Practical Session on R

Louis-Claude Canon and Emmanuel Jeannot

January 22, 2024

Abstract

The objective of this practical session is to present some basic features provided by R. Additionally, methodological principles are given in order to acquire autonomy in using R functions. The participant is assumed to be familiar with programming, file editing and Unix shell.

1 Getting used to R interactive environment

Launch RStudio:

```
$ rstudio
```

The prompt looks like this in the bottom-left part:

```
>
```

You can do a simple computation

```
>3^2
```

You can now enter your first assignment:

```
> X <- c(5, 1)
```

Type the name of the variable X to show its content:

```
> X
[1] 5 1
```

R is *vectorial*ÃÂ language. This means that operations are applied on each elements of the vector. Try :

```
> X <- c(1, 2, 3, 4, 5)
> X^2
> 2^X
> X^2 > 6
```

Type the name of the variable X to show its content:

```
> X
```

Look the structure of the documention of one function:

```
> ?print
```

Test the following functions without argument: *ls*, *rm*.

You can then see the output of the following functions on any variable (try with the variable X): *mode*, *class* and *length*.

At any step, you can access the previous instructions you have entered by using the *up* key.

Exercise :

Q1: compute your BMI (Body Mass Index) that is your body mass (in Kg) divided by the square of the body height (in m)

Q2: put your BMI in a variable and check if it is larger than 40

Q3: compute `x<-runif(100,0,10)`. Use `mean`, `max`, `min`, `var`, `sd`, `median` or `length` to compute some statistic on `x`.

Q4: To compute a function, the simplest way is one to generate a vector x of all the abscissa values (by using `seq`). Then compute the vector y as a function of x. Compute $sin(x)/x$ for $x \in [-10, 10]$ with a step of 0.1

Q5: You can find the indexes of the values of a vector that have certain properties. For instance `which(x>=0)` gives you all the indexes of the positive values in x. `length` gives you the length of a vector. Use these two functions to compute the ratio of positive values of question Q4 above.

Q6: you can change the values of a vector with a vector of indexes.

```
> X <- seq(0,10)
> I <- c(1,2,4)
> X[I]<- -X[I]
> X
```

As `which` return a vector `X[which...]` is valid. Use this to transform the y vector computed in Q4 to truncate the function to 0.5 (all the values greater than 0.5 become 0.5).

Q7: A friend propose the following game. You draw a 6-faces dice. If the result is 5 or 6 you win 3 €, if the result is four you win 1€and if it is less than 3, you loose 2.5 €. Before accepting to play, try to simulate this game to see your chance to get richer. Conclusion? Is this coherent with the theory?

Use: `runif`, `floor` (because `runif` is for the continuous uniform distribution), `which` and `mean`.

2 Data types

In this section, you will perform simple operations on data types. Save relevant R instructions in a script (top-left part of the window) for future usage during the session.

Q0: use the `c` function for creating two vectors (`v1, v2`) of the same number of elements (e.g. 3 ints). Compare `M1<-rbind(v1, v2)` with `M2<-cbind(v1,v2)` and multiply `M1` with `M2` using `%*%`.

Q1: First, create a vector of 100 successive intergers from 1 to 100.

Q2: First, create a vector of 100 numbers randomly drawn from a uniform distribution. We refer you to the manual pages of the following functions: *runif*.

Q3: Count how much of them are greater than 0.5 (*which* and *length* functions).

You have now two vectors of distinct types (check this with the *class* function on both vectors).

Q4: Build a bi-dimensional structure with both vectors (*data.frame*). Test this method and use *attributes* on the produced variable.

Q5: Give an appropriate name to each column either at the construction of the structure (when calling *data.frame* for example) or by modifying the attributes (*names* or *colnames*).

You can now access one column with the following syntax: *data$colname*.

Q6: Now, we discard the rows in which the second value (the numeric one) is above 0.5. Note that T[1:2,] returns the two first rows.

