Assignment 2
Inference about mean vectors

Kurskod och namn:	732A97 Multivariate Statistical Methods
Delmomentsansvarig:	Krzysztof Bartoszek, Hao Chi Kiang
Instruktioner:	This assignment is part of the examination for the Multivariate Statistical Methods course You will work in groups. Submit your report as a **.PDF** file Be concise and do not include unnecessary printouts and figures produced by the software and not required in the assignments. **All code (R) should be included as an appendix into your report.** A typical report should contain 2–4 pages of text plus some amount of figures plus appendix with codes. In the report reference **ALL** consulted sources and disclose **ALL** collaborations. The report should be handed in via LISAM (or alternatively in case of problems e–mailed to hao.chi.kiang@liu.se or krzysztof.bartoszek@liu.se), by **23:59 1 December 2019** at latest. **Late submission may result in an additional penalty assignment.** The report should be written in English. Notice there is a final deadline of **23:59 2 February 2020** after which no submissions nor corrections will be considered and you will have to redo the missing labs next year.

Assignment developed by Ann–Charlotte Hallberg and Bertil Wegmann.

Learning objectives

After reading the recommended text and doing the assignment the student shall be able to:

- formally test outliers in a data set
- test a given mean vector
- compute a confidence region (ellipsoid) and simplifications of it in the form of simultaneous confidence intervals
- compare two or more mean vectors

Recommended reading

Chapters 4–6 in *Johnson, Wichern*

Focusing on the multivariate normal distribution, we will study methods for estimating, testing hypotheses about and comparing mean vectors. These methods are the multivariate generalizations of the univariate methods.

Question 1: Test of outliers

Consider again the data set from the `T1-9.dat` file, National track records for women. In the first assignment we studied different distance measures between an observation and the sample average vector. The most common multivariate residual is the Mahalanobis distance and we computed this distance for all observations.

a) The Mahalanobis distance is approximately chi–square distributed, if the data comes from a multivariate normal distribution and the number of observations is large. Use this chi–square approximation for testing each observation at the 0.1% significance level and conclude which countries can be regarded as outliers. Should you use a multiple–testing correction procedure? Compare the results with and without one. Why is (or maybe is not) 0.1% a sensible significance level for this task?

b) One outlier is North Korea. This country is not an outlier with the Euclidean distance. Try to explain these seemingly contradictory results.

Question 2: Test, confidence region and confidence intervals for a mean vector

Look at the bird data in file `T5-12.dat` and solve Exercise 5.20 of *Johnson, Wichern*. Do not use any extra R package or built–in test but code all required matrix calculations. You **MAY NOT** use loops!

Question 3: Comparison of mean vectors (one–way MANOVA)

We will look at a data set on Egyptian skull measurements (published in 1905 and now in `heplots` R package as the object `Skulls`). Here observations are made from five epochs and on each object the maximum breadth (mb), basibregmatic height (bh), basialiveolar length (bl) and nasal height (nh) were measured.

a) Explore the data first and present plots that you find informative.

b) Now we are interested whether there are differences between the epochs. Do the mean vectors differ? Study this question and justify your conclusions.

c) If the means differ between epochs compute and report simultaneous confidence intervals. Inspect the residuals whether they have mean 0 and if they deviate from normality (graphically).

Tip: It might be helpful for you to read Exercise 6.24 of *Johnson, Wichern*. The function `manova()` can be useful for this question and the residuals can be found in the `$res` field.