NMST434 – Modern statistical methods (2016–17)

Homework assignments

General instructions

- The solution of the assignment should be delivered at **the beginning of the exercise class** (or at any lecture preceding the exercise class).
- **Hand written solution** are completely fine, but the solution should be written in a **readable** way.
- The language of the homework reports can be either **English** or **Czech/Slovak**.
- Generally only pdf-documents prepared using a suitable software for document preparation is acceptable for sending solutions of assignments via **email**. Exceptionally it is possible to send a photo of your handwritten solution. If so, be careful that the resulting image is not too dark. The solution should be readable after printing in a grayscale mode. Files sent via email should be named with your surname and the number of the homework, e.g. `novak-2.pdf`. Please, write `nmst434` in the subject of the email.
- 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 will 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 let stand for the number of your student identity card.

Homework 1 (14 p) - deadline 8. 3. 2017

Suppose we observe independent and identically distributed random variables $X_1, \ldots, X_n$ from a 'zero-inflated Poisson distribution', that is a mixture of a Poisson distribution and Dirac measure at the point zero so that

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

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

(i) Using the moment method write down the estimating equations for the unknown parameter $\boldsymbol{\theta} = (\lambda, w)$.

(ii) Derive the asymptotic distribution of the moment estimator $\widehat{\boldsymbol{\theta}}_n$.
The asymptotic variance matrix can be in a form of a multiplication of matrices, provided that it is clear how each of the matrices is calculated. You can also make use of the higher moments of the Poisson distribution that you find in reliable resources (i.e. you do not need to calculate them as a part of your solution).

(iii) Use the dataset defects that is available at the webpage of the course and generate your data as

```
load('defects.RData')
set.seed(AAA);
X <- sample(defects, size=5000)
```

Now X contains the number of defects produced by 5 000 machines of the same kind during the same time interval. It is supposed that each machine is either properly aligned or misaligned. If the machine is properly aligned then it 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). Based on you data, find a confidence interval for the probability that the machine is properly aligned.

Homework 2 (11 p) - deadline 15. 3. 2017

Suppose you observe independent identically distributed random variables $X_1, \ldots, X_n$ from the following discrete distribution

$$\mathsf{P}(X_1 = 0) = 1 - p - p^2, \quad \mathsf{P}(X_1 = -1) = \mathsf{P}(X_1 = 1) = \tfrac{p}{2}, \quad \mathsf{P}(X_1 = -2) = \mathsf{P}(X_1 = 2) = \tfrac{p^2}{2},$$

where $p \in \left(0, \frac{\sqrt{5}}{2} - \frac{1}{2}\right)$.

(i) Consider the moment estimator $\widetilde{p}_n$ of the parameter p and derive its asymptotic distribution.

(ii) Consider the maximum likelihood estimator $\widehat{p}_n$ of the parameter p and derive its asymptotic distribution.

(iii) Based on the above results suggest a test of the null hypothesis $H_0 : p = \frac{1}{2}$ against the alternative $H_1 : p \neq \frac{1}{2}$.

Homework 3 (13 p) - deadline 22. 3. 2017

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 has 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. Next 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 $\beta_0, \beta_1, \beta_2, \sigma^2$ are unknown. Further 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, 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 4 (10 p) - deadline 29. 3. 2017

Let $(Y_1, X_1)^\mathsf{T}, \ldots, (Y_n, X_n)^\mathsf{T}$ be independent random vectors such that the conditional distribution of Y_1 given X_1 is Poisson $\mathsf{Po}\big(\lambda(X_1)\big)$, where $\lambda(x) = \exp\{\beta_0 + \beta_1\, x\}$ and parameters β_0 and β_1 are unknown. Further suppose that the distribution of X_1 does not depend on parameters β_0 and β_1. Derive the expression for the profile log-likelihood of parameter β_1.

Let generate data in the following way:

```
set.seed(AAA);
n <- 50;
X <- runif(n);
beta0 <- 1;
beta1 <- 2;
Y <- rpois(n, lambda = exp(beta0 + beta1*X));
```

For generated data plot the profile log-likelihood for parameter β_1 and find the 95 %-confidence interval based on the likelihood ratio test. Compare this asymptotic confidence interval with the asymptotic confidence interval based on the Wald approach.

Hint. For evaluating the confidence interval numerically, the function `uniroot` *might be of interest.*

Homework 5 (5 p) - deadline 5. 4. 2017

Suppose we are in a situation for which Cochran–Mantel–Haenszel test was derived. Consider the special case $n_{i0} = n_{i1} = 1$ for each $i = 1, \ldots, I$. Introduce

$$N_{jk} = \sum_{i=1}^{I} \mathbb{I}\{Y_{i0} = j, Y_{i1} = k\}, \qquad j = 0, 1;\ k = 0, 1.$$

Show that then the test statistic $R_n^{(c)}$ of Cochran-Mantel-Haenszel test simplifies to

$$R_n^{(c)} = \frac{(N_{01} - N_{10})^2}{N_{01} + N_{10}},$$

which is known as McNemar's test.

Homework 6 (8 p) - deadline 5. 4. 2017

The table below summarises the results of a clinical study in 8 centers. The study compared two cream preparations, an active drug and a control, on their success in curing an infection.

Center	Treatment	Response	
		Success	Failure
1	Drug	11	25
	Control	10	27
2	Drug	16	4
	Control	22	10
3	Drug	14	5
	Control	7	12
4	Drug	2	14
	Control	1	16
5	Drug	6	11
	Control	0	12
6	Drug	1	10
	Control	0	10
7	Drug	1	4
	Control	1	8
8	Drug	4	2
	Control	6	1

Formulate a suitable model that assumes the common effect of the drug in the eight centers. With the help of the conditional likelihood estimate this common effect of the drug and find also a confidence interval for this effect. Interpret the results.

