$\hat{\mu} = \bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$

$\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2$

MAD 2019-20, Assignment 3

Bulat Ibragimov, Fabian Gieseke, Kim Steenstrup Pedersen

hand in until: 10.12.2019 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. 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.

Exercise 1 (**2 points. (from 10.7.6) Inequalities).** Let X be a random variable with mean μ and variance σ^2. Show that

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

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

$\bar{X}$ = sample mean

S^2 = sample variance

Sample standard deviation

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

a) Pretend that (1) holds, even if we replace σ by its estimate $\hat{\sigma} := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2}$. Construct a $(1-\gamma)$-confidence interval for $\hat{\mu}$.

b) Modify `confidenceinterv.ipynb` (here, $n = 9$) and report, how often (out of 10000) the correct parameter lies outside the confidence interval. (Hint: python's `np.var` divides by n, not by $n-1$.)

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. Show the code. (Hint: the r-quantile of a t_{n-1} distribution is denoted by $t_{n-1;r}$ and can be computed by `scipy.stats.t.ppf(r, n-1)`.)

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 (flowering time replicate 1 minus flowering time replicate 2) are normally distributed with mean μ and variance σ^2.

a) Choose the null hypothesis and briefly justify your answer.

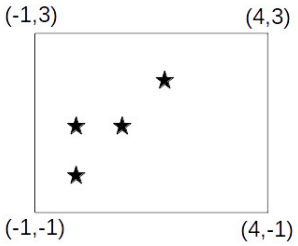
b) Perform the corresponding t-test to the level 0.05 (Hint: use the "six steps").

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. (You can use that $t_{f;1-\alpha/2}$ converges against $z_{1-\alpha/2}$ with increasing f.)

Exercise 4 