Sheet 4

Hand in solution by April 4 in the lecture room

1. **Distribution of the estimated correlation.** In this exercise we explore how well we can estimate the correlation between two (jointly normal) random variables.

   a) Generate $n = 100$ random samples from a bivariate normal random vector with correlation $\rho = 0.7$.
   Hint: To generate multivariate normal random vectors, use the function `mvrnorm` from the package `MASS`. Load the package with `library(MASS)`. `mvrnorm(n,m,Sigma)` produces a sample of size `n` from the multivariate normal distribution with mean `m` and covariance matrix `Sigma`.

   b) Compute the sample correlation between the two variables of the sample generated in **a)**. Compare the estimate to the theoretical true value of 0.7. Why are they not equal?

   We define the following function:

```
generateCorValue <- function(n,m,Sigma){
  data <- mvrnorm(n,m,Sigma)
  cor(data[,1],data[,2])
}
```

   Calling `generateCorValue(n,m,Sigma)`, we get the sample correlation for a new sample of size `n` from the multivariate normal distribution with mean `m` and covariance matrix `Sigma`.

   c) Use the function `generateCorValue` to estimate the correlation a thousand times, and depict the estimated values with a histogram. How big is the sample standard deviation of the estimated correlations?
   Hint: `t.rho <- replicate(1000,generateCorValue(n,m,Sigma))` calls our function 1000 times with the given arguments and stores the 1000 values in the vector `t.rho`.

   d) (*) The more observations we generate (that is, the larger n grows), the closer the estimators are to the true value 0.7. In particular, the standard deviation of the estimators should go to zero. To check this behaviour, write a new function that performs all the previous steps and provides the sample standard deviation of the correlation estimators as output. Compute the sample standard deviation of the correlation estimators for $n = 10, 20, 50, 100, 200, 500$ (sample size). Plot the results.

   e) (*) From the theory we know that the standard deviation of the estimators should be approximately proportional to $n^{-1/2}$. Verify whether our experiment matches this. For this purpose, plot the logarithm of the sample standard deviations of the estimators vs. $\log(n)$. You should obtain a slope of approximately -0.5.

2. Mahalanobis distances, Q-Q plot.

Load the function `f.qqchi` using:
`source("http://stat.ethz.ch/education/semesters/ss2011/ams/f.qqchi.R")`.
Take a look at the function `f.qqchi`. You can display the source code by typing `f.qqchi` in the R console. Try to understand what the code does.

a) Generate a random sample from a bivariate normal distribution and make a Q-Q plot of the Mahalanobis distances. The plot will look different for every new sample. Use different sample sizes, and try to develop a feeling for what Q-Q plots using normal data should look like.
Hint: Generate the samples using `mvrnorm` as in Exercise 1. `mvrnorm` returns a matrix. Convert the result into a data frame using `data <- as.data.frame(mvrnorm(?))`, then you can use `mean` and `cov` as usual. To compute the distances, use
`D <- sqrt(mahalanobis(data,center=?,cov=?))`. For `center` and `cov`, use the sample mean and sample covariance matrix, respectively. The function `f.qqchi(D,df)` plots the empirical quantiles of `D` against the square roots of the corresponding theoretical quantiles of the χ^2-distribution with `df` degrees of freedom. Choose the appropriate value for `df`.

b) To develop a feeling for "bad" Q-Q plots, take normal samples as in **a)**, apply a nonlinear transformation (e.g., `exp`, `abs`, squaring), and look at the Q-Q plot for the transformed data.

The file `clayton.dat` contains 250 realizations from an unknown bivariate distribution. Load the data:

```
t.url <- "http://stat.ethz.ch/education/semesters/ss2011/ams/clayton.dat"
data <- read.table(t.url, header=TRUE)
```

We want to explore whether the data are bivariate normal.

c) Provide a scatter plot of the data, and histograms of each of the variables. Do these "marginal distributions" look normal? Is it plausible that the data are bivariate normal?
Hint: For a nice plot, use `pairs(data,diag.panel=panel.hist)` as in 1.c) on Sheet 1.

d) Make a Q-Q plot of the Mahalanobis distances using the function `f.qqchi`. Does the result contradict your findings from **c)**?

3. Hotelling's T^2-test. We consider again the data `clayton.dat` from the previous exercise. We want to test whether the mean is significantly different from $\underline{0}$ using Hotelling's T^2-test. Hotelling's T^2-test assumes normality for the data—which may not be the case here, but for large sample sizes, the test is still approximately correct due to the central limit theorem.

a) Write down the definition of the T^2 test statistic used for Hotelling's T^2-test of the null hypothesis $H_0 : \underline{\mu} = \underline{0}$. What is the distribution of this statistic under $H_0 : \underline{\mu} = \underline{0}$?

b) Compute the value of T^2 for the given data.
Hint: Use the function `mahalanobis`.

c) Compute the p-value for the given data. Can we reject H_0 on the 5% level?
Hint: Recall that the p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. `pf(q,df1,df2)` gives the value of the cumulative distribution function of the F-distribution $\mathcal{F}(df1, df2)$ at the point `q`.