NMST434 – Modern statistical methods (2017–18)

Homework assignments

General instructions

- All solutions must be delivered **by the time the exercise class starts**.
- Solutions can be delivered either at the beginning of the exercise class, or left in the mailbox of S. Nagy in the corridor of the Department of Probability and Math. Statistics (first floor, the door on the right from the staircase). PDF documents created in LaTeX can be submitted also via e-mail to `nagy@karlin.mff.cuni.cz`. In all cases, in the header of the document clearly state ***NMST434, S. Nagy***.
- Hand written solutions are completely fine, but must be written in a **readable** way.
- The language of the homework reports can be either **English** or **Czech/Slovak**.
- If the number of your student card is needed for the assignment, then include this number at the beginning of your solution of the assignment.
- In case of **plagiarism** all authors get zero points.
- If the homework includes analysis of (real or simulated) data, it is expected that you also **numerically calculate** the required estimators, confidence intervals, test statistics ... Do not also forget to **specify the assumed model** and give **the formulas** so that it is clear how the result is calculated.
- If not stated otherwise use 5 % as the level (prescribed probability of type I error) of the test and 95 % as the coverage of the confidence interval.

In what follows AAA stands for the number of your student identity card.

Homework 1 (8 p) - deadline 6. 3. 2018

Let $X_0, X_1, X_2, \ldots$ be a sequence of random variables such that $|X_n| < 1$ almost surely for all $n = 0, 1, 2, \ldots$. Prove, or find a counter-example:

(i) $X_n \xrightarrow[n\to\infty]{d} X_0$ implies $\mathsf{E}\, X_n \xrightarrow[n\to\infty]{} \mathsf{E}\, X_0$.

(ii) $X_n \xrightarrow[n\to\infty]{d} X_0$ implies $\mathsf{E}\, |X_n - X_0| \xrightarrow[n\to\infty]{} 0$.

(iii) $\sqrt{n}X_n \xrightarrow[n\to\infty]{d} X_0$ implies $\mathsf{E}\, \sqrt{n}X_n \xrightarrow[n\to\infty]{} \mathsf{E}\, X_0$.

Homework 2 (10 p) - deadline 6. 3. 2018

We observe independent random variables $X_1, \ldots, X_n$ generated by the following procedure. First, a random sample $W_1, \ldots, W_n$ is drawn from Bernoulli distribution with the probability of success $1/2$. If $W_i = 0$, the distribution of X_i is given by

$$\mathsf{P}\,(X_i = 0) = 1 - p_1, \quad \mathsf{P}\,(X_i = 1) = p_1,$$

with $p_1 \in (0, 1)$; if $W_i = 1$, the distribution of X_i is given by

$$\mathsf{P}\,(X_i = 1) = 1 - p_2, \quad \mathsf{P}\,(X_i = 2) = p_2,$$

for $p_2 \in (0, 1)$. Parameters $\boldsymbol{\theta} = (p_1, p_2)^\mathsf{T}$ are unknown.

(i) Using the moment method, find an estimator $\widehat{\boldsymbol{\theta}}_n$ of the unknown vector $\boldsymbol{\theta}$.

(ii) Derive the asymptotic distribution of $\widehat{\boldsymbol{\theta}}_n$.

(iii) Use the dataset hw2data that is available on the webpage (**here**) of the course and generate your data as

```
load("hw2data.RData")
set.seed(AAA)
X <- sample(hw2data, size=200)
```

Variable X contains a dataset of size $n = 200$ from the model above. Provide a confidence region for the parameter $\boldsymbol{\theta}$. Does the point $(1/2, 1/2)^\mathsf{T}$ lie inside that region? Interpret your result.

Homework 3 (12 p) - deadline 13. 3. 2018

Consider a random sample $X_1, \ldots, X_n$ as in Homework 2.

(i) Construct the maximum likelihood estimator $\widetilde{\boldsymbol{\theta}}_n$ of the parameter $\boldsymbol{\theta}$ and derive its asymptotic distribution.

(ii) Based on either the estimator $\widehat{\boldsymbol{\theta}}_n$ from Homework 2 or $\widetilde{\boldsymbol{\theta}}_n$, construct a test of the null hypothesis $H_0 : p_1 \leq p_2$ against the alternative $H_1 : p_1 > p_2$.

