NMST434 – Modern statistical methods (2019–20)

Homework assignments

General instructions

To obtain the course credit for the exercise class, you need at least 100 points out of 140 possible for the solutions of the homework assignments. At the same time, you must successfully solve the two indicated compulsory tasks (bootstrap and EM algorithm).

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 to 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, include this number at the beginning of your solution of the assignment. If you do not have this number, use your date of birth in the format `YYYYMMDD`.

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.

Unless stated otherwise, it is acceptable that mathematical software (`Wolfram|Alpha`, `R`, `Mathematica` etc.) is used for the solution of partial problems (for instance, for computation of complicated integrals and sums). But, it must always be clear from the report how and why such a computation was performed, what was its input and output, and what is its relevance to the problem.

If not stated otherwise use 5 % as the level (prescribed probability of type I error) of the tests and 95 % as the coverage of the confidence intervals.

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

Homework 1 (8 p) - deadline 28. 2. 2020

Let $\{x_n\}_{n=1}^{\infty}$ be a sequence of constants, and $\{r_n\}_{n=1}^{\infty}$ a sequence of positive constants. Recall that $x_n = O(1/r_n)$ if the sequence $\{r_n\, x_n\}_{n=1}^{\infty}$ is bounded.

Prove or find a counter-example:

(i) For a sequence of positive random variables $X_n = o_{\mathsf{P}}(1/\sqrt{n})$ if and only if $\sqrt{n}X_n \xrightarrow[n\to\infty]{d} 0$.

(ii) For a sequence of positive random variables $X_n = O_{\mathsf{P}}(1/\sqrt{n})$ if and only if $\sqrt{n}X_n \xrightarrow[n\to\infty]{d} X$ for some random variable X.

(iii) $\left(1 + O_{\mathsf{P}}\left(1/\sqrt{n}\right)\right)^{-1} = 1 + O_{\mathsf{P}}\left(1/\sqrt{n}\right)$.

(iv) $\left(2\, O_{\mathsf{P}}\left(1/\sqrt{n}\right)\right)^{\alpha} O(1) = O_{\mathsf{P}}\left(n^{-\alpha/2}\right)$ for all $\alpha > 0$.

(v) $\exp\left(O_{\mathsf{P}}\left(1/\sqrt{n}\right)\right) = O_{\mathsf{P}}\left(\exp\left(1/\sqrt{n}\right)\right)$.

Homework 2 (8 p) - deadline 28. 2. 2020

We observe independent and identically distributed random variables $X_1, \ldots, X_n$ from a mixture of a Poisson distribution and a Dirac measure at zero, that is

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

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

(i) Use the method of moments to derive an estimator $\widehat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta} = (\lambda, w)^{\mathsf{T}}$.

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

(iii) Use the dataset defects available at the **website** of the course and generate your data using

```
load("defects.RData")
set.seed(AAA)
X = sample(defects, size=1000)
```

Vector X contains the number of defects produced by 1 000 machines of the same kind within one week. 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 a randomly chosen machine is properly aligned.

Homework 3 (10 p) - deadline 5. 3. 2020

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

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

where p and q are unknown quantities that make this expression a probability distribution.

(i) Find a moment estimator of the unknown parameters and derive its asymptotic distribution.

(ii) Derive the maximum likelihood estimator of the unknown parameters and derive its asymptotic distribution.

(iii) Based on results from parts (i) and (ii), suggest two tests of the null hypothesis $H_0 : \mathsf{P}(X_1 = 1) \leq \mathsf{P}(X_1 = 2)$ against the alternative $H_1 : \mathsf{P}(X_1 = 1) > \mathsf{P}(X_1 = 2)$ based on the moment estimator and the maximum likelihood, respectively. Which test do you prefer, and why?