Q7: Let call noise the second values. Make a vector of all the noise lower than 0.5

3 Functions

One of the coolest thing of `R` is how it handles parameters in function. You can give a default value to a parameters that can be overwritten only if specified. For instance look at :

```
mytan<-function(angle, deg = 0, method = 0){
  if(deg != 0){
    x <- pi*angle/180
  } else{
    x <- angle
  }
  if(method == 0){
    tan(x)
  }else{
    (x-x^3/6+x^5/120)/(1-x^2/2+x^4/24)
  }
}
mytan(pi/3,method=0) ; mytan(pi/3,method=1) ; mytan(60,deg=1)
```

4 Inputs

We would like to save in a file the first data frame that was created.

Q1: Use the function *write.table* (the first argument in the data frame and the second is the name of the output file) to perform this operation and open the generated file to note the default behavior.

Q2: Use *read.table* on the same file with the default parameters (precise only the name of the file) and compare the content and the attributes with the initial data frame.

5 Plot

In R, it is possible to specify the device driver to use for each graphical output. By default, it is the one of your graphical environment. You can generate postscript, pdf, xfig, png, ... For instance, you may use the pdf device driver by initializing the device with *pdf("test.pdf")* before calling a plotting primitive and by closing the device afterward with *dev.off()*.

5.1 *plot* function

The first notable graphical function, *plot*, allows several types of plot to be drawn: points, lines, both, steps, ...

Q1: We would like to plot the value used in the previous sections with the first column on the x-axis and the second column on the y-axis. Experiment various types of plot.

Q2: Most of the parameters are described on the manual page of *par*. For instance, change the points aspect with parameters *cex*, *col* and *pch*. Lastly, experiment logarithmic scale (*log*).

Q3: Specify the title, x-axis and y-axis labels (with parameters *main*, *xlab* and *ylab*). Add a legend (*legend* function).

Q4: plot $sin(x)/x$ in $[-2\pi, 2\pi]$ with a resolution of 0.1. Hint design the vector `x` with `seq` computey `<- sin(x)/x` and plot `y` vs `x`.

5.2 Plotting distributions

In this section, we suppose that we have a collections of values and we want to plot them without excessive aggregation. The data is the second column of the data frame previously used.

Q1: Initialize a plot by calling the function *hist*. Specify to use density scale (*freq = FALSE*). The breaks can be adjusted with the parameter *breaks* (10 to 20 breaks should be appropriate).

Low-level plotting does not initialize a plot but draw additional elements on existing plot (then, they use the existing axises).

Enter *lines(density(VALUES, bw = 0.5))*, where *VALUES* is your collection of values. This draws an estimate of the density of the distribution from which comes the values. Parameter *bw* denotes the sensitivity of the method to each single sample (try larger and smaller values).

Q2: R allows to plot empirical cumulative distribution functions. You have to call *plot(ecdf(VALUES))* (or the *plot.ecdf* function).

Copy and paste the example given on the manual page of function *ecdf*. It shows what can be obtained through customization.

Q3: You can also superpose histogram with distribution. Use `rbinom` to draw 1000 samples of binomial variable with 50 trials and 0.5 of probability. Then draw its histogram (density mode as Q1 above) with 30 breaks and superpose the normal distribution with the `curve(dnorm(x,...), ...)` function (with `add = TRUE`). In this case the mean is 25 and the standard deviation is $\sqrt{0.5 \times 0.5 \times 50} \approx 3.535534$.

5.3 Tests

Questions Q1-Q5 are taken from `Statistical inference for data science` by B. Caffo.
Q1 Load the `father.son` dataset. `library(UsingR); data(father.son)`. `father.son$fheight`

contains the father height and `father.son$sheight` contains the son height. If we want to test the confidence interval for the difference of height between father and son, what kind of test do we have to do? Test the normality of this difference. Perform the test in R. Interpret the result.

Q2 Load the mtcars dataset with `library(datasets); data(mtcars)`. Extract a vector m4 (resp m6) that contains the MPG (miles per galon) for 4 (resp. 6) cylinder cars. Can you conclude with 99% confidence that 4 cylinders are ore economical than 6 cylinders. What kind of tests have been performed. Assume a constant variance.