(iii) In R, simulate a random sample $X_1, \ldots, X_n$ of size $n = 100$ for several choices of p_2, with $p_1 = 1/2$. In each simulation, perform the test proposed in part (ii). Repeat this procedure $B = 1000$ times for each chosen p_2. Report the average rejection rate for the chosen values of p_2, either in the form of a table, or in the form of a plot. Comment on your results.

Homework 4 (13 p) - deadline 20. 3. 2018

Let $(Y_1, \boldsymbol{X}_1^\mathsf{T}, \mathbf{Z}_1^\mathsf{T})^\mathsf{T}, \ldots, (Y_n, \boldsymbol{X}_n^\mathsf{T}, \mathbf{Z}_n^\mathsf{T})^\mathsf{T}$ be independent random vectors that have the same distribution as the generic vector $(Y, \boldsymbol{X}^\mathsf{T}, \mathbf{Z}^\mathsf{T})^\mathsf{T}$, where $\boldsymbol{X} = (X_1, \ldots, X_d)^\mathsf{T}$ and $\mathbf{Z} = (Z_1, \ldots, Z_q)^\mathsf{T}$ are random vectors. Suppose that the conditional distribution of Y given $(\boldsymbol{X}^\mathsf{T}, \mathbf{Z}^\mathsf{T})^\mathsf{T}$ is normal $\mathcal{N}\big(\beta_0 + \boldsymbol{\beta}_X^\mathsf{T}\boldsymbol{X} + \boldsymbol{\beta}_Z^\mathsf{T}\mathbf{Z}, \sigma^2\big)$, where the parameters $\beta_0, \boldsymbol{\beta}_X, \boldsymbol{\beta}_Z, \sigma^2$ are unknown. The distribution of $(\boldsymbol{X}^\mathsf{T}, \mathbf{Z}^\mathsf{T})^\mathsf{T}$ does not depend on the parameters $\beta_0, \boldsymbol{\beta}_X, \boldsymbol{\beta}_Z, \sigma^2$. Derive the likelihood ratio test, Wald test and Rao score test for testing the null hypothesis $H_0 : \boldsymbol{\beta}_Z = \mathbf{0}_q$ against the alternative $H_1 : \boldsymbol{\beta}_Z \neq \mathbf{0}_q$. Compare the derived tests. Finally, compare the derived tests with the standard test that you know from the linear regression course.

Homework 5 (12 p) - deadline 27. 3. 2018

Let $(Y_1, X_1, Z_1)^\mathsf{T}, \ldots, (Y_n, X_n, Z_n)^\mathsf{T}$ be a random sample such that the conditional distribution of Y_1 given X_1 and Z_1 has density

$$f(y|x,z) = \begin{cases} \mathrm{e}^{\alpha+\beta x+\gamma z} \exp\left(-y\,\mathrm{e}^{\alpha+\beta x+\gamma z}\right) & \text{if } y > 0, \\ 0 & \text{otherwise,} \end{cases}$$

and the distribution of X_1 and Z_1 does not depend on the unknown parameters α, β, and γ.

(i) Derive the expression for the profile log-likelihood of parameters β and γ.

(ii) Generate data in the following way:

```
set.seed(AAA);
n <- 25;
X <- -.5*rexp(n);
Z <- rexp(n);
alpha = 1; beta = 2; gamma = 3;
Y <- rexp(n, rate = exp(alpha + beta*X + gamma*Z));
```

For the generated dataset, plot the profile log-likelihood of β and γ. Find the maximum likelihood estimate of α, β, and γ, and visualise the 95 %-confidence region for β and γ based on the likelihood ratio test.

(iii) Find the asymptotic confidence ellipse for β and γ based on some application of the Wald approach. Plot this ellipse in the same figure as the confidence region from part (*ii*). Which confidence region do you prefer and why?

Hint: For the numerical optimization and visualisation, `R` functions `optim`, `ellipse` (package `car`), and `contour` might be of interest.

Homework 6 (15 p) - deadline 10. 4. 2018

Use the dataset hw6data that is available on the webpage of the course (**here**) and run

```
load("hw6data.RData")
set.seed(AAA)
ii <- sample(nrow(hw6data$Y), size=I<-20);
Y = hw6data$Y[ii,]; N = hw6data$n[ii,]
```

Entries Y_{ij}, $i = 1, \ldots, 20$, $j = 1, 2$ of table `Y` contain independent realisations of binomial experiments $Y_{ij} \sim \mathsf{Bi}(n_{ij}, p_{ij})$ where

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \psi_i + \tau\, \mathbb{I}\{j = 2\}$$

as in Example 30 in the course notes. Table `N` contains the sizes n_{ij}.

