Statistics and R

Ana Carolina Gontijo Graça

December 2019

1 Introduction

This is a HarvardX course based R utils summary, with some of the important functions and properties of R that I found during this course (as well as basic functions).

2 R Utils

2.1 Basic Functions

2.1.1 Creating a vector [1,2,3,4,5] :

```
version1 <- c(1,2,3,4,5)
version2 <- 1:5
version3 <- seq(1,5,by=1)
version4 <- seq(1,5,lenght.out=5)
```

- **version1**: Uses the basic vector creator
- **version2**: Uses the range vector creator (only for sequential integers)
- **version3**: Uses the *seq* function, defining the initial and final elements and the increments between them (by).
- **version4**: Uses the *seq* function, defining the initial and final elements and the number of elements in the resulting vector.

2.1.2 Looping in elements of a vector printing them:

```
for (element in vector) {
  print(element)
  paste(element)
}
```

The for loop simply iterates through a list/vector/dataframe or any other structure with elements. The functions *print* and *paste* both print strings on the screen, but I mostly used *paste* in the exercises since it was easier to mix strings and variables in it.

2.1.3 Installing and using packages, using datasets from its package and analysing classes:

```
install.packages("datasets")
library(datasets)
cars
class(cars)
```

First and second line are preparing for the use of the **datasets** package. When you are using it, some datasets are available for use like they were variables (see [1]). One of those is **cars**. By using the function *class* in any object it will return the class of that object.

[1] https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html

2.1.4 Functions length, nrow and colnames:

```
length(cars)
nrow(cars)
colnames(cars)
```

The *length* function returns the number of columns (or the size of the last dimension of the dataframe), while the *nrow* function returns the number of rows. The *colnames* function returns a list of names of the columns of the dataframe.

2.1.5 Accessing with *[]* and using function *mean*:

```
cars[1,2]
cars[,2]
mean(cars[,2])
```

It's possible to access elements of objects using *[]*, when it's used with integers in all spaces, it simply chooses one element by its index in each dimension. When some indexes are missing, this chooses all elements in that dimension. The function *mean* simply takes an 1 dimension object and takes the mean of its elements.

2.1.6 Function *which* and operator *$*:

```
which(cars$dist == 85)
```

The operator *$* chooses a column of the data frame by its name. The function *which* recieves a logic expression as a parameter, returning the indexes where this expression is true.

2.2 Dplyr Package Functions

2.3 Qqplot Package Functions

3 Statistics

3.1 Central Limit Theorem

The Central limit theorem says that in a sample $\{X_1, X_2, \cdots, X_n\}$ of size n taken from a set with mean μ and variation σ, the random variable given by: $\frac{\sum_{i=1}^{N} X_i}{n} = \overline{X}$ is approximated by a normal distribution with mean μ and variation $\frac{\sigma^2}{n}$.

4 Reference

- 2019, *Statistics and R*, Online Course on EDX, Harvard