Homework 4 (15 p) - deadline 12. 3. 2020

In Homework 3 we considered a random sample $X_1, \ldots, X_n$ from a special multinomial distribution $X \sim \mathrm{Mult}_K\left(1, \boldsymbol{x}, \mathbf{p}\left(\boldsymbol{\theta}\right)\right)$ defined by

$$\mathsf{P}\left(X = [\boldsymbol{x}]_k\right) = \left[\mathbf{p}\left(\boldsymbol{\theta}\right)\right]_k \quad \text{for } k = 1, \ldots, K,$$

where $[\boldsymbol{x}]_k$ is the k-th coordinate of vector $\boldsymbol{x}$. The unknown vector parameter is $\boldsymbol{\theta} \in \mathbb{R}^p$, and the function $\mathbf{p}\colon \mathbb{R}^p \to (0,1)^K$ is known. We saw that in Homework 3, for $\boldsymbol{\theta} = (p,q)^\mathsf{T}$ and $\boldsymbol{x} = (-2,-1,0,1,2)^\mathsf{T}$, the moment estimator based on the second moment $m_2 = \sum_{i=1}^n X_i^2/n$ and the fourth moment $m_4 = \sum_{i=1}^n X_i^4/n$ was

$$\widetilde{\boldsymbol{\theta}}_n = \left(\frac{1}{6}\left(4m_2 - m_4\right), \frac{m_4 - m_2}{4\left(4m_2 - m_4\right)}\right)^\mathsf{T}.$$

The maximum likelihood estimator took the form

$$\widehat{\boldsymbol{\theta}}_n = \left(\frac{n_2 + n_4}{2\,n}, \frac{n_1 + n_5}{n_2 + n_4}\right)^\mathsf{T}$$

for $n_k = \sum_{i=1}^n \mathbb{I}\left\{X_i = [\boldsymbol{x}]_k\right\}$.

(i) Show that the two estimators above are indeed equal.

(ii) Is this a coincidence? If possible, find an example of a properly parametrized multinomial distribution where the moment estimator $\widetilde{\boldsymbol{\theta}}_n$ does not equal the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_n$.

(iii) Formulate sufficient conditions for function $\mathbf{p}$ under which the two estimators $\widetilde{\boldsymbol{\theta}}_n$ and $\widehat{\boldsymbol{\theta}}_n$ are the same for general multinomial distributions. Prove your claim.

Homework 5 (9 p) - deadline 19. 3. 2020

Read **this** excerpt from Zvára's *Regrese* on the general derivation of Breusch-Pagan's test. Explain in detail:

(i) How the part of the Fisher information matrix (9.9) that corresponds to $\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}'$ is computed.

(ii) How the final formula (9.11) is obtained from the two matrices on page 124.

(iii) Why S_f can be interpreted as the regression sum of squares in certain models (page 125).

(iv) Why $S_{f,\text{Koenker}}$ can be written in terms of a sample correlation coefficient (page 125 below).

Homework 6 (6 p + bonus points) - deadline 26. 3. 2020

Let $\mathcal{F} = \{f(\mathbf{x};\theta) : \theta \in \mathbb{R}\}$ be a system of densities that is regular in the sense of [**R0**]–[**R6**] from Section 2.1 in the course notes. For a smooth, strictly increasing function $g \colon \mathbb{R} \to \mathbb{R}$ denote by $\psi = g(\theta)$ the transformed parameter of this family. We want to test $H_0 \colon \theta = \theta_0$ against $H_1 \colon \theta \neq \theta_0$ for θ_0 given, which is equivalent with $H_0^* \colon \psi = g(\theta_0)$ against $H_1^* \colon \psi \neq g(\theta_0)$ if the family $\mathcal{F}$ is considered as parametrized by ψ.

We know that the likelihood ratio test is invariant under reparametrization, i.e. the two likelihood ratio tests derived for θ and ψ always give the same result (see formula (29) in the course notes).

(i) Are the Wald test and the Rao score test invariant under such reparametrizations? Prove or find a counterexample.

(ii) Can you state an analogous result for p-dimensional parameters? How does the Fisher information matrix of the transformed parameters relate to the original Fisher information matrix?

Note: Part (ii) may be difficult, and is for bonus points. An original proof will be appreciated, but also correct interpretation of results found in the literature is completely fine.

