# Computer problem 1

Tailor-made `m`-files e.g. `jackknif.m`, `jackstat.m`, `bca.m`, `neigen.m`, `eigen.m` and `yuwaest.m` can be found on the course home page.

`http://www.math.kth.se/matstat/gru/sf2955/uppgift.html`

What to hand in
Hand in your m-files and the Matlab-log which can be obtained by using the command "`diary` *filename*" (is turned off with the command `diary off`).

In addition I want you to write a small description of what has been done and what conclusions can be made.

# 1 Simulation

1)
We are interested in examining the properties of the estimate $\widehat{\lambda} = 1/\bar{x}$ of the failure rate for exponentially distributed data when we have $n = 10$ observations. The function `exprnd` can be used. Simulate a large number of such samples and examine the distribution of $\widehat{\lambda}$. Show the result in the form of a histogram and calculate the expected value and variance.
One can show that $E(\widehat{\lambda}) = \lambda \cdot n/(n-1)$ - is this compatible with your simulation results?
2)
Simulate samples consisting of 20 two-dimensional normally distributed random variables which have correlation $\rho$. In order to get a two-dimensional observation $(X, Y)$ which has $E(X) = E(Y) = 0$ and $V(X) = V(Y) = 1$ and where the correlation is $\rho$. You can use the Choleski-method by getting two $N(0,1)$-distributed random variables $Z_1$ and $Z_2$ and then form $X = Z_1$ and $Y = \rho Z_1 + \sqrt{1-\rho^2} Z_2$. Estimate $\rho$ with the plug-in-estimate $\widehat{\rho}$ i.e.

$$\widehat{\rho} = \frac{\sum_{i=1}^{20}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{20}(x_i - \bar{x})^2 \sum_{i=1}^{20}(y_i - \bar{y})^2}}.$$

Simulate this for $\rho = 0.9$ and make a histogram. Study also the distribution of $h(\widehat{\rho})$ where

$$h(x) = \frac{1}{2} \ln\left(\frac{1+x}{1-x}\right)$$

i.e.the Fisher-transformation (see section 5.5). Make a histogram and calculate the expected value and variance in the distribution of $h(\widehat{\rho})$. Compare this with the theoretical asymptotic result for the Fisher-transformation.

# 2 Bootstrap - bias correction

Let $X_1, X_2, \cdots, X_{10}$ be independent uniformly distributed random variables on the intervall $[0, \theta]$ with $\theta = 1$. Let $X_{(1)}, X_{(2)}, \cdots, X_{(10)}$ be these sorted in increasing order. We estimate $\theta$ with the plug-in-estimate $X_{(10)}$.
a) Get the distribution of $X_{(10)}$, i.e. for the maximum of 10 U(0,1)-distributed random variables. Calculate the expected value and the variance. What is the bias of $\widehat{\theta} = x_{(10)}$?
b) Show that in using the bootstrap from the observations $x_1, x_2, \cdots, x_{10}$ we get for $\widehat{\theta}^* = X^*_{(10)}$ the bootstrap-distribution

$$P(\widehat{\theta}^* = x_{(i)}) = (\frac{i}{10})^{10} - (\frac{i-1}{10})^{10}, \text{ for } i = 1, 2, \cdots, 10.$$

c) Calculate the bias-corrected estimate $\bar{\theta}$ of $\theta$ based on $\widehat{\theta} = x_{(10)}$ which is obtained using the bootstrap and calculate its (theoretical) expected value. Compare with the result in the a-part.
d) Simulate the distribution of $\bar{\theta}$ and calculate from this its expected value and variance and compare with the corresponding values for $\widehat{\theta} = X_{(10)}$.

# 3 Bootstrap and jackknife

Generate two samples with 20 and 100 observations from some distribution, e.g. the exponential distribution or the normal distribution. This can be done with the procedures `exprnd` and `normrnd` respectively in Matlab's statistics toolbox (Stats-module).

I want you to play around with these data and get a feeling for how the bootstrap and the jackknife methods work. Very little has to be documented but draw some conclusions about how they work for simple estimates and how the bootstrap results depend on the number of bootstrap samples.

## 3.1 Jackknife

Estimate for the two samples the following quantities with jackknife using `jackknif` och `jackstat` which can be obtained from the course home page.

Jackknife:

1) Standard error with Jackknife

2) The bias correction with Jackknife

## 3.2 Bootstrap

Use e.g. $B = 50, 100, 500, 1000$ and (if possible 10000). Note that subseries can be used from the same bootstrap-generation.

1) Standard error with Bootstrap.

2) Bias correction with Bootstrap

3) Plot the distribution of $\widehat{\theta}^*$ i.e. the bootstrap distribution with `hist`.

Do the above for the following parameters after writing down the plug-in-estimates.

a) $\theta = E(X)$ i.e. the expected value

b) $\theta = D(X)$ i.e. the standard deviation

c) $\theta$ =the median, i.e. the 50%:point of the distribution

d) $\theta$ =the inter-quartile distance, i.e. the difference between the two points where 25% of the probability mass is to the left and to the right. You need to write your own `m`-file. You can use `prctile`. Another alternative is to sort the vector and take the appropriate quantities.

Hints: You can write your own `m`-file which gives the desired results and where you only change in-data and estimation function.