(i) Estimate parameter τ using the usual maximum likelihood method, the profile likelihood, and the conditional likelihood.

(ii) For each method from part (i) construct one confidence interval for τ, and evaluate it numerically.

(iii) For `N` as above, simulate a new table `Y` with $\psi_1, \ldots, \psi_{20}$ an independent random sample from $\mathcal{N}(0, 100)$, for $\tau = 1$. Find the estimates of τ from parts (i) and (ii). Repeat this procedure $B = 1000$ times, and report the average mean squared error of the point estimators, and the average coverage of the interval estimators. Which method is the best and why?

Hint: You can use the `R` function `mantelhaen.test`. But, you need to specify in your solution how this function is called with data `Y` and `N`, and how the estimates are obtained.

Homework 7 (12 p) - deadline 10. 4. 2018

You observe a random sample $\mathbf{Z}_1 = (Y_1, \mathbf{X}_1^\mathsf{T})^\mathsf{T}, \ldots, \mathbf{Z}_n = (Y_n, \mathbf{X}_n^\mathsf{T})^\mathsf{T}$, where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^\mathsf{T}$ and

$$Y_i = \boldsymbol{\beta}^\mathsf{T} \boldsymbol{X}_i + \varepsilon_i,$$

where $\mathsf{E}\left[\varepsilon_i \,|\, \boldsymbol{X}_i\right] = 0$ for $i = 1, \ldots, n$. Consider the following estimator of the unknown parameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\mathsf{T}$:

$$\widehat{\boldsymbol{\beta}}_n = \operatorname*{arg\,min}_{\mathbf{b} \in \mathbb{R}^p} \sum_{i=1}^{n} \left[Y_i - \mathbf{b}^\mathsf{T} \boldsymbol{X}_i\right]^4.$$

(i) Derive the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_n$.
(It is not necessary to check the regularity assumptions, but do not forget to show that the identified parameter is really $\boldsymbol{\beta}$.)

(ii) Construct a confidence set for the parameter $\boldsymbol{\beta}$.

(iii) Describe a test of the null hypothesis $H_0 : \beta_p = 0$ against the alternative $H_1 : \beta_p \neq 0$.

(iv) Suppose that you can assume that $\varepsilon_1, \ldots, \varepsilon_n$ are independent random variables such that ε_i is independent of $\boldsymbol{X}_i$ and the distribution of ε_1 is symmetric around zero. Derive the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_n$ in this more specific model and compare it with the asymptotic distribution of the least-squares estimator

$$\widetilde{\boldsymbol{\beta}}_n = \operatorname*{arg\,min}_{\mathbf{b} \in \mathbb{R}^p} \sum_{i=1}^{n} \left[Y_i - \mathbf{b}^\mathsf{T} \boldsymbol{X}_i\right]^2.$$

Homework 8 (10 p) - deadline 10. 4. 2018

The density of a random vector $\mathbf{X}$ in $\mathbb{R}^d$, $d \geq 1$, is said to be (spherically) symmetric if it takes the form $f(\mathbf{x}) \propto g(\|\mathbf{x}\|^2)$ for all $\mathbf{x} \in \mathbb{R}^d$, where $\propto$ means "up to a constant multiple", $\|\cdot\|$ is the Euclidean norm, and g is a univariate function. It can be shown that the distribution of $R = \|\mathbf{X}\|^2$ then has a density $f_R(r) \propto r^{d/2-1} g(r)$, for $r \geq 0$.

Consider the following distributions:

(i) Standard multivariate t-distribution with $\nu \geq 1$ degrees of freedom defined as

$$\mathbf{X} = \frac{\mathbf{Z}}{\sqrt{W/\nu}},$$

where $\mathbf{Z}$ has a standard d-variate normal distribution, $W \sim \chi^2_\nu$, and $\mathbf{Z}$ is independent of W. The density of $\mathbf{X}$ is generated by $g(t) = (1 + t/\nu)^{-(\nu+d)/2}$;

(ii) Uniform distribution on the unit ball, whose density is generated by $g(t) = \mathbb{I}\{t \in [0, 1]\}$;

(iii) Uniform distribution on the unit sphere $\{x \in \mathbb{R}^d \colon \|x\| = 1\}$.

