NMST424 – Mathematical Statistics 3 (2022–23)

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 compulsory task on EM algorithm.
- Solutions can be delivered during the class on Friday or submitted in the Moodle system as **one pdf-file**. In the latter case, please, take care that the names of the files are in the form `yourfamilyname_hw1.pdf`. If R is used to calculate the results, please submit also `R` source file (and all the other files) that are needed to run your code.
- 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 (13 p) - deadline: 24. 2. 2023 at the class (or 25. 2. 2023 at 12:00 in moodle)

We observe independent and identically distributed positive random variables $X_1, \ldots, X_n$ and we are interested in the coefficient of variation, i.e.

$$\theta_X = \frac{\sqrt{\mathsf{var}\, X_1}}{\mathsf{E}\, X_1}.$$

Suggest an estimator $\widehat{\theta}_n$ of the parameter θ_X. Use the Δ-theorem to derive the asymptotic linear approximation for $\widehat{\theta}_n$. Derive the (asymptotic) confidence interval for θ_X.

Homework 2 (16 p) - deadline: 3. 3. 2023 at the class (or 4. 3. 2023 at 12:00 in moodle)

We observe independent and identically distributed random variables $X_1, \ldots, X_n$ from the following (truncated Poisson distribution)

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

where $\lambda > 0$ is an unknown parameter to be estimated.

(i) Use the method of moments to derive an estimator $\widehat{\lambda}_n$ of λ.

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

(iii) Use the dataset accidents available either at this link or in the assignment of the moodle website

```
load("accidents.RData")
set.seed(AAA)
X = sample(accidents, size=634)
```

Assume that a given government has an unknown number of police cars. Nevertheless from some records you were able to identify police cars that have at least one accident in the last year. You have found that the number of such police (with at least one accident in the last year) cars is 634 and your dataset X contains the number of accidents for each of these cars. Assume that the number of accidents of a randomly chosen police car follows a Poisson distribution with the unknown parameter λ. Based on your dataset (X) calculate the confidence interval for λ. Estimate also the total number of police cars.

Hint. For solving a nonlinear equation numerically the function uniroot might be of interest.

Homework 3 (14 p) - deadline: 17. 3. 2023 at the class (or 18. 3. 2023 at 12:00 in moodle)

Let $(Y_1, X_1)^\mathsf{T}, \ldots, (Y_n, X_n)^\mathsf{T}$ be independent and identically distributed 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. Note that it is not sufficient to give only the numerically calculated value of the Wald confidence interval but you should give also the formula for Wald confidence interval and explain how you can calculate each quantity in this formula in this specific situation).

Hint. For evaluating the confidence interval numerically, the function `uniroot` might be of interest. To maximize the profile likelihood you can use the function `optimize`. To find the standard maximum likelihood estimate (and its estimate of asymptotic variance) you can use `glm(Y X, family="poisson")`.

Homework 4 (15 p) - deadline: 31. 3. 2023 at the class (or 1. 4. 2023 at 12:00 in moodle)

Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be two independent random samples from the Poisson distributions. Let λ_X be the true value of the parameter for the first sample and λ_Y for the second sample. Note that $\mathbf{S} = (S_1, S_2)^\mathsf{T} = \left(\sum_{i=1}^{n_1} X_i, \sum_{i=1}^{n_2} Y_i\right)^\mathsf{T}$ is a sufficient statistic for the parameter $\boldsymbol{\theta} = (\lambda_X, \lambda_Y)^\mathsf{T}$. Derive the conditional distribution of S_1 given $S_1 + S_2$. Use this result to find an 'exact' test of

$$H_0 : \lambda_X = \lambda_Y, \qquad H_1 : \lambda_X \neq \lambda_Y. \tag{1}$$

Further derive an 'exact' confidence interval for the ratio $\frac{\lambda_X}{\lambda_Y}$.

Now use the derived test of (1) and the confidence interval on the data that are generated in the following way

```
set.seed(AAA);
n1 <- sample(20:50, size=1);
n2 <- sample(20:50, size=1);
X <- rpois(n1, lambda=1);
Y <- rpois(n2, lambda=1+runif(1));
```

Now `X` contains the realizations of $X_1, \ldots, X_{n_1}$ and `Y` of $Y_1, \ldots, Y_{n_2}$.

Compare the results of testing with some standard test that can be used in this situation.

Homework 5 (15 p) - deadline: 31. 3. 2023 at the class (or 8. 4. 2023 at 12:00 in moodle)

Use the dataset `DATA` from **this file** that is available also at the **moodle website** of the course and generate your data using

```
set.seed(AAA);
load("match_paired_study.RData");
```

```
n <- 200;
iii <- sample(unique(DATA$pair), n);
DATA <- droplevels(DATA[is.element(DATA$pair, iii),]);
```

The dataset `DATA` now summarizes the results of a matched pair study consisting of 400 volunteers matched into 200 pairs based on the characteristics that may be relevant for a treatment of a given infection. The study compared two cream preparations, an active drug (`treatment = 1`) and a control (`treatment = 0`), on their success (`outcome = 1`) in curing an infection.

(i) Formulate a suitable model that assumes the common effect of the drug in the pairs.

(ii) Using the results from the exercise class write down the conditional likelihood and use it to estimate the common effect of the drug. Find a confidence interval for this effect. Interpret the results.

(iii) Using the following command calculate the standard profile likelihood confidence interval for the parameter of interest and compare the results.

```
print(confint(glm(outcome~pair+treat, data=DATA, family=binomial), parm="treat"))
```

Do not forget to submit the R-code as a part of your solution.

