LAB-2

Getting to know Data: Iris blossoms

The data set iris contains measurements of the length and the width (in cm) of petals and sepals of three iris species: 1: Setosa, 2: Versicolor and 3: Virginica.

a) This data set iris is already part of the standard R installation. Consider the object iris. How is it structured? How many observations (lines) does it contain? How many variables (columns)? Hint: nrow(), ncol(), dim(), str()

b) To get an overview of the range of values, look at the summary() of the data set. Which information on the data set does it provide?

c) For the variable Sepal.Length check the results above by using the R-functions min(), max(), mean(), median(), quantile(). If necessary, make use of the help functions ?quantile etc.

Missing Values

Statistics needs data. Unfortunately, data often cannot be collected fully. Therefore many data sets contain \gaps", non-existing measurements, so-called NAs (not available). In this exercise you will get to know how R deals with NAs. We work with the data set iris. Make a copy of the iris data set by d.iris <- iris.

a) Assume that we were unable to take the second observation of Petal.Length and Petal.Width, and for the fifth observation, the data for Sepal.Length, Sepal.Width and Petal.Width are missing. Replace these five fields by NA.
Hint: Replace the values by NAs using e.g. d.iris[2, 3:4] <- NA

b) Consider the rst eight observations of the modified data set, to observe how the NAs are displayed by R. The commands class(), nrow(),ncol(), dim(), str() also work for the data set with missing values. What changes in the summary()?

c) Try to confirm the given values for the variable Sepal.Length using min(), max(), mean(), median(), quantile(). Is there a difference?

d) There are functions that cannot handle NAs (Result 'NA' or 'Error: missing observations'). There is a trick to make them calculate the correct results: simple functions such as min(), max(), mean(), median(), quantile(), range() etc. can take an argument na.rm. When you set its value to TRUE, the NAs will not be considered in the calculation.
Try to confirm the values provided by summary() again, using this new argument.

e) Why should missing values always be coded by NA, and not, for instance, filled with a zero? Explain for the case of the mean() function.

f) Experiment with missing values in the statistical functions var(), sd(), cor(). Can you ex-plain the behaviour of R?

g) Select only those observations which have missing values in either Sepal.Length or in Petal.Length.
Hint: is.na()

h) The function na.omit() eliminates all observations from the data frame for which any(!) variable contains NAs. Save the result of na.omit(d.iris). How many observations remain? How many remain using na.omit(d.iris[,1:3])?

Note:

Higher-level functions such as t.test() or wilcox.test() have an argument na.action, with which the reaction to NAs can be determined. na.action=na.omit first deletes all lines (observations) with NAs before anything is calculated.