Compute the expectation, and the variance of these distributions. In each case, find also the distribution of $\|X\|^2$. Does it belong to a known parametric family?

Homework 9 (12 p) - deadline 17. 4. 2018

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be two independent random samples from exponential distributions with means $1/\lambda$ and $1/\mu$, respectively. Find the distribution of

$$T_{n,m} = \frac{\overline{X}_n}{\overline{Y}_m}$$

and show that it depends on parameters λ and μ only through their ratio λ/μ. Suggest tests of the null hypothesis $H_0 : \lambda \leq 3\mu$ against the alternative $H_1 : \lambda > 3\mu$ based on

(i) the exact distribution of $T_{n,m}$,

(ii) the Monte Carlo version version of the exact test, and

(iii) the asymptotic distribution of $T_{n,m}$ under the assumption that $n \to \infty$, $m \to \infty$, and $n/m \to q \in (0, \infty)$ (q is not known).

On a small simulation study compare the performance of these three tests. In detail, describe the design of your simulation study, and report your findings in an appropriate way.

Homework 10 (10+10 p) - deadline 24. 4. 2018

Parts (i)–(iii) (10 p) of this homework are compulsory. See the requirements to get the course credit. Please send your R code (via e-mail) together with your solution. Before each resampling procedure set `set.seed(AAA)`. It must be clear from your solution how the resampling is done for each particular task.

Load the dataset pisa.csv (available **here**)

```
PISA = read.csv("pisa.csv",sep=";") # contains NA values
```

The dataset contains four variables observed in each world country — the average PISA[1] scores in reading (`Reading`) and in science (`Science`), and the estimated average IQ (`IQ`) and median age (`Age`) of the population of that country.

(i) Denote by ρ_R and ρ_S the correlation coefficients between the IQ, and the PISA scores in reading and science, respectively. Use bootstrap to approximate the distribution of an estimator of $\rho_R - \rho_S$. Devise a confidence interval for $\rho_R - \rho_S$, and test $H_0 : \rho_R \geq \rho_S$ against $H_1 : \rho_R < \rho_S$.

[1] Programme for International Student Assessment

(ii) Describe and perform a bootstrap test of the null hypothesis that the distribution of the PISA scores in science is normal (with unknown parameters).

(iii) Suppose that the joint distribution of `Age` and `IQ` is normal. Describe and perform a permutation test of the null hypothesis that the average IQ is independent of the median age in the population.

(iv) Perform a bootstrap test of the null hypothesis that the joint distribution of the PISA scores in science and reading is bivariate normal (with unknown parameters).

(v) The spatial median of a d-variate random vector $\boldsymbol{X}$ is the point $\boldsymbol{\theta}_{\boldsymbol{X}}$ that minimizes $\mathsf{E}\,\|\boldsymbol{X}-\boldsymbol{\theta}\|$ as a function of $\boldsymbol{\theta}$, for $\|\cdot\|$ the Euclidean norm. Estimate the spatial median $\boldsymbol{\theta}_{\boldsymbol{X}}$ of the bivariate random vector of the PISA scores, and derive some bootstrap-based confidence region for $\boldsymbol{\theta}_{\boldsymbol{X}}$. Compare this region with the confidence ellipse for $\boldsymbol{\theta}_{\boldsymbol{X}}$ based on the asymptotic normality result. Which region do you prefer and why? How would you construct a confidence region for the covariance matrix of the PISA scores?

Homework 11 (14 p) - deadline 15. 5. 2018

Use the dataset pisa.csv (available **here**) from Homework 10.

(i) Perform quantile regression to describe how the average IQ in a country (variable `IQ`) depends on the two PISA scores (variables `Science` and `Reading`) and the median age (`Age`). Consider appropriate transformations of the variables. Compare the results when modelling different conditional quantiles and interpret the differences. Produce at least one figure that visualises the results. Compare with the results obtained by the least squares method.

(ii) For the first-quartile-regression (0.25-quantile), test the hypothesis that the average IQ is not affected by the median age. Suggest, and perform a resampling test of this hypothesis, and compare the results with those using a test based on an asymptotic normality result. Report the p-values, and comment on possible differences between them.

Hint: `R` functions `anova.rq` and `summary.rq` from library `quantreg` may be of use. But, you need to specify in your solution how those functions are called with your data, and how the p-values are obtained.