Homework 6 (15 p) - deadline 14. 4. 2023 at the class (or 15. 4. 2023 at 12:00 in moodle)

Suppose you observe independent and identically distributed random vectors $\binom{Y_1}{\boldsymbol{X}_1}, \ldots, \binom{Y_n}{\boldsymbol{X}_n}$ and define the estimator

$$\widehat{\boldsymbol{\beta}}_n = \arg\min_{\mathbf{b}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \left[- Y_i \, \boldsymbol{X}_i^\mathsf{T}\mathbf{b} + 2\log\left(1+\exp\{\boldsymbol{X}_i^\mathsf{T}\,\mathbf{b}\}\right)\right].$$

Further suppose that there exists $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ such that

$$\mathsf{E}\left[Y_i|\boldsymbol{X}_i\right] = \frac{2\,\exp\{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}_0\}}{1+\exp\{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}_0\}}. \tag{2}$$

(i) Find the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_n$.

(ii) Find an estimator of the asymptotic variance of $\widehat{\boldsymbol{\beta}}_n$.

(iii) Show that the estimator $\widehat{\boldsymbol{\beta}}_n$ corresponds to the maximum likelihood estimator when the conditional distribution of Y_i given $\boldsymbol{X}_i$ is binomial with parameters 2 and $\pi(\boldsymbol{X}_i, \boldsymbol{\beta})$, where $\pi(\boldsymbol{X}_i, \boldsymbol{\beta})$ is the probability of success given by $\pi(\boldsymbol{X}_i, \boldsymbol{\beta}) = \frac{\exp\{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}\}}$ and the distribution of $\boldsymbol{X}_i$ does not depend on $\boldsymbol{\beta}$.

(iv) Suggest a model in which (2) holds and $\mathsf{P}(Y_i \in \{0,1,2\}) = 1$, but which is different from the model in (iii).

Homework 7 (14 p) - deadline 21. 4. 2023 at the class (or 22. 4. 2023 at 12:00 in moodle)

Let $X_1, \ldots, X_n$ be a random sample from a distribution F. Consider the estimator defined as

$$\widehat{\theta}_n = \arg\min_{t\in\mathbb{R}} \sum_{i=1}^{n} \log\left(1+(X_i-t)^2\right).$$

Derive the asymptotic distributions of $\widehat{\theta}_n$ (under appropriate regularity assumptions that you do not need to check).

Suppose that the distribution F is a Student t-distribution with ν degrees of freedom. Based on the asymptotic variances compare the estimator $\widehat{\theta}_n$ with the sample mean $\overline{X}_n$ and the sample median $\widetilde{m}_n$. Plot the asymptotic variances of these three estimators as the function of ν. Discuss which estimator is the most suitable for a given degrees of freedom.

Hint. For calculating the integrals numerically the function `integrate` *might be of interest.*

nal 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.

Homework 8 (14 p) - deadline 5. 5. 2023 at the class (or 6. 5. 2023 at 12:00 in moodle)

For the dataset foodexp.RData that is available **here** use quantile regression to describe how the food expenditure (variable `foodexp`) of households is influenced by their income (variable `income`, both variables in Belgian francs).

(i) Find an appropriate quantile regression model. Specify the linear part (predictor) of the model so that also the intercept is interpretable.

(ii) Interpret the estimates of the regression parameters (both intercepts as well as slopes).

(iii) Compare the results when modelling different conditional quantiles and interpret differences.

(iv) Visualise the quantile regression fits, and compare them with the results obtained by the least squares method.

Homework 9 (16 p + 8 p) - deadline 12. 5. 2023 at the class (or 16. 5. 2023 at 23:59 in moodle)

THIS HOMEWORK IS COMPULSORY; SEE THE REQUIREMENTS TO GET THE COURSE CREDIT. DO NOT FORGET TO SUBMIT R-CODE AS A PART OF YOUR SOLUTION.

Suppose that you observe independent random variables $Y_1, \ldots, Y_n$, such that

$$\mathsf{P}(Y_i = k) = \sum_{j=1}^{G} w_j \binom{n_i}{k} p_j^k (1-p_j)^{n_i-k}, \qquad k \in \{0, \ldots, n_i\},$$

where $0 < p_1 < p_2 < \ldots p_G < 1$ and $w_j \in (0,1)$ such that $\sum_{j=1}^{G} w_j = 1$. I.e. Y_i is a mixture of G binomial distributions.

(i) Describe in detail how the EM-algorithm is used to calculate the unknown parameters $w_1, \ldots, w_G$ and $p_1, \ldots, p_G$. I.e. introduce the complete likelihood, describe the E-step and M-step. Suggest initial estimators of the parameters.

Use the dataset election2018.RData **available here** (as well as in the moodle modul of this homework) and generate your data using

```
load(file="election2018.RData");
set.seed(AAA)
DATA <- election[sample(nrow(election))[1:3000],];
```

`DATA` now gives the results of elections in 3 000 electoral districts in 3 regions.

(ii) Implement the EM-algorithm described above in R software where Y_i is the number of votes for one of the party you choose and n_i be `total.votes`. Consider $G = 3$. Comment on the way how you choose the stopping rule for your algorithm.

(iii) Use the results to divide the electoral districts into three clusters. Compare these clusters with the regions of the electoral districts. How successfully one would be in predicting regions using the election results?

Extra task (for extra 8 points). Use the election results of all parties (at the same time) and fit the mixture of multinomial distributions. It is sufficient to describe the EM-algorithm only very briefly (explaining only the differences to fiting mixture of binomial distributions). Again use this results to create clusters of electoral districts and compare them with regions of the electoral districts.

Homework 10 (8 p) - deadline 19. 5. 2023 at the class (or 20. 5. 2023 at 12:00 in moodle

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 is negative definite.