Chapter 1: Data exploration

1 Session 1

Objectives: install the softwares, get started with R.

Steps of the session:

1. Start by installing the softwares R and Rstudio, following the links on `https://www.r-project.org` and `https://www.rstudio.com`.
2. Download the dataset and the instructions on `https://toltex.u-ga.fr/bio`.
3. Install the packages `survival` and `stoda`.
4. Individual work: read Chapter *Getting started with R.*
5. Individual work: Practice with the instructions given on the web site of the course.
6. Share and work: discuss what you understand.
7. Individual work: Practice with the three exercises of Session 1.
8. Share and work: discuss the scripts together.
9. Individual work at home: read Chapter *Descriptive statistics*, sections *Dataset and variables* and *Summarizing/presenting variables.*

Exercise 1.1.

1. Open a new script. Save your file as `Lab1.r`.
2. Create the vector $(1, 2, 3, 4, 5)$.

   Assign the previous vector to `X` .

   Check the contents of `X`.
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 1.2.

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 1.3.

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 (`rep(1,5)`), the vector `Z` containing the sequence from 3 to 11 by step 2 (`seq(3,11,by=2)`). 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 Session 2

Objectives: Load and understand a data file, name and sort variables; describe discrete variables, compute absolute, relative, conditional frequencies; describe continuous variables; understand the notions of mean, variance, standard deviation, quantiles, distribution function, boxplot, histogram, empirical cumulative distribution function (ecdf).

Steps of the session:

1. Individual work: Practice with the instructions given on the web site.
2. Share and work: discuss what you understand.
3. Individual work: Practice with exercises of Session 2 (at least exercise 1).
4. Share and work: discuss the scripts together.

Exercise 2.1.

1. Upload `bosson.csv`, read data description. Assign column `country` to variable `C`, column `gender` to variable `G`, column `aneurysm` to variable `A`, column `bmi` to variable `B`, column `risk` to variable `R`. Sort the five variables as discrete or continuous.
2. Display the first 6 rows. Display the data for rows 28,34,78, and columns 2,4. Display the data for Vietnamese patients. Display the body mass index of men.
3. Compute the absolute and relative frequencies of the two countries. What proportion of patients are Vietnamese? Compute the absolute and relative frequencies of the two genders. What proportion of patients are women? Compute the absolute and relative frequencies of the six risk levels. What proportion of patients have at least two risk factors? Display bar plots for the three variables.
4. Display bar plots for `C` per value of `G` and for `G` per value of `C`. How many men are Vietnamese? What proportion of all patients are Vietnamese men? What proportion among men are Vietnamese? What proportion among Vietnamese are men?
5. Display bar plots for `C` per value of `R` and for `R` per value of `C`. How many Vietnamese have 0 risk factor? What is the proportion of Vietnamese with 0 risk factor among all patients? What proportion of Vietnamese have 0 risk factor? What proportion of patients having 0 risk factor are Vietnamese?
6. Display bar plots for `G` per value of `R` and for `R` per value of `G`. How many men have 0 risk factor? What is the proportion of men with 0 risk factor among all patients? What proportion of men have 0 risk factor? What proportion of patients having 0 risk factor are men?

7. Display summary, boxplot, histogram, ecdf, for variable `A`.
8. Compute the mean, variance, standard-deviation, median, quantiles at 1/3 and 2/3, inter-quartile range.
9. Assign to `Az` the z-scores of variable `A` (standardize). Repeat questions 2 and 3 for variable `Az`. What are the differences with `A`?
10. How many patients had aneurysm size below 40? above 60? What proportion of patients had aneurysm size below 40? above 60? What size of aneurysm is such that 10% of patients are below? What size of aneurysm is such that 10% of patients are above?
11. 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.
12. On the histogram of `A`, superpose a red vertical line marking the mean, a blue vertical line marking the median, two green vertical lines marking the first and third quartiles.
13. On the ecdf plot of `A`, superpose a red vertical line marking the mean, a blue vertical line marking the median, two green vertical lines marking the first and third quartiles. Add a blue horizontal line at 0.5, two green horizontal lines at 0.25 and 0.75.
14. Repeat questions 2. to 7. for variable `B`; for question 4, replace 40 and 60 with 22 and 30.

Exercise 2.2.

1. Upload `ferretti.csv`, read data description. Assign column `height` to variable `H`, column `diameter` to variable `DA`, column `density` to variable `DE`, column `invasive` to variable `I`. Sort the four variables as discrete or continuous.
2. Display the first 6 rows. Display the data for rows 10 to 15, and columns 2,4. Display the data for invasive tumors. Display the height of tumors having positive density.
3. Compute the absolute and relative frequencies of the three classes of density. What is the proportion of tumors with positive density? Compute the absolute and relative frequencies of invasive and non-invasive. What is the proportion of invasive tumors? Display bar plots for the two variables.
4. Display bar plots for `DE` per value of `I` and for `I` per value of `DE`. How many invasive tumors have positive density? What proportion of all tumors are invasive with positive density? What proportion among tumors with positive density are invasive? What proportion among invasive tumors have positive density?

Exercise 2.3.

1. Upload `tauber.csv`, read data description. Assign column `gender` to variable `G`, column `age` to variable `A`, column `height` to variable `H`, column `weight` to variable `W`. Sort the four variables as discrete or continuous.

2. Display summary, boxplot, histogram, ecdf, for variable `A`.

3. Compute the mean, variance, standard-deviation, median, quantiles at 1/3 and 2/3, inter-quartile range.

4. How many children are younger than 5 years? older than 6 years? What proportion of children are younger than 5 years, older than 6 years? What age is such that 10% of children are younger? What age is such that 10% of patients are older?

