Data Analysis and Exploration
A.A. 2010/11 First set of problems.

Some of the problems require the use of file from the (zipped) directory "dati" that can be found on the Web of the course
`http://www.science.unitn.it/AnalisiInfoTLC/LAED`.

If not already done (we did that in class), download it to her/his own account and decompress it.

1. Load (using the command "load"') the file `BBBClub.rda` from the directory "dati". If everything has worked, you will find among the variables [you can check using the command "ls()"] the data.frame BBBClub.

   (a) List the variables contained in the data.frame;

   (b) Prepare a summary table of all variables of the data.frame;

   (c) For each of the two numeric variables `amount` and `first`, draw a histogram and a graph of the (estimated) density, then superimpose the two plots; finally build a box–and–whiskers plot.

   From these graphs can you say anything about the symmetry of the distributions, and the presence of outliers.

   (d) Compare the quantiles of the two distributions with those of a normal distribution, From these can you conclude that we can approximate the distributions with a normal?

   (e) Build a variable `x` that contains the values of `BBBClub$amount` for the observations corresponding to females (i.e. "gender: female"). Compare, through suitable summary indications, the distribution of `x` to that of `BBBClub$amount`.

2. Read the file "studenti.txt" (it is a text file to be read using "read.table") from the subdirectory `Crivellari_data` putting the output in a dataframe of R; the names of the variables are "Voto" (= mark), "Luogo" (= place, with an unknown code) and "Sesso" (=sex). Compute median and mean of mark for the whole sample, and for males and females separately. Plot a histogram of marks, separately for males and females.

   One may note that some students have mark 0; presumably, this is a code for students who have not passed the exam. Compute the proportion of students that have not passed the exams, and compute mean and median of the other ones.

3. Read the file "Tests.txt" (it is a file formatted CSV which has to be read using the command "read.csv" or "read.csv2") from the subdirectory `Crivellari_data` putting the output in a dataframe of R; the variables are named `Peso;Sex;Eta;LivScol;StCiv;Test.A;Test.B;Test.C;Test.D`. Rename the variable "Peso" as "Weight" and "Eta" as "Age". Make a bivariate plot of "Weight" and "Age" (some data appear very unlikely...). Transform "Sex" and "LivScol" into qualitative variables (the numbers in the set are codes for a class); separate the plot of "Weight" vs. "Age" according to the values of "Sex".

4. Give a definition of a quantile for a continuous (theoretical) distribution, and for an empirical sample.

5. Let $Y = X\beta + \varepsilon$, where ε are Gaussian (=normal) random variables with $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Cov}(\varepsilon) = W$, where W is a known positive definite matrix[1] X is a constant (known) matrix, while β is the vector of (unknown) parameters.

   (a) Show that the maximum likelihood estimate of β is the point of minimum of $(Y - X\beta)^t W^{-1}(Y - X\beta)$.

   (b) Write explicitly this minimum problem in the case when W is a diagonal matrix with positive entries σ_i^2.

   (c) Show that this can be phrased as a problem of minimum distance from a subspace in the Euclidean (Hilbert) space $\mathbb{R}^n$ with scalar product $(u, v) = \langle W^{-1/2}u, W^{-1/2}v\rangle$ where $\langle\cdot,\cdot\rangle$ is the standard scalar product, $W^{1/2}$ is the unique [2] positive definite matrix such that $W^{1/2}W^{1/2} = W$ and $W^{-1/2}$ is its inverse.

   (d) Find the minimum point through an orthogonal projection (relatively to the scalar product $(\cdot,\cdot)$), showing that the estimate of β is
   $$\hat{\beta} = (X^t W{-}1X)^{-1}X^t W^{-1}Y.$$

6. Consider the linear regression model $Y = a + bX + \varepsilon$ where X is a variable that takes on only the values 0, 1 and -1.

[1] Look at my notes (in Italian) at `http://www.science.unitn.it/AnalisiInfoTLC/LAED/olds/note/note.pdf` or at any book of probability theory for the density function of a multivariate normal with covariance matrix $(n \times n)$ W. Otherwise analyse the simpler case where W is a diagonal matrix with positive entries.

[2] its existence and uniqueness can be proved through the diagonalization of W.

(a) Assume for simplicity that the first n_0 values are equal to 0, the following n_1 are equal to 1 and the final $(n - n_0 - n_1)$ ones are equal to -1. Which is the matrix X when writing the linear model $Y = X\beta + \varepsilon$?

(b) Which are the resulting estimates for a and b? What is their interpretation?

(c) If X is interpreted as a qualitative variable, the correct matrix X to use in a linear model[3] has n rows and 3 columns where the first column has all 1's, the second column has 1 on the first n_0 rows and 0 on all the others, the third column has 1 on the second n_1 rows and 0 on all the others. What is then the interpretation of the elements of the vector $\hat{\beta}$? Can you write explicitly the formulae?

7. The file "savings.txt" in the directory "/dati/ascdata" contains economic data (averages in the period 1960-70) on 50 states in the world.

   The variables are 'sr' (savings rate), 'pop15' (fraction of the population under 15 years), 'pop75' (fraction of the population above 75 years), 'dpi' (available income per person, in dollars), 'ddpi' (percentage increase of available income per person).

   (a) Plot all variables, finding cases of particularly asymmetric distributions that suggest a transformation of data.

   (b) Perform a regression of 'sr' on 'pop15' plotting the points together with the regression line.

   (c) Add the variable 'ddpi' as predictor variable; perform the new regression, and discuss whether it is justified adding 'ddpi' as predictor variable.

   (d) Plot the residuals of this last regression, and perform some diagnostic (visual) test on them.

[3] R may actually produce a somewhat different matrix, but the conclusions are the same