many, may times? Say, after $10^{10}$ times?

(G) Let $x$ be a number and $n$ a non-negative integer. We then have that

$$\sum_{i=0}^{n} x^i = \frac{1 - x^{n+1}}{1 - x}$$

Figure out what the following R-code does, then use it to verify the above formula for a selection of $x$ and $n$ values.

```
n <- 15
x <- 0.5
series <- x^(0:n)
sum(series)
(1-x^(n+1))/(1-x)
```

Note that the above R-code is already included in `1.R`.

*Hint 1:* Look at `series` (by just writing `series` in the R-console after executing the above commands) and see if you can figure out what the first element is, the second element is, and so on, related to the formula $x^i$.

*Hint 2:* Recall the so-called scientific notation, so that e.g. $1.000000e+00 = 1.000000 \times 10^0 = 1$ and $5.000000e-01 = 5.000000 \times 10^{-1} = 0.5$, and so on.

## 2. An elementary problem in Big Data analytics

*Note:* This part of the problem set is just mathematics, and is solved with a pen and paper.

Statistics is a very broad discipline, which many see as encompassing so-called machine learning and big data techniques that underlie the generation and behaviour of websites such as Facebook, Google, Amazon and similar pages.

When doing statistics with very big data-sets, we have to be a bit inventive, as we may end up with computations that otherwise would not be possible to do on standard computers within a reasonable time. This is especially the case for websites. If computations needed to generate the user-specific website takes, say,

15 seconds, this is clearly unacceptable! Some computations underlying popular websites would take several days if done on a single computer.

One important technique in big data is to split up problems into manageable *chunks* that can be dealt with using standard computers, and then somehow combine the results from each computer into a finished computation on the entire dataset. The current philosophy in big data applications is to not use expensive super-computers, but instead form a gigantic network of cheap computers that together have vast computing power. However, each cheap computer is relatively slow, and has limited storage capacity, and this pose practical challenges for programmers.

Some data-sets are so large that this may require many, many computers. According to

`http://www.datacenterknowledge.com/data-center-faqs/facebook-data-center-faq`,

Facebook's website was running on around 60,000 computers in 2010, and the number is much higher today. Google's data-centres are much more massive, see

`https://en.wikipedia.org/wiki/Google_Data_Centers`

We will here look at the following elementary problems: If we distribute a big dataset with just a single variable onto several computers, how can we get them to compute a grand average? That is, each separate computer only use a fraction of the total data, and they are only able to do computations on this data. Then they send some type of summary to us, which we wish to combine into the average for the entire data-set.

(A) Recall the following problem from the preliminary course (problem 1, video 4, part I): Suppose we study the height of $n = 200$ people. The average of the first 100 people is 1.80. The average of the remaining 100 people is 1.78. Find the average height of all $n = 200$ people. Solve this problem again, without consulting the course notes or the video (at least in the start, unless you are stuck for a long time).

(B) Suppose we study the height of $n = 200$ people $X_1, X_2, \ldots, X_n$. The average of the first 100 people is $y$. The average of the remaining 100 people is $z$. Show that the average height of all $n = 200$ people is the average of $y$ and $z$, i.e., show that

$$\bar{X}_n = \frac{y+z}{2}.$$

(C) Suppose we study the height of $n = 100$ people $X_1, X_2, \ldots, X_n$. The average of the first 30 people is $y$. The average of the remaining 70 people is $z$. Show that the average height of all $n = 100$ people is

$$\bar{X}_n = \frac{30}{100}y + \frac{70}{100}z.$$

(D) Suppose we study the height of $n$ people $X_1, X_2, \ldots, X_n$. Let an integer $a$ be such that $1 \leq a \leq n$. The average of the first $a$ people is $y$. The average of the remaining $n - a$ people is $z$. Show that the average height of all $n$ people is

$$\bar{X}_n = \frac{a}{n}y + \frac{n-a}{n}z.$$

(E) Suppose we have observations $X_1, X_2, \ldots, X_n$. Let $u, v$ be integers fulfilling $1 \leq u \leq v \leq n$. Let us define[1]

$$\bar{X}_{u:v} = \frac{1}{v-u+1}\sum_{i=u}^{v} X_i. \tag{1}$$

[1]That is, let us assign meaning to the symbol "$\bar{X}_{u:v}$" by what is given on the right hand side of eq. (1).

Recalling that $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, we see that $\bar{X}_n = \bar{X}_{1:n}$, and we also see that $\bar{X}_{u:u} = X_u$ (to check your understanding of the above introduced notation, please confirm that this is so).

Now suppose we have integers $a, b$ fulfilling $1 \leq a < b < n$. Show that

$$\bar{X}_n = \frac{a}{n}\bar{X}_{1:a} + \frac{b-a}{n}\bar{X}_{(a+1):b} + \frac{n-b}{n}\bar{X}_{(b+1):n}.$$

Supposing that $\bar{X}_{1:a}, \bar{X}_{(a+1):b}, \bar{X}_{(b+1):n}$ were to be computed separately on three equally fast computers, how would you choose $a$ and $b$?

(F) What is known as a *weighted average* of numbers $Z_1, Z_2, \ldots, Z_n$ is given by

$$\sum_{i=1}^n w_i Z_i$$

where the weights $w_1, w_2, \ldots, w_n$ are such that $w_i \geq 0$ for each $1 \leq i \leq n$ and $\sum_{i=1}^n w_i = 1$.

Show that the weighted average is the classical average when the weights are constant, meaning the weights do not change with $i$.

(G) Show that $\bar{X}_n$ is a weighted average of $\bar{X}_{1:a}, \bar{X}_{(a+1):b}, \bar{X}_{(b+1):n}$. You do this by identifying weights, and showing that they conform to the requirements given when we defined a weighted average.

(H) *(Optional)* Provide a general formula for the grand average of $X_1, \ldots, X_n$ if we split the computation up into $M$ parts.

Department of Economics, BI Norwegian School of Management, Nydalsveien 37, Oslo, Norway 0484, Norway

*Email address*: Steffen.Gronneberg@bi.no