Homework 12 (8 p) - deadline 15. 5. 2018

Consider the quantile regression model fitted to the Infant Birth Weight data as described **here**. Based on the results in Figure 1.11, answer the following questions:

(i) Provide a (rough) estimate of the conditional median, the conditional $\tau = 0.05$ quantile, and the conditional expectation, of the birth weight of a boy whose mother is unmarried, white non-smoker, who is 20 years old, has only elementary education, her first prenatal visit was in the first trimester of the pregnancy, and who gained 20 Lbs of weight. Interpret the possible differences in these three quantities.

(ii) Describe in detail how the curves in Figure 1.12 are plotted, based on the information in Figure 1.11. Provide an approximate formula for the curve in Figure 1.12 that corresponds to $\tau = 0.1$ quantile.

(iii) Describe in detail how Figure 1.13 is plotted, based on the information in Figure 1.11. Interpret, in your own words, the meaning of Figure 1.13, and the conclusions that can be drawn from it.

Homework 13 (15 p) - deadline 22. 5. 2018

This is a compulsory homework. See the requirements to get the course credit. Please send the R *code with your implementation of the EM-algorithm (via e-mail) together with your solution.*

Use the file defects2.RData (available **here**) and generate your data as

```
load('defects2.RData')
set.seed(AAA);
X <- sample(defects, size=100)
```

Variable X contains the number of defective products manufactured by 100 machines of the same kind during the same time interval. It is supposed that each machine is either properly aligned, or misaligned. A properly aligned machine produces no defects. If the machine is misaligned, then the number of defects produced by the machine follows a Poisson distribution (with an unknown parameter λ). Thus, we observe a random sample from the so-called zero-inflated Poisson distribution given by

$$\mathsf{P}(X_1 = k) = w\,\mathbb{I}\{k = 0\} + (1 - w)\,\frac{\lambda^k \mathrm{e}^{-\lambda}}{k!}, \qquad k = 0, 1, 2, \ldots,$$

where $\lambda > 0$ and $w \in (0, 1)$ are unknown parameters to be estimated.

(i) Describe in detail, and perform the EM-algorithm that finds the maximum likelihood estimator of the unknown parameters (w, λ).

(ii) Test the hypothesis that the proportion of the aligned machines is 0.1.

(iii) Test the hypothesis that the proportion of the aligned machines is zero.

Homework 14 (12 p) - deadline 22. 5. 2018

Consider the bivariate dataset Xo generated by

```
set.seed(AAA)
library(mvtnorm)
X = rmvnorm(n <- 1000,sigma=matrix(c(1,0.5,0.5,1), 2, 2)) # complete data
Xo <- X
qqq <- (X[,1]<=-0.5)
Xo[qqq,2] <- NA # replace some data by NA
```

(i) Read Sections 1 and 2 of the paper of Cohen (1955, available **here**). In your own words, describe the findings from that paper that are relevant to the analysis of our data. Please do not copy the derivations of Cohen; only describe which formulas from the paper are useful and why.

(ii) For your dataset Xo, estimate the unknown parameters (expected value and variance matrix of the bivariate normal distribution) using Cohen's method. Compare your results with those obtained by the methods considered at the exercise class (see the R scripts **here** and **here**). Comment on the results.

(iii) Is it possible to extend Cohen's method to the situation "3. Missing not at random" considered at the exercise class (**here**)? Will such an extension perform better than the EM-algorithm programmed in function binormal.EM from **this script**? Why?

Homework 15 (4 p) - deadline 22. 5. 2018

Find an example of a bivariate dataset with missing data, such that the sample variance matrix computed from this dataset using the available case analysis (see Example 61 in the course notes) is negative definite.

Homework 16 (10 p) - deadline 29. 5. 2018

Let $X_1, \ldots, X_n$ be a random sample from a distribution with the density f (with respect to the Lebesgue measure). Let x be a fixed point, and let the derivative of f be continuous at x. Let K be a differentiable kernel function. Consider the following estimator of $f'(x)$:

$$\widehat{f}'_n(x) = \frac{1}{n\,h_n^2} \sum_{i=1}^{n} K'\big(\tfrac{x-X_i}{h_n}\big).$$

Prove that $\widehat{f}'_n(x)$ is a (weakly) consistent estimator of $f'(x)$. Specify the assumptions on the bandwidth h_n and, if necessary, also further assumptions about the kernel function K.