Applied probability lab session 2

Prepare a report including your code and the answers and interpretation.

The lecture http://www.dyndoc.fr/teaching/ssd proposes

- how to get started with R
- how to perform descriptive statistics with R
- how to summarize and present variables or couples or variables

1 Starting with R

Exercise 1

1. Open a new script. Save your file as `Lab2.r`.
2. Create the vector (1, 2, 3, 4, 5) with the command:

   ```
   c(1,2,3,4,5)
   ```

   `c(.)` is the concatenation function which links different values into a new vector. Assign the previous vector to `X` by:

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

   Check the contents of `X`, by typing `X` then return.
3. Create the vector `Y` with values `(1, 4, 9, 16, 25)`.
4. Check that `X` and `Y` have the same length.
5. Plot the points defined by the two vectors `X` and `Y` by `plot(X,Y)`. Change the symbol: `pch=2`, then `pch=3`, etc. Change the type: `type="b"`, then `type="l"`. Change the color: `col="red"`, then `col="blue"`, etc. Add a title, add labels on both axes.
6. Add the curve $y = x^2$ by

   ```
   curve(x^2,add=TRUE)
   ```

Exercise 2

1. Create the vector `X` containing all integers from 0 to 7.
2. Multiply `X` by 5, divide it by 5, add 5 to it.
3. Compute the sum of `X`, its cumulative sums.
4. Compute the square root of `X`, its third power.

Exercise 3

1. Create the vector `X` containing (0, 1, 4, 9, 16). Extract from `X` the subvector with indices 3 and 5. Extract all values larger than 2. Extract all values larger than 2 and smaller than 10.

2. Create the vector `Y` containing 5 ones (with function `rep`), the vector `Z` containing the sequence from 3 to 11 by step 2 (with function). Concatenate `X`, `Y`, `Z`. Bind them as rows. Bind them as columns, and assign the result to `XYZ`.

3. Compute row sums and column sums of `XYZ`.

4. Extract from `XYZ`:

   (a) row number 4,

   (b) column number 3,

   (c) rows with indices 3, 5, columns with indices 2, 3,

   (d) rows such that `X` is larger than 2.

   (e) columns named `"Y"` and `"Z"`.

2 Descriptive statistics

1. Upload titanic.csv, read data description. Assign column pclass to variable `P`, column survived to variable `S`, column gender to variable `G`, column age to variable `A`. Sort the four variables as discrete or continuous.

2. Display the first 6 rows. Display rows 250 to 257. Display the data for rows 28,34,78, and columns 2,4. Display the data for all first class passengers. Display the data for babies (under 1 year old).

3. Compute the absolute and relative frequencies of the three passenger classes. What was the proportion of first class passengers?

4. Compute the absolute and relative frequencies of survivors and non-survivors. What was the proportion of survivors?

5. Compute the absolute and relative frequencies of men and women. What was the proportion of women?

6. Display bar plots for all three variables.

7. Display bar plots for `P` per value of `S` and for `S` per value of `P`.

   (a) How many first class passengers survived?

   (b) What proportion of all passengers were first class passengers and survived?

   (c) What proportion among first class passengers survived?

   (d) What proportion of survivors were first class passengers?

   (e) Repeat for second and third class passengers.

   (f) What are the conditional frequencies of the three passenger classes given they survived? given they did not survive?

8. Display bar plots for `P` per value of `G` and for `G` per value of `P`.

(a) How many first class passengers were men?

(b) What proportion among first class passengers were men?

(c) What proportion of men were first class passengers?

(d) Same question for second and third class passengers.

(e) What are the conditional frequencies of the three passenger classes given they were men? given they were women?

9. Display summary, boxplot, histogram, ecdf, for variable A.

    (a) Compute the mean, variance, standard-deviation, median, quantiles at 1/3 and 2/3, inter-quartile range, for variable A.

    (b) How many passengers were older than 40? younger than 20? What proportion of passengers were older than 40? younger than 20? What age is such that 5% of passengers were younger? What age is such that 5% of passengers were older?

    (c) On the boxplot of A, superpose a red horizontal line marking the mean, a blue horizontal line marking the median, two green horizontal lines marking the first and third quartiles.

10. Display summaries of A by values of P. In which class were passengers older?

11. Display summaries of A by values of S. Were survivors older or younger than non survivors?

12. Display summaries of A by values of G. Were women older or younger than men?