Q3 Assume that the data set mtcars is a random sample. Compute the mean MPG, $\bar{x}$, of this sample. You want to test whether the true MPG is μ_0 or smaller using a one sided 5% level test. ($H_0 : \mu = \mu_0$ versus $H_a : \mu < \mu_0$). Using that data set and a Z test: Based on the mean MPG of the sample $\bar{x}$, and by using a Z test: what is the smallest value of μ_0 that you would reject for (to two decimal places)?

Q4 Consider again the mtcars dataset. Use a two group t-test to test the hypothesis that the 4 and 6 cyl cars have the same mpg. Use a two sided test with unequal variances. Do you reject?

Q5 You believe the coin that you're flipping is biased towards heads. You get 55 heads out of 100 flips. Do you reject at the 5% level that the coin is fair?

Q6 Consider a weighing balance that displays weights in grams and a precision of miligrams. However each time it takes a measure it makes a gaussian error (following a normal law) with a standard deviation of 1. Simulate in R a set of 100 measures (using `round`) and display the 95% confidence interval. Assume infinite precision, how many measures do you have to do if you want to a the confidence interval to be smaller than 0.1. Each time you want to divide the uncertainty by 10 how many more measures do you have to do?

Q7 In 2018, 302 982 girls were passing the bacalaureat as well as 297 960 boys. The success rate was respectively 91.04% and 86.02%. Use `Prop.test` to compute the 95% confidence interval of the difference of the proportion. What can you conclude?

Q8 We measure the weight of apples in grams. We in a sample of 7 apples we find : 154, 165, 151, 171, 148, 155, 162. Based on this sample what is the probability that an apple weights more than 175g.

Q9 We sample the weight (in grams) of 15 strawberries of two fields. On field one we find:
$48.73, 43.44, 46.71, 51.62, 47.24, 54.64, 47.00, 48.40, 45.86, 47.70, 46.14, 47.68, 44.73, 51.69, 50.54$.
On field two we find:
$44.89, 34.31, 42.74, 53.36, 41.98, 41.64, 47.24, 37.86, 45.89, 40.88, 40.85, 38.60, 44.38, 44.52, 38.26$.
Compare the two hypothesis: "H_0 the two fields produce the strawberries of same weights" with H_1 "the two fields produce the strawberries of different weights" with a confidence of 98%.

Q10 Consider the gain in weight of 19 female rats between 28 and 84 days after birth. 12 were fed on a high protein diet and 7 on a low protein diet.

High protein	Low protein
134	70
146	118
104	101
119	85
124	107
161	132
107	94
83	
113	
129	
97	
123	

Can we say than one diet is different from the other, with 95% confidence?

Q11 : χ^2 test You and a friend are munching on a bag of Harvest Blend M&M's, when your friend says, "There seems to be more yellow and brown candies than red and maroon candies. In fact, I claim there are 30% yellow, 30% brown, and only 20% red and 20% maroon." Together you count the remaining M&M's in the bag with the results below. Use the critical value method with significance level 0.05 to test your friend's claim.

Color	Yellow	Brown	Red	Maroon	Total
Number	58	61	55	46	220

Q12 : χ^2 test

A sample of coin flips is collected from three different coins. The results are below. Use one hypothesis test to test the claim that all three coins have the same probability of landing heads. Use the critical value method with significance level 0.10.

	Coin A	Coin B	Coin C
Heads	88	93	110
Tails	112	107	90

6 Correction

7 Getting started

Q4: `x<-seq(-10,10,0.1) ;y<-sin(x)/x`
Q5: `length(which(y>0))/length(y)`
Q6: `y[which(y>0.5)]<-0.5`
Q7: The theoretical mean is -0.08333333. To do simulate. Generate a vector `x` with 6000 integers uniformly picked in [1,6]. You can check with `length(x[which(x==1)])` ... `length(x[which(x==6)])`. With `which` change each values by its gain (be careful to do this in the right order) and compute the average of x (hence the average gain).