# 4 Multi-dimensional observations

The procedure `neigen.m` produces a matrix of multidimensional normally distributed observations - one for each row - with a prescribed $A$-matrix. This means that row $k$ is an observation of $\boldsymbol{X} = A\boldsymbol{Z}$ where $\boldsymbol{Z}$ consists of independent N(0,1)-distributed random variables. Thus $\boldsymbol{X}$ has the covariance matrix $C = AA^T$. The procedure `neigen.m` also gives the eigen values for $C$ and the value of

$$\theta = \frac{\max(\text{eigen values})}{\text{sum of eigen values}}$$

which is the parameter to be estimates using the observed data.

Select an $A$-matrix and generate a suitable number of observations ($n = 20$ or $n = 100$) using `neigen.m`.

Estimate $\theta$ using `eigen.m` and perform jackknife and bootstrap using the procedure `eigen.m` and estimate the standard error for $\widehat{\theta}$.

Calculate also a confidence interval using the percentile method (directly from the bootstrap distribution) and also using the $BC_a$-method (use the procedure `bca.m`.

## 4.1 Correlation

Generate two-dimensional data - e.g.the LSAT/GPA data (Efron's data) which can be loaded using the command `load lawdata` in Matlab which gives two vectors `lsat` and `gpa` which contain the data. If you want them in one vector `x` you can write

`x=[lsat gpa]`

An alternative way to generate two-dimensional data is to use `neigen.m` as above.

Estimate from your data the following quantities:

a) The correlation coefficient. The procedure `corrcoef` is useful. Remember that is gives the whole correlation matrix.

b) Estimate the parameters $a$ and $b$ in a linear regression $y = a + bx$ for your data. The procedure `regress` can be useful.

Estimate for these parameters the bias and the standard error using jackknife and bootstrap (bootstrap of points)

# 5 Regression

Linear regression with $y = a + bx + cx^2$ as the theoretical regression relation where we are interested in $\theta = -b/2c$ (which gives the maximum value of the regression function).

Generate e.g. $y = 10 + 50x - x^2$ for $x = 1, 2, \cdots, 50$ and with random errors which are $N(0, 20^2)$.

## 5.1 Jackknife

Estimate the bias for $\theta = -b/2c$ using the jackknife.

Estimate the standard error for $\theta = -b/2c$ using the jackknife.

## 5.2 Bootstrap

Use
a) the method of bootstraping points

b) the method of bootstraping residual
for the following problems

a) Estimate standard error and bias for $\theta = -b/2c$

b) Draw a histogram for the bootstrap distribution for $\theta = -b/2c$.

Calculate a confidence interval for $\theta = -b/2c$ using the $BC_a$-method.

# 6 Pivotal-based confidence intervals

1) Generate e.g. 20 Exp($\theta$)-distributed (expected value $\theta$) observations with e.g. $\theta = 5$. Use `exprnd`.

a) Calculate with the pivotal method a 90% confidence interval for $\theta$ based on $\widehat{\theta}/\theta$.

b) Calculate a confidence interval using the $BC_a$-method.

c) Calculate an exact confidence interval by using the $\Gamma$-distribution. The procedure `gaminv` can be used.

Compare the bootstrap distribution in the a-part with the true distribution of $\widehat{\theta}/\theta = \bar{X}/\theta$, i.e. (in principle) a $\Gamma$-distribution - use `gampdf` or `gamcdf`.

2) Generate e.g. 20 normally distributed observations from $N(\theta, \sigma)$. Use `normrnd`.

a) Calculate a confidence interval for $\theta$ based on

$$\frac{\widehat{\theta} - \theta}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - \theta}{\frac{s}{\sqrt{n}}}$$

b) Calculate a confidence interval using the $BC_a$-method.

c) Calculate an exact confidence interval based on the $t$-distribution - use `tinv` or a table.

Compare the $t$-distribution (use `tpdf` or `tcdf`) with the bootstrap-distribution.

3) Create two normally distributed samples of length 10 and 20 from $N(m_1, \sigma_1^2)$ and $N(m_2, \sigma_2^2)$. Calculate using the pivotal method a confidenceinterval for $m_1 - m_2$ based on a suitable (approximate) pivotal quantity. You may not asume that $\sigma_1 = \sigma_2$.

# 7 Time series

Generate an $AR(1)$-time series of length e.g. 100, i.e.

$$y_{t+1} = ay_t + \epsilon_t, \ t = 0, 1, 2, \cdots, 100$$

with e.g. $a = 0.9$ and where the $\epsilon_t$:s are independent $\mathrm{N}(0, 5^2)$. Start the recursion with $y_0 = 0$.

Estimate $a$ with the `m`-file `yuwaest.m` which estimates the parameters in an $AR(p)$-process (i.e. we have $p = 1$). This file can be found in the `funk`-directory or can be downloaded from the home page. `yuwaest` uses the Yule-Walker-equations with the syntax

```
[aest sest C]=yuwaest(y,1)
```

where `aest` is the estimate of $a = 0.9$, `sest` is an estimate of $\sigma = 5$ and `C` is the traditional estimate of the standard error of the $a$-estimate taught e.g. in courses in time series.

Use bootstrap to get an idea of the distribution of the $a$-estimate. You should therefore generate new time series by using the estimated value of $a$ and bootstraping the estimated residuals in the original time series. Is the estimate unbiased? What is its variance?