Statistical Data Analysis
Exercise Session 3

1. (a) Generate several times $n = 20$ data points from a normal distribution with mean $\mu = 2$, and standard deviation $\sigma = 3$. Make a normal QQ plot and view the variation in the results.

   (b) Compare with the situation $n = 100$ and $n = 500$.

2. (a) Generate several times $n = 20$ data points from a 4−variate normal distribution with mean $\mu = \mathbf{0_4}$ and $\mathbf{\Sigma} = \mathbf{I}_4$. Make a χ^2 QQ plot of the Mahalanobis distances and view the variation in the results.

   (b) Compare with the situation $n = 100$ and $n = 500$.

3. (a) Load the ces.csv dataset.

   (b) Create the matrix FDHOFDAW which contains values of the variables FDHO and FDAW for all data points for which both observations are strictly greater than 0.

   (c) For each of the variables in FDHOFDAW:

      i. Investigate the normality of the original variable.

      ii. Determine the optimal Box-Cox transformation. Which transformation would you implement in practice?

      iii. Investigate the normality of the transformed variable.

   (d) For both variables in FDHOFDAW together:

      i. Investigate the bivariate normality of the original variables.

      ii. Investigate the bivariate normality of the transformed variables.

4. (a) Load the hills data set from the MASS library.

   (b) Make a scatter plot of the variables dist and climb. Add the tolerance ellipse to this, based on the classic average and the classic covariance matrix. Also add the MCD-based tolerance ellipse. Compare the robust distances with the Mahalanobis distances. Do this for alpha = 0.5 and alpha = 0.75 and compare the results.

5. (a) Generate $n = 100$ data points from an exponential distribution with $\lambda = 4$. Define a function that calculates the negative log likelihood of *lambda*, given the data. Calculate a numerical approximation for the MLE estimate for λ with the mle function in package stats4.

   (b) Generate $n = 100$ data points from a normal distribution with $\mu = 3$ and $\sigma = 2$. Define a function that calculates the negative log likelihood of (μ, σ) given the data. Calculate a numerical approach for the MLE estimates for (μ, σ).

   (c) Generate n = 100 data points from a bivariate normal distribution with $\mu = (3, 4)^t$ and

   $$\Sigma = \begin{bmatrix} 2 & -0.5 \\ -0.5 & 1 \end{bmatrix}.$$

   Define a function that calculates the negative log likelihood of (μ, Σ), given the data. Calculate a numerical approach for the MLE estimates for (μ, Σ).

Some useful functions:

1. qqplot() and qqline()
2. read.csv() and read.table()
3. which()
4. shapito.test()
5. bcPower in package car and boxcox in package MASS
6. covMcd() in the robustbase library.
7. rmvnorm and dmvnorm in package mvtnorm