MAD 2020-21, Assignment 3

Bulat Ibragimov, Kim Steenstrup Pedersen

hand in until: 13.12.2021 at 23:59

General comments: The assignments in MAD must be completed and written individually. You are allowed (and encouraged) to discuss the exercises in small groups. If you do so, you are required to list your group partners in the submission. The report must be written completely by yourself. In order to pass the assignment, you will need to get at least 40% of the available points. The data needed for the assignment can be found in the assignment folder that you download from Absalon.

Submission instructions: Submit your report as a PDF, not zipped up with the rest. Please add your source code to the submission, both as executable files and as part of your report in appendix. To include it in your report, you can use the `lstlisting` environment in LaTeX, or you can include a "print to pdf" output in your pdf report. In some exercises we will ask you to include a code snippet as part of your solution text - a code snippet is only the most essential lines of code needed for solving the problem, this does not include import statements, other forms of boiler plate code, as well as plotting code.

Exercise 1 (2 points. (based on Blitzstein & Hwang Exercise 10.7.6) Inequalities). Let X be a random variable with mean μ and variance σ^2. Show that

$$E\left[(X-\mu)^4\right] \geq \sigma^4 \ .$$

Hint: Consider if you can use Jensen's inequality.

Deliverables. Include the steps and argumentation of the proof of the expression.

Exercise 2 (4 points. Confidence Intervals). Let $\gamma \in \mathbb{R}$ be fixed and let $X_1, \ldots, X_n$ be i.i.d. with Normal distribution $\mathcal{N}(\mu, \sigma^2)$. We estimate μ by the sample mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$. In the lecture, we have seen that

$$\sqrt{n}\,\frac{\hat{\mu}-\mu}{\sigma} \sim \mathcal{N}(0,1). \qquad (1)$$

a) Pretend that (1) holds, even if we replace σ by the estimator $\hat{\sigma} := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\hat{\mu})^2}$ (i.e. the sample standard deviation). Construct a γ-confidence interval for μ by using the procedure explained in the lecture.

b) We can make a simulation of many experiments at a fixed n, and compute the probability of the correct μ value being outside the confidence interval – if our estimated confidence interval is a good fit then this probability should be $< (1-\gamma)$. Modify the code in `confidenceinterv.py` (here, $n = 9$) and report, how often (out of 10000 experiments) the correct parameter lies outside the confidence interval. (Hint: Numpy's `np.var` divides by n, not by $n-1$. To correct for this, set the parameter `ddof=1` of `np.var`.)

c) In fact, (1) does not hold if we replace σ with $\hat{\sigma}$. Instead, we have

$$\sqrt{n}\,\frac{\hat{\mu}-\mu}{\hat{\sigma}} \sim t_{n-1}$$

where t_{n-1} is a student-t distribution with $n-1$ degrees of freedom. Again, report the corresponding confidence interval, modify the notebook, and report, how often the correct parameter is not covered. (Hint: Use `scipy.stats.t.ppf(q, n-1)` to compute the critical value by reverse look up in the CDF of the t distribution of $n-1$ degrees of freedom.)

Deliverables. a) Write the correct expression for the γ-confidence interval under these assumptions, b) your modified version of the code and your answer to the question, c) write the correct expression for the γ-confidence interval under these assumptions, and include your modified code and your answer to the question.

Exercise 3 (4 points. Hypothesis Testing). A scientist claims that he has found a single gene that has an influence on the flowering time of a plant. In order to see whether his claim is true, he obtains five pairs $(X_1, Y_1), \ldots, (X_5, Y_5)$ of two genetically identical replicates. In each second replicate $(Y_1, \ldots, Y_5)$ he has knockedout the gene. The following table shows the flowering time (in days).

Plant	1	2	3	4	5
Replicate 1 without knockout	4.1	4.8	4.0	4.5	4.0
Replicate 2 with knockout	3.1	4.3	4.5	3.0	3.5

Assume that the differences $X_i - Y_i$ (that is, flowering time replicate X minus flowering time replicate Y) are normally distributed with mean μ and variance σ^2.

a) Choose the null hypothesis and briefly justify your answer.
b) Perform the corresponding t-test to the level 0.05 (Hint: Use the "six steps" from the lecture).
c) Can the scientist change the test result by (illegally) copying the data set, that is, by writing down each data point k times and pretending that he has investigated $5 \cdot k$ independent pairs of plants? Justify your answer.

Deliverables. a) Your justified answer, b) explain the six steps you go through in the t-test and whether or not you can reject the null hypothesis, c) your justified answer.

Exercise 4 (2 points. (from 27.11.) Maximum Likelihood). Let the random variables $X_1, \ldots, X_n$ be i.i.d. following the geometric distribution $\mathrm{Geo}(\theta)$ with PMF $p_\theta(x) = (1-\theta)^{x-1}\theta$ and that the values $x \in \mathbb{Z}_+$ (set of positive integers excluding zero), which then also holds for each of $X_1, \ldots, X_n$. Prove that

$$\hat{\theta}_n = \frac{n}{\sum_{i=1}^{n} x_i}$$

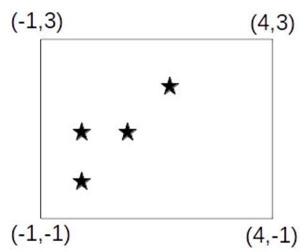
is the maximum likelihood estimator (MLE) for the parameter θ. (Hint: Sometimes it is easier to maximize the logarithm of a function instead of the function itself.)

Deliverables. Include the steps of your proof of the maximum likelihood estimator for the parameter θ, including showing that the estimator is indeed a maximum of the likelihood function.

Exercise 5 (4 points. 4-dimensional Maximum Likelihood). During night, a prisoner sits in a completely dark room and faces a dark wall. He knows that the wall has a window, but he neither knows the window's size nor its exact position. The only thing that he sees are four stars, i.e., light points, at positions $(0,0)$, $(0,1)$, $(1,1)$ and $(2,2)$, which obviously must be within the window. He then wants to infer the boundary of the window by maximum likelihood. Assume that the points $(X_1, Y_1), \ldots, (X_4, Y_4)$ representing stars visible in the window are independent and identically distributed (i.i.d.) by a uniform distribution with the parameters $\theta := (x_{\min}, x_{\max}, y_{\min}, y_{\max})$. That is, the probability density function (PDF) for distribution of stars seen through the window is

$$f_\theta(x, y) = \begin{cases} c & \text{if } x_{\min} \le x \le x_{\max} \text{ and } y_{\min} \le y \le y_{\max} \\ 0 & \text{otherwise.} \end{cases}$$

a) Find the correct value for $c \in \mathbb{R}$ (Hint: Recall the defining properties of a probability density function).
b) Compute the likelihood for the two sets of parameter values $\theta_1 = (-1, 4, -1, 3)$ (see illustration to the right) and $\theta_2 = (-2, 5, -3, 6)$.
c) Find the maximum likelihood estimator (MLE) for the parameter set, $\hat{\theta}^{\mathrm{ML}} = (\hat{x}_{\min}, \hat{x}_{\max}, \hat{y}_{\min}, \hat{y}_{\max})$ given the observed stars at positions $(0,0)$, $(0,1)$, $(1,1)$ and $(2,2)$. (Hint: Consider for which cases the likelihood > 0, and use the fact that the PDF is uniform).



Deliverables. a) Include the derivation steps and the result, b) write the expression for the likelihood function as well as computing its values for the two sets of parameters, c) include argumentation for the result.