Homework 7 (11 p) - deadline 26. 3. 2020

Let $(Y_1, X_1, Z_1)^\mathsf{T}, \dots, (Y_n, X_n, Z_n)^\mathsf{T}$ 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));
```

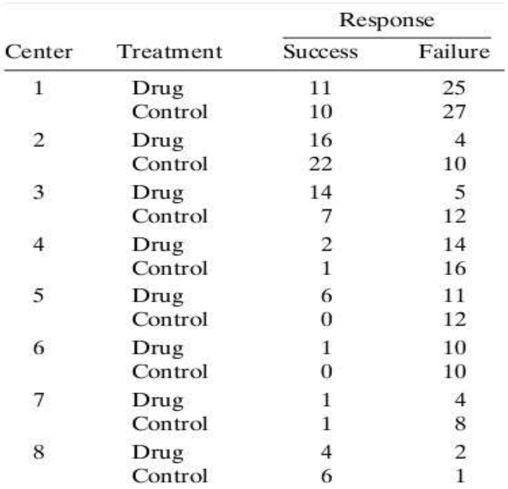
For the generated dataset, plot the profile log-likelihood of β and γ. Find the maximum likelihood estimate of α, β, and γ, and visualise the 95 %-confidence region for β and γ based on the likelihood ratio test.

(iii) Find the asymptotic confidence ellipse for β and γ based on 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 numerical optimization and visualisation, R functions `optim`, `ellipse` (package `car`), and `contour` might be of interest. An example of a call of function `ellipse` can be found here.

Homework 8 (8 p) - deadline 2. 4. 2020

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.



		Response	
Center	Treatment	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 a confidence interval for this effect. Interpret the results.

Homework 9 (7 p) - deadline 2. 4. 2020

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

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

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

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

This test statistic is known from **McNemar's test** of marginal homogeneity of 2 × 2 contingency tables.

(ii) Explain how does the general model under which the Cochran-Mantel-Haenszel test was derived relate to the model under which McNemar's test is introduced. What is the interpretation of the contingency table $\{N_{jk} \colon j = 0, 1;\ k = 0, 1\}$ used in McNemar's test?

Homework 10 (10 p) - deadline 9. 4. 2020

For $\varepsilon \in [0, 1/2)$ and $\nu \geq 1$ an integer, denote by $F_\varepsilon(x) = (1 - \varepsilon)\Phi(x) + \varepsilon F_\nu(x)$ the cumulative distribution function (cdf) of a mixture of the standard normal distribution (with cdf Φ) and the t-distribution with ν degrees of freedom (with cdf F_ν). Consider three estimators of location of F_ε — sample mean, sample median, and Huber's M-estimator.

(i) Find the asymptotic variance of the sample mean and the sample median in this model.

(ii) Plot the asymptotic variance of the Huber estimator as a function of its tuning constant k for various choices of ν and ε. Compare with the asymptotic variances from part (i).

(iii) Interpret your findings from parts (i) and (ii). Which location estimator is the best, and under which conditions? Describe a real-life scenario that justifies the considered model (i.e. give an example of a dataset that could be modelled using F_ε). How would your conclusions be affected if, instead of a standard normal distribution, a general normal distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ was considered?

Hint: See here.

Homework 11 (12 p) - deadline 16. 4. 2020

You observe a random sample $\boldsymbol{Z}_1 = (Y_1, \boldsymbol{X}_1^\top)^\top, \ldots, \boldsymbol{Z}_n = (Y_n, \boldsymbol{X}_n^\top)^\top$, where $\boldsymbol{X}_i \in \mathbb{R}^p$ and

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

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

$$\widehat{\boldsymbol{\beta}}_n = \underset{\mathbf{b} \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^n \left(Y_i - \mathbf{b}^\top \boldsymbol{X}_i\right)^4. \tag{1}$$

(i) Derive the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_n$. Does this M-estimator identify $\boldsymbol{\beta}$?
(It is not necessary to check the regularity assumptions.)

(ii) Suppose that $\varepsilon_1, \ldots, \varepsilon_n$ are independent, ε_i is independent of $\boldsymbol{X}_i$, and the distribution of ε_1 is symmetric around zero. Compare the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_n$ in this situation with the asymptotic distribution of the least squares estimator

$$\widetilde{\boldsymbol{\beta}}_n = \underset{\mathbf{b} \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^n \left(Y_i - \mathbf{b}^\top \boldsymbol{X}_i\right)^2. \tag{2}$$