7.1 Data types

Q0: `v1 <- c(1,2,3) ; v2 <- c(4,5,6) ; M1 <- rbind(v1,v2); M2 <- cbind(v1,v2); M1; M2 ; M1%*%M2`
Q1: `X <- 1:100`
Q2: `Y <- runif(100)`
Q3: `length(which(Y > 0.5))`
Q4: `V <- data.frame(X, Y)`
Q5: `V <- data.frame(int=X, noise=Y)` or `V <- data.frame(X, Y) ; names(V)[1] <-"int" ; names(V)[2] <-"noise"`
Q6: `W <- V[V[,2] < 0.5,]` or `W <- V[-which(df$noise>0.5),]`
Q7: `X<-V$noise[V$noise>0.5]`

7.2 Inputs

Q1: `write.table(V, "test.txt")`
Q2: `read.table("test.txt")`

7.3 Plot

7.3.1 *plot* function

Q1:
```
plot(V, type = "b")
```
Q2:
```
plot(V, type = "b", cex = 0.5,
     col = "red", pch = 12)
```
Q3:
```
plot(V[order(V[,1], V[,2]),], type = "b", cex = 0.5,
     col = "red", pch = 12, main = "Main title",
     xlab = "integer", ylab = "integer with noise")
 legend("bottomright", "random value", pch = 12, col = "red", pt.cex = 0.5)
```
Q3: `x <- seq(-2*pi,2*pi,0.1) ; y <- sin(x)/x ; plot(x,y,type='l')`

7.3.2 Plotting distributions

Q1: `hist(Y, freq = FALSE, breaks = 20)`
Q2: `plot(ecdf(Y), do.points = FALSE, xlim = range(Y),`

```
       xaxt = "n", yaxt = "n", ann = FALSE) Q3: v<-rbinom(1000,50,0.5); hist(v,breaks
= 30,freq = FALSE) ;curve(dnorm(x,mean=25,sd=sqrt(0.5*0.5*50)), col="green", lwd=2,
add= TRUE)
```

7.3.3 Tests

Q1: The p-value of the Shapiro Wilk test is 0.5087 and the interval is $[0.8310296, 1.1629160]$
Q2: 99% confidence interval is $[2.500601, 11.340958]$
Q3: You would reject if TS$<z_{0.05}$. $\bar{x} = 20.09$. The alternative hypothesis is that $\bar{x}$ is lower than μ_0. Hence, $\mu_0 > 20.09$. Here, $s = 6.026948$ and $z_{0.05} = -1.644854$. Hence, the answer is 21.84
Q4: The interval is $[3.751376, 10.090182]$ and the p-value is 0.0004048 so we reject.
Q5: The interval is $[0.462983, 1.000000]$ and the p-value is 0.1841. So, we do not reject the possibility that the coin is fair.
Q6 You have to solve $2 \times z_{0.975} \times /\sqrt{n} < 0.1$. Hence, $n > 1536$. The precision is inversely proportional to $\sqrt{n}$, hence I have to do 100 times more experiments to divide the precision by 10.
Q7: The 95% confidence interval is $[0.04858899, 0.05181101]$. This this means that with 95% of confidence the success rate of girls is between, 4.86% and 5.18% higher than the one of the boys.
Q8: The probability is 0.04243692.
Q9: The t.test p.value is 0.000503 so we can reject H_0.
Q10: assuming unequal variance leads to a 95% confidence interval of the difference of the mean of: $[2.193679, 40.193679]$. Assuming equal variances leads to: $[2.469073, 40.469073]$. In both cases the interval contains 0, hence we cannot conclude that one diet is different than the other.
Q11: $T = 4.189394$ is lower than $Q = 7.814728$ so we fail to reject.
Q12: $H_0 : p_A = p_B = p_C$ (all three coins have the same probability of landing heads.) $T = 5.3248$ is greater $Q = 4.60517$ so we reject H_0.