# assignment_1

*Agustín Valencia*

*11/5/2019*

## Question 1: Describing individual variables

Consider the data set in the T1-9.dat file, National track records for women. For 55 different countries we have the national records for 7 variables (100, 200, 400, 800, 1500, 3000m and marathon ). Use R to do the following analyses.

a) Describe the 7 variables with mean values, standard deviations e.t.c.

b) Illustrate the variables with different graphs (explore what plotting possibilities R has). Make sure that the graphs look attractive (it is absolutely necessary to look at the labels, font sizes, point types). Are there any apparent extreme values? Do the variables seem normally distributed? Plot the best fitting (match the mean and standard deviation, i.e. method of moments) Gaussian density curve on the data's histogram. For the last part you may be interested in the hist() and density() functions.

## Question 2: Relationships between the variables

a) Compute the covariance and correlation matrices for the 7 variables. Is there any apparent structure in them? Save these matrices for future use.

b) Generate and study the scatterplots between each pair of variables. Any extreme values?

c) Explore what other plotting possibilities R offers for multivariate data. Present other (at least two) graphs that you find interesting with respect to this data set.

## Question 3: Examining for extreme values

a) Look at the plots (esp. scatterplots) generated in the previous question. Which 3–4 countries appear most extreme? Why do you consider them extreme? One approach to measuring "extremism" is to look at the distance (needs to be defined!) between an observation and the sample mean vector, i.e. we look how far one is from the average. Such a distance can be called an multivariate residual for the given observation.

b) The most common residual is the Euclidean distance between the observation and sample mean vector, i.e.

$$d(\overrightarrow{x}, \overline{x}) = \sqrt{(\overrightarrow{x} - \overline{x})^T (\overrightarrow{x} - \overline{x})}$$

This distance can be immediately generalized to the $L^r$, r > 0 distance as

$$d_{L^r}(\overrightarrow{x}, \overline{x}) = \left( \sum_{i=1}^{p} |\overrightarrow{x}_i - \overline{x}_i|^r \right)^{1/r}$$

where p is the dimension of the observation (here p = 7).

Compute the squared Euclidean distance (i.e. r = 2) of the observation from the sample mean for all 55 countries using R's matrix operations. First center the raw data by the means to get $(\overrightarrow{x} - \overline{x})$ for each country. Then do a calculation with matrices that will result in a matrix that has on its diagonal the requested squared

distance for each country. Copy this diagonal to a vector and report on the five most extreme countries. In this questions you MAY NOT use any loops.