(iii) If the errors in part (ii) are normal, which of the estimators $\widetilde{\boldsymbol{\beta}}_n$ and $\widehat{\boldsymbol{\beta}}_n$ has smaller asymptotic variance? In a situation where $\boldsymbol{\beta}$ is identified, give examples of distributions of errors such that the estimator $\widehat{\boldsymbol{\beta}}_n$ is preferable to $\widetilde{\boldsymbol{\beta}}_n$.

(iv) Do you see a general pattern in the results from part (iii)? For which distributions does it appear that higher exponents in equations (1) and (2) yield more efficient estimators? Compare with the asymptotic variance of the LAD estimator derived in Section 4.3.2 in the course notes.

Homework 12 (6 p) - deadline 16. 4. 2020

Consider a random sample $(Y_1, X_1, Z_1)^\top, \ldots, (Y_n, X_n, Z_n)^\top$ that satisfies

$$Y_i = \alpha + \beta X_i + \gamma Z_i + \varepsilon_i,$$

where $\alpha, \beta, \gamma \in \mathbb{R}$, and $\varepsilon_1, \ldots, \varepsilon_n$ are independent. In the least squares estimator of the regression coefficients, we ignore the contribution of variables Z_i and define

$$\left(\widehat{\alpha}_n, \widehat{\beta}_n\right)^\top = \underset{(\mathrm{a},\mathrm{b})^\top \in \mathbb{R}^2}{\arg\min} \sum_{i=1}^n \left(Y_i - \mathrm{a} - \mathrm{b} X_i\right)^2.$$

(i) Under which conditions do both M-estimators $\left(\widehat{\alpha}_n, \widehat{\beta}_n\right)^\top$ identify the vector $(\alpha, \beta)^\top$?

(ii) Under which conditions does $\widehat{\alpha}_n$ identify α? Under which conditions does $\widehat{\beta}_n$ identify β?

Homework 13 (8 p) - deadline 23. 4. 2020

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

(i) Give 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 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 14 (15 p) - deadline 7. 5. 2020

EM-algorithm (Semi-supervised clustering)

This homework is compulsory; see the requirements to get the course credit. Please send your R code (via e-mail) together with your report. Because of compatibility issues, please make sure that you use R of version at least 3.6.0. You can check the version of your R installation by running `R.Version()$version.string`.

Consider a random sample $(Y_1, \mathbf{Z}_1), \ldots, (Y_{n+m}, \mathbf{Z}_{n+m})$ that follows the same model as a generic pair $(Y, \mathbf{Z})$. Vector $\mathbf{Z} = (Z_1, Z_2, Z_3)^{\mathsf{T}} \in \{0,1\}^3$ is a vector of group indicators satisfying

$$Z_g = \begin{cases} 1, & \text{if } \mathbf{Y} \text{ belongs to group } g, \\ 0, & \text{otherwise,} \end{cases} \qquad \text{for all } g = 1, 2, 3.$$

We suppose that $\mathsf{P}(Z_g = 1) = \pi_g$, where $\pi_1 + \pi_2 + \pi_3 = 1, 0 < \pi_g < 1$ for all $g = 1, 2, 3$. The distribution of Y given that is comes from group g is normal, i. e.

$$Y | Z_g = 1 \quad \sim \quad \mathsf{N}\left(\mu_g, \sigma_g^2\right),$$

where $\mu_g \in \mathbb{R}$ and $\sigma_g > 0$ for each $g = 1, 2, 3$. Denote

$$\boldsymbol{\theta} = \left(\pi_1, \pi_2, \pi_3, \mu_1, \mu_2, \mu_3, \sigma_1^2, \sigma_2^2, \sigma_3^2\right)^{\mathsf{T}}$$

the set of unknown parameters.

Consider the following situation. All $Y_1, \ldots, Y_{n+m}$ are observed, however, only the first n indicators $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are available to you. The rest of the indicators $\mathbf{Z}_{n+1}, \ldots, \mathbf{Z}_{n+m}$ remain **unobserved**.

(i) Write down the explicit formula for the complete log-likelihood $\ell_C(\boldsymbol{\theta})$.

(ii) Derive and describe in detail the EM-algorithm used for the estimation of $\boldsymbol{\theta}$.

(iii) Think about the two special cases $n = 0$ and $m = 0$, respectively. How does the EM-algorithm behave in these situations?

Download the `wine` dataset into R environment. This dataset contains 6 physical and chemical measurements on 178 wine samples. There are three types of wine: Barolo, Grignolino and Barbera, and the type is known for each of the samples. You select a subset of $n = 50$ known type indicators using the following commands:

```
load("wine.RData");
set.seed(AAA);
N <- dim(wine)[1]
n <- 50
m <- N - n
knowntypes <- sort(sample(1:N, n))
unknowntypes <- setdiff(1:N, knowntypes)
```

where `AAA` stands for your student ID number. Our task is to predict the type of wine ($\mathbf{Z}$) for the observations from `unknowntypes`, based on the available information (Y) given in one variable out of the following six: ’Flavanoids’, ’Color_Intensity’, ’Alcalinity_of_ash’, ’pH’, ’Magnesium’ and ’Alcohol’. You are allowed to chose your variable Y, or run the algorithm for several variables separately and compare the final results.

(iv) Implement the EM-algorithm according to your theoretical derivations. Beware of numerical issues.

(v) Comment on the methods you use to initiate and terminate the algorithm.

(vi) Report the estimated parameters and visualize the resulting distribution.

(vii) Based on the estimated unobserved indicators classify observations $i = n+1, \ldots, n+m$ into clusters with the highest individual probability. Compare the classification based on your model with the true type of the wine. How successful is it? Visualize the difference between the true and estimated types. Try to add some measure of uncertainty.

(viii) Try different values of n, including $n = 0$ and $n = 178$. Do the estimated normal distributions substantially differ? What are the benefits of knowing at least some n indicators? Is the clustering without any known wine type successful?

(ix) BONUS: Implement an analogous EM algorithm based on a bivariate (or other multivariate) vector $\mathbf{Y}$ chosen from the additional information available, assuming that $\mathbf{Y}|Z_g = 1$ follows a multivariate normal distribution for each $g = 1, 2, 3$. Compare the results of the classification task with those above. Comment on your findings.

Homework 15 (12 p) - deadline 14. 5. 2020

Use the dataset lottery and generate your data as

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

Variable `X` contains 100 random numbers that a runner of a lottery claims to be generated from the uniform distribution on the interval $(0, 1)$. Suggest tests of this null hypothesis that would be powerful against the alternative that the dataset was generated from a beta distribution (different from a uniform distribution).

(i) Describe and perform a test based on some estimators of the parameters in beta distribution. Use both an asymptotic test, and its Monte Carlo version. Report the two p-values and comment on possible differences.

(ii) Similarly as in part (i), perform a suitable nonparametric test of uniformity. Compare its asymptotic and Monte Carlo version.

(iii) Which of the four considered tests do you prefer and why? If possible, support your claims by appropriate numerical results.

Hint: See Example 8 in the course notes.

Homework 16 (12 p) - deadline 21. 5. 2020

Bootstrap

This homework is compulsory; see the requirements to get the course credit. Please send your R code (via e-mail) together with your report. Because of compatibility issues, please make sure that you use R of version at least 3.6.0. You can check the version of your R installation by running R.Version()$version.string.

Use the **diabetes** dataset containing several covariates measured on diabetes patients. In our analysis, we will focus on three variables — `Glucose`, `Insulin`, and `BMI`. Note that for these variables, missing data appear to be indicated by zeros.

Do not forget to fix the random seed (`set.seed(AAA)`) before performing your analysis in each part (i)–(iv) below.

(i) At first, focus on the variable `Glucose`. We are interested in describing its scale to see how spread out the glucose levels are among the diabetes patients. We are interested in the interquartile range (IQR) rather than in the standard deviation.

(a) Give a point estimate of the IQR.

(b) Use bootstrap to estimate the mean squared error (MSE) of your estimated IQR.

(c) Use bootstrap to calculate a confidence interval for the true IQR.

(ii) From past experience, we can assume that the true spread of the glucose level among diabetes patients is approximately equal to 50. Using nonparametric bootstrap, perform a test of the hypothesis

$$H_0 : \text{IQR} = 50,$$
$$H_1 : \text{IQR} \neq 50.$$

It is not acceptable to use the confidence interval for IQR obtained from part (i) to perform this test. Explain why.

(iii) Now, focus on the variable `Insulin` denoting the insulin level of each patient. We are interested in testing whether this variable follows a gamma distribution $\text{Gamma}(\alpha, \beta)$ for some $\alpha, \beta > 0$. Formulate a test procedure and give the final conclusion.

(iv) We suspect that the insulin levels (`Insulin`) are positively correlated with the patients' body mass index (`BMI`). Formulate and perform a permutation test of this hypothesis. Report your results.

Homework 17 (10 p) - deadline 28. 5. 2020

Let $X_1, \ldots, X_n$ be a random sample from a distribution with density f (with respect to the Lebesgue measure on $\mathbb{R}$). Let $x \in \mathbb{R}$ be fixed and let the second derivative of f be continuous at x. The kernel K is twice differentiable everywhere. 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.