Hint. For finding the estimate and the confidence interval numerically, the functions `optimize` *and* `uniroot` *might be of interest.*

Homework 7 (12 p) - deadline 12. 4. 2017

Suppose you observe independent and identically distributed random vectors $\mathbf{Z}_1 = (Y_1, \mathbf{X}_1^\mathsf{T})^\mathsf{T}, \ldots, \mathbf{Z}_n = (Y_n, \mathbf{X}_n^\mathsf{T})^\mathsf{T}$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^\mathsf{T}$. Suppose that you can moreover assume that

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

where $\mathsf{E}\,[\varepsilon_i^3 \,|\, \mathbf{X}_i] = 0$ for $i = 1, \ldots, n$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\mathsf{T}$. Consider the following estimator of the parameter $\boldsymbol{\beta}$:

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

(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 moreover assume that $\varepsilon_1, \ldots, \varepsilon_n$ are independent random variables such that ε_i is independent of $\mathbf{X}_i$ 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-square estimator

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

Homework 8 (10 p) - deadline 19. 4. 2017

Use the dataset random-numbers that is available at the webpage of the course and generate your data as

```
set.seed(AAA);
load("random-numbers.RData");
X <- sample(random.numbers, size=100);
```

Now X contains 100 random numbers that are claimed (by a runner of a lottery) to have uniform distribution on the interval $(0, 1)$. Suggest a test of this null hypothesis that would be powerful against the alternative that the dataset were generated from a Beta distribution (different from a uniform distribution).

(i) Provide a p-value of the test computed by the Monte Carlo principle.

(ii) Compare the p-value calculated in (i) with a p-value computed by the asymptotic approximation.

Do not forget to explain how the p-values are computed.

The function `gamma` and related functions might be of interest.

Homework 9* (12 p) - deadline 26. 4. 2017

This is a compulsory homework. See the requirements to get the course credit. Do not forget to send the R-code (via email) together with your solution. Your code should also include setting the seed for the random generation (`set.seed(AAA)`) before each of the resampling method. Further, it should be clear from your solution how the resampling method is done for each particular task.

Use the dataset lq-en.RData from the exercise class. Let concentrate on the IQ of the students (variable `iq`).

(i) Recall that the skewness of the distribution is defined as $\gamma = \mathsf{E}\left(\frac{X_1-\mu}{\sigma}\right)^3$, where $\mu = \mathsf{E}\, X_1$ and $\sigma^2 = \mathsf{var}(X_1)$. With the help of bootstrap find the confidence interval for the skewness of the IQ of students.

(ii) Describe and perform a bootstrap test of the null hypothesis that the IQ of students is normally distributed (with unknown mean and variance). Compare the resulting p-value with a p-value of a standard test for this problem.

(iii) Describe and perform a permutation test of the null hypothesis that the distribution of IQ for boys is the same as the distribution of IQ for girls.

Homework 10 (12 p) - deadline 10. 5. 2017

Use the dataset foodexp that is available also at the webpage of the course. Use the quantile regression to describe how the food expenditure (variable `foodexp`) of households (in Belgian francs) is influenced by the income (variable `income`) of households (in Belgian francs). Think what could be used as a response. Think about appropriate transformations of the variables. Compare the results when modelling different conditional quantiles and interpret the (possible) differences. Produce at least one figure that visualises the results. Compare with the results obtained by the least square methods. Write the model in such a form so that one can also interpret the estimate of the intercept.

Homework 11* (15 p) - deadline 17. 5. 2017

This is a compulsory homework. See the requirements to get the course credit. Do not forget to send the R-code (via email) together with your solution.

Use the file defect2s.RData and generate your data as.

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

Now `X` contains the number of defects produced by 100 machines of the same kind during the same time interval. It is supposed that each machine is either properly aligned or misaligned. If the machine is properly aligned then it 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).

(i) Describe in detail the EM-algorithm that finds the maximum likelihood estimator of the unknown parameters (w, λ) of the zero inflated Poisson model (see Homework 1).

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

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

Homework 12 (12 p) - deadline 24. 5. 2017

Suppose you observe independent and identically distributed random vectors $(X_1, Y_1)^\mathsf{T}, \ldots, (X_n, Y_n)^\mathsf{T}$ from a (general) bivariate normal distribution. Suppose that X_i is always observed, but some of Y_i are missing. Describe in detail the EM-algorithm that finds the maximum likelihood estimator of the unknown parameters.

Homework 13 (10 p) - deadline 2. 6. 2017

Use the dataset failures.RData available at the webpage of the course and generate your dataset as follows.

```
set.seed(AAA);
load("failures.RData");
DATA <- failures[sample(1:length(failures), size=200)];
```

The dataset `DATA` now contains failure times of some devices. Estimate the unknown density failure times provided you know that the true density will be 'close' to (but not exactly in) the family of densities

$$f(x) = \lambda^2 \, x \, \exp\{-\lambda \, x\} \, \mathbb{I}\{x > 0\},$$

with λ being an unknown parameter.

(i) Give the estimate of the density at the points 1 and 2.

(ii) Provide a figure of the $\widehat{f}_n(x)$ together with a histogram.

In your solution do not forget to explain how you choose the bandwidth.

Homework 14 (10 p) - deadline 2. 6. 2017

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

$$\widehat{f}_n''(x) = \frac{1}{n\, h_n^3} \sum_{i=1}^{n} K''\big(\tfrac{x-X_i}{h_n}\big).$$

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.