DATA SCIENCE

Introduction to R

Practical 1

M.José Ramírez-Quintana
(with modifications by José Hernández-Orallo)
ETSINF
Universitat Politècnica de València

September 16, 2015

---

1. Generate the numbers 1, 2,. . ., 12, and store the result in the vector x.

2. Generate four repetitions of the sequence of numbers (6, 2, 4).

3. Generate the sequence consisting of six 9s, then five 2s, and finally four 5s. Store the numbers in a 5 by 3 matrix (populating it columnwise).

4. Generate a vector consisting of 20 numbers generated randomly from a normal distribution. Use the value 100 as seed (in order to be able to replicate the experiments). Setting the seed is done as follows

   > set.seed(100)

   Then, calculate the following statistics about the generated vector: mean, median, variance and the standard deviation.

   Repeat the generation of the vector and the statistics with and without changing the seed and observe what happens.

5. From the resources folder at poliformat, download the file "data1.txt" that contains information about students.

   (a) Read the data into an R object named `students` (data is in a space-delimited text file and there is no header row).

   (b) Add the following titles for columns (see section 9):
   `height, shoesize, gender, population`

   (c) Check that R reads the file correctly.

   (d) Print the header names only.

   (e) Print the column `height`.

(f) What is the gender distribution (how many observations are in each groups) and the distribution of sampling sites (column `population`) ?

(g) Show the distributions in the above item at the same time by using a contingency table.

(h) Make two subsets of your dataset by splitting it according to gender. Use data frame operations first and then do the same using the function `subset`. Use the help to understand how `subset` works.

(i) Make two subsets containing individuals below and above the median height. Use data frame operations first and then do the same using the function `subset`.

(j) Change height from centimetres to metres for all rows in the data frame. Do this using in three different ways: with basic primitives, a loop using `for` and the function `apply`.

(k) Plot height against shoesize, using blue circles for males and magenta crosses for females. Add a legend.