5. 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.

6. On the histogram of `A`, superpose a red vertical line marking the mean, a blue vertical line marking the median, two green vertical lines marking the first and third quartiles.

7. On the ecdf plot of `A`, superpose a red vertical line marking the mean, a blue vertical line marking the median, two green vertical lines marking the first and third quartiles. Add a blue horizontal line at 0.5, two green horizontal lines at 0.25 and 0.75.

8. Repeat questions 2. to 7. for variable `H`. For question 4, replace "younger than 5 years" and "older than 6" with "smaller than 1.10 meter" and "taller than 1.20 meter".

9. Repeat questions 2. to 7. for variable `W`. For question 4, replace "younger than 5" and "older than 6" with "lighter than 20 kilograms" and "heavier than 25 kilograms".

3 Session 3

Objectives: start your project by identifying questions, compute the first summary statistics.

Steps of the session:

1. Choose your group and data set.
2. Team work: Search information on the dataset of your project. Identify the questions you want to answer.
3. Share and work: discuss the questions of all the projects.
4. Team work: Compute the summary statistics of your variables.
5. Individual work at home: Read the chapter *Descriptive statistics*, section *Summarizing/presenting couples of variables*.

4 Session 4

Objectives: describe couples of variables; interpret graphically associations of variables; understand the notion of correlation.

Steps of the session:

1. Share and work: discuss the chapter and its instructions.
2. Individual work: Practice with exercises of Session 4 (at least exercises 1 and 3).
3. Share and work: discuss your scripts together.
4. Individual work at home: Read the chapter *Probability*, sections *Introduction, Probability distributions and theoretical characteristics, Discrete distributions*

Exercise 4.1.

1. Upload `bosson.csv`, read data description. Assign column `country` to variable `C`, column `gender` to variable `G`, column `aneurysm` to variable `A`, column `bmi` to variable `B`, column `risk` to variable `R`. Sort the five variables as discrete or continuous. In variable `R`, group classes `3,4,5` into a single class, interpreted as "`3` or more".
2. Assign to variables `A0`, `A1`, `A2`, `A3`, the aneurysm sizes of patients with 0, 1, 2, 3 or more risk factors respectively. Plot the ecdf of `A0` in green. Superpose the ecdf of `A1` in blue, of `A2` in red, of `A3` in black. Interpret the graphic.
3. Assign to variables `AV`, `AF`, the aneurysm sizes of Vietnamese and French patients respectively. Plot the ecdf of `AV` in green. Superpose the ecdf of `AF` in red. Interpret the graphic.
4. Assign to variables `AM`, `AF`, the aneurysm sizes of men and women respectively. Plot the ecdf of `AM` in blue. Superpose the ecdf of `AF` in red. Interpret the graphic.
5. Display boxplots of `A` against `R`. How do they confirm the observation of question 2?
6. Display boxplots of `A` against `C`. How do they confirm the observation of question 3?
7. Display boxplots of `A` against `G`. How do they confirm the observation of question 4?
8. Repeat the same study for the body mass index.

9. Display a scatterplot of `A` against `B`. Do you see a strong dependence between aneurysm and body mass index? Does aneurym size tend to increase with body mass index? Compute the correlation of `A` and `B`. Repeat the plot, coloring in green the points corresponding to Vietnamese patients, in orange the others (use function `ifselse`). Repeat the plot, coloring in blue the points corresponding to men in red to women. Are the observations of the previous questions confirmed?

Exercise 4.2.

1. Upload `fires.csv`, read data description. Assign column `month` to variable `M`, column `day` to variable `D`, column `temp` to variable `TE`, column `relhum` to variable `RH`, column `wind` to variable `W`, column `area` to variable `A`. Sort the six variables as discrete or continuous. Assign the logarithm of `A` to variable `lA`.

2. Display boxplots of `A` against `M`, `A` against `D`. Why do the box plots look so flat? What could be done to improve them?

3. Display scatter plots pf `A` against `TE`, `RH`, `W`. Why are most points at the bottom of the graphic?

4. Define `Az` as the z-scores of variable `A`. Define `Wz` as the z-scores of variable `W`. Plot the ecdf of `Az` in black. Add the ecdf of `Wz` in green. Interpret the differences between the two curves.

5. Compute the correlations of `A` and `TE`, `A` and `RH`, `A` and `W`.

6. Repeat questions 2, 3, 4, 5, replacing `A` with `lA`.

Exercise 4.3.

1. Upload `tauber.csv`, read data description. Assign column `gender` to variable `G`, column `age` to variable `A`, column `height` to variable `H`, column `weight` to variable `W`. Sort the four variables as discrete or continuous.

2. Define `co` as a color vector of `"blue"` for boys, `"red"` for girls. Display scatterplots of height and weight against age, of weight against height, coloring the points with `co`. What differences do you see between the 3 scatterplots?

3. Define `Hz` as the z-scores of variable `H`. Define `Wz` as the z-scores of variable `W`. Plot the ecdf of `Hz` in black. Add the ecdf of `Wz` in green. Interpret the differences between the two curves.

4. Compute the correlations of age and height, of age and weight, of height and weight. Can you understand why the correlation of weight with height is higher than the correlations of height and weight with age?

5. Assign to `AHW` the last three columns of the data matrix. Display paired scatterplots of `AHW` (function `pairs`), coloring points with `co`. Compute the correlation matrix.

5 Session 5

Objectives: summarize the couples of variables of your project.

Steps of the session:

1. Team work: Summarize the couples of statistics of your project dataset.
2. Team work: Write the report corresponding to your results.