Exam 1 SCR (The midterm)

R-team

October 30, 2019

Exam Instructions

The current exam consists of two tasks for a total of 100 points. The third task is a bonus consisting of an extra 20 points. Behind each (sub)task you will see the number of points that can be earned. The grading model can always be modified in favor of all students.

Except for the package `readxl`, only use functions from the libraries that automatically load with the default settings of `RStudio`.

Furthermore, load the variables that are stored in `0_data/Resit_answer_variables.RData`. You can use these variables to check your answers, or use them to be sure you are working with the correct variables from the answers of the (previous) subtasks.

Your style of coding can affect your final exam grade. A correct answer is preferred above beautiful code, however, it may cost you points when very complicated code is provided as an answer at the place where basic `R` functions would suffice. Adhere to a consistent and neat programming style, e.g.,
+ https://google.github.io/styleguide/Rguide.xml,
+ http://style.tidyverse.org.

Change the name of the file `Lastname_ULCN.Rmd` accordingly (give it your real last name and ULCN number), and write down your answers in this file. Make sure you write your `R` code in `R` code chunks. When you wish or need to write text to answer one of the questions, don't use `R` comments, but just write your text outside the `R` code chunk.

Upload your changed `Lastname_ULCN.Rmd` file to Blackboard, and do so **before** 13.00 hours.

- Go to Statistical Computing with R –> Exams & Assignments –> Midterm Exam October 30, 2019.
- Every minute later than 13:00 hours will cost you 10 out of 100 points: when you submit your `.Rmd` file at 13:05 hours spot on, it means you already lost 50 points. Files submitted after 13:05 hours will not be graded.
- To be sure, you can also e-mail your .Rmd file to `rteam@stat.leidenuniv.nl`.
  Last, you are allowed to consult the internet and all files on your computer, as well as your own physically prepared written notes. However, keep in mind that any form of communication that is being sent to others is not allowed, and is considered **FRAUD**.

Success!

the R-team.

1. Last Year's Grades

Each worksheet in `filename.xlsx` is a data set representing the obtained grades of last year's students on the first exam (`Exam1`), the assignment (`Assignment`), and the second exam (`Exam2`). The data does not take into account the resit assignment, and the resit exam. For privacy reasons, we show fake names and fake ULCN numbers. Moreover, we added some noise to the data.

The first fours columns of each data set (= work sheet of the excel file) have the following names in the following order:`First Name` (column 1), `Last Name` (column 2), `ULCN` (column 3), and `Final Grade` (column 4).

1.1 Reading and Exploring the Data (20)

1.1a

To read all the data sets in one-go, we use the function `readxl::read_excel()` and a `for` loop, and the function `assign()`. The data sets are stored eventually stored in the object variables `Exam1`, `Assignment`, and `Exam2`.

```
library(readxl) # line 1
sheets <- c("Exam1", "Assignment", "Exam2") # line 2
for (sheet in sheets) { # line 3
  file_name <- "0_data/190108_SCR_Grades.xlsx" # line 4
  assign(sheet, read_excel(file_name, sheet = sheet)[, 1:4]) # line 5
}; rm(sheet)
```

Also run this code on your own console. Could you explain in at most two sentences per line what is happening at each line of code?

1.1b The same Column names?

Verify with `R` code that the column names of the data sets `Exam1`, `Assignment`, and `Exam2` are the same. Your answer code should result in a logical vector of length 1 with the value `TRUE`.

1.1c How many students?

The number of observations (students) are 47 for the first exam, 41 for the assignment, and 40 for the second exam.

Use "ULCN" columns of the three data sets, `Exam1`, `Assignment`, and `Exam2`, and show with code

i. that there are 49 (unique) students in these data sets?

ii. that there are 37 students that took part in both exams AND the assignment.

1.1d

Use a `for` loop to show the structure of each of the three data sets, make sure that the output for each data set is separated with a new line `"\n"`.

Hint: Check out what happens in your console when you run get("Exam1")

1.1e Checking out Percentages

A student has passed the course for sure when 55 or more points were obtained out of each exam and assignment. Around 74.5% of the students have obtained 55 or higher on the first exam, and around 75.6% obtained 55 or higher on the Assignment, and around 87.5% obtained 55 or higher on the second exam.

Could you reproduce these percentages in R? (There is no need for any kind of loop here, you can just write three lines of code)

1.2 Merging data sets (10)

Without merging the data sets into one, it is less obvious to check how many students eventually scored 55 (or higher) on the exams and the assignment. Therefore, we want you to create one data set containing all the grades together. Since the `merge()` function is only made to merge two data frames, we need merge the three data frames into two steps.

1.2a

To ease the merging proces, first change the name of each `Final Grade` column of the data sets.

Do the following in R: set the `Final Grade` column of the `Exam1` data set to `"Exam1"`, and the `Final Grade` column of the `Assignment` data to `"Assignment"`, and the `Final Grade` column of the `Exam2` data to `"Exam2"`.

1.2b

Merge the `Exam1` data and `Assigment` data with each other on the first name, last name, and ULCN. Ensure that all information of the merged data sets remains present in the new dat set.

Hint: if you are not familiar yet with the function `merge()`, take a quick look at its helpfile and its last examples.

1.2c

Perform the merge again, but now merge your new data set with the remaining `Exam2` data set. Your final data set should be equal to the `the_grades` data set from the model answer variables. Check whether this is true by using the `all.equal()` function in R.

1.3 Reproduce a Visualization (10)

If you did not succeed in 1.2, then use the data set `the_grades` from the model answer variables as your data set.

Can you reproduce Figure 1 which shows the visualization of the obtained grades by the students? The dashed red lines are plotted at a grade equal to 50 (indicating neccessary, but not sufficient requirements to pass the course).

Hint: Don't bother how things seem to scale out in your own produced plot

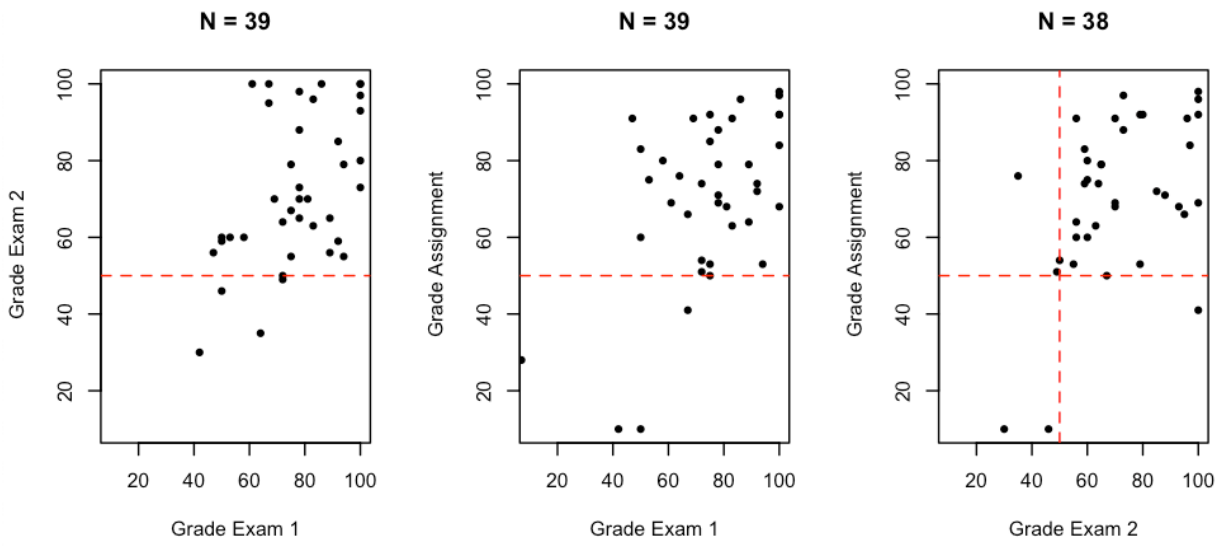

Figure 1: Scatterplots of the grades for Exam1, Exam2 and the Assignment. The red dashed lines indicate the necessary (but not sufficient) minimum requirements to pass the course.

1.4 Pass / resit and the final course grade (20)

If you did not succeed in 1.2, then use the data set `the_grades` from the model answer variables as your data set.

1.4a

In the first question we've seen that there are 37 students who took part in both exams and the assignment. Could you verify this with code by using the newly merged data and by counting the number of observations that have an `NA` value for either one of the three grades?

1.4b

To pass the course without any resit, a final SCR course grade of 55 (out of 100) is required. Let

E_1 := Grade Exam 1
E_2 := Grade Exam 2
$E = (E_2 + \max(E_1, E_2))/2$
A := Grade Assignment
Y := Final SCR course grade (without resits)

Then,

$$Y = \begin{cases} (2/3)E + (1/3)A, & \text{if } (E \geq 50) \cap (E_2 \geq 50) \cap (A \geq 50) \\ \min\left((2/3)E + (1/3)A,\ 50\right), & \text{otherwise.} \end{cases}$$

If either E_1, E_2, or A has a missing value (`NA`), then replace it with a 10 (out of 100).

Write a function which tells us whether the students have passed the course in one-take, or should (at least) do one resit. The function takes as input the `the_grades` data set, and gives as output a `list` that contains

two entries: 1. a factor (pass/resit) of `nrow(the_grades)` elements, and 2. a numeric vector (final grade) of `nrow(the_grades)` elements. The labels (and levels) of the factor indicate whether a student has passed the course (`passed`) or needs to do at least one resit (`resit`).

Hint: don't bother about rounding the final course grade. Right now you don't need to understand why the grade is calculated this way, just implement the formula.

1.4c

Create a table of the `factor` pass/resit to see how many people passed the course in one go, and also create a barplot of the `factor` with the title `"SCR Course in-one-go"`.

If you did not succeed in the previous task(s), then use the `pass` column `final_grades` that is in the model answer variables.

1.4d

Use the function `aggregate()` or `tapply()` to show the average grade of the students that passed the course, and of those that would have to do the resits.

Hint: if you did not succeed in the previous tasks, you may use the `CourseGrade` column from the data set `final_grades`

2 Generate(d) Data

The data from the previous task consisted of fake names and fake ULCN numbers for reasons of privacy protection. In this second task we will ask you to generate UCLN numbers yourself, and to code your own function with which you can encrypt names.

2.1 About ULCN numbers (20)

Your ULCN number starts with an "s" and then a number of 7 digits. Let us describe the *population* of ULCN numbers from which we can generate a sample (like the ULCN numbers of the SCR course of last year). In this population 25% of the numbers are uniformly distributed over the interval `[1000001, 1699999]`, consisting of discrete values only. The ULCN numbers in this interval are also referred to as the *old* ULCN numbers. The remaining 75% of the numbers are uniformly distributed over the interval `[1700000, 2499999]`, also consisting of discrete values only. The numbers in this latter interval are referred to as the *new* ULCN numbers.

Knowing our "population" of ULCN numbers, we will ask you later to generate `49` unique ULCN numbers from this population.

2.1a

Create a vector object `ULCN_nrs` that consists of the numbers 1000001 to 2499999. Check whether the lenght of this vector is `1499999`. Then, use this `ULCN_nrs` vector to confirm with a logical expression that `699999` of the numbers are smaller than `1700000`, and `800000` are larger or equal to `1700000`.

2.1b

For the SCR-class we estimate that the ULCN numbers come from the same population as described above. Thus, around 25% of the students have an old ULCN number, and around 75% have a new ULCN number, the probability of drawing one specific old ULCN number is

```
prob_old <- 0.25 / 699999
```

and the probability of drawing specific new ULCN number is

```
prob_new <- 0.75 / 800000
```

Use the function `rep()` to create a vector object `probs_ULCN` that contains the probability of each ULCN number in `ULCN_nrs`. Also show that the sum of all probabilities in `probs_ULCN` is 1.

2.1c

Use `ULCN_nrs` and `probs_ULCN` to generate 49 unique ULCN numbers accordingly. Set a seed `20191030` (representing the date of this exam).

2.1d

Change your generated ULCN numbers into a **sorted** character vector where each number also has the "s" up-front (see `my_ULCNs` from the model answer variables).

2.2 Decrypting Names (20)

It is common practice to make sure that privacy related data is encrypted. For encryption and decryption of the data a 'key' is used. In this task you will have to encrypt the data with such a key. Let the key be of class `data.frame`, of which `key20191030` here below is an example.

```
key20191030 <- {
  set.seed(20191030)
  data.frame(
    decrypt = c(letters, LETTERS, 0:9, " ", NA, ".", "/", "-"),
    encrypt = sample(c(letters, LETTERS, 0:9, " ", NA, ".", "/", "-")),
    stringsAsFactors = FALSE
  )
}
```

You will have to use this particular `key20191030` to encrypt the names of the students we have used before.

2.2a Import the data in two columns

The names of the students can also be found in the `names.txt` file from the `0_data` folder. Import the content of this file into a `data.frame` of two columns (First name and Last name). Ensure that the columns of the `data.frame` are `character` vectors.

2.2b Using the `key` to encrypt the names

Write a function that encrypts any character vector (= first argument), based on a key `data.frame` (= second argument). As a default, set the second argument equal to the `key20191030` variable.

Show that your function can encrypt a character vector like `c("Tommy", "Francina")` correctly into `c("IkddN", "pKGunXuG")`.

Hint: first try to write the function just for one name only. When you've succeeded, then try to see on how you could apply this function in a loop on each element of the character vector (or list).

2.2c

Create a new `data.frame` with the encrypted names by running an implicit loop, `lapply()`, over the two columns of the original `data.frame` with the (decrypted) names. *Note that this needs a bit more coding then just a `for` loop.*

Your encrypted `data.frame` should contain the exact same names as `encr_students` from te model answer variables. If you did not succeed to load the orignal (decrypted names) into a data.frame, use the `student_names` data.frame from the model answer variables. Similarly if you did not manage to write your own encrypt function, then use the following function instead:

```
Surrogate <- function(names_decr, key) {
  load("0_data/model_answer_vars.RData")
  if(length(names_decr) != nrow(encr_students)) {
    stop("wrong input")
  }
  if(names_decr[1] == "Tommy") {
    out <- encr_students[, 1]
  } else if(names_decr[1] == "Tejera") {
    out <- encr_students[, 2]
  } else {
    stop("wrong input")
  }
  return(out)
}
```

2.2d

Write your encrypted names (or the `encr_students` object) into a text file and store them in the `0_data` folder. Your encrypted names should be the same as those in the file `encrypted_names.txt` from the `0_data` folder.

3 Bonus: A Visual illusion (20)

Take a look at Figure 2, you probably see black dots appearing and dissappearing. This Figure is a typical visual illusion that you could create with `R` as well.

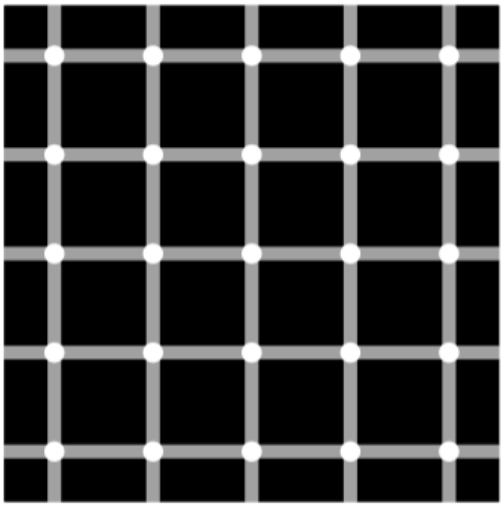

Figure 2: Visual illusion with Black Dots

Recreate Figure 2 with your own code. Use the color `"#FFFFFF90"` for plotting the see-through white lines (that appear grey) on the black rectangular background. Keep the aspect ratio of the vertical axis and that of the horizontal axis equal to 1 (`?plot`).

All 20 points can be obtained for this task if you do **NOT** use any implicit or explicit loops.

Hint:Check out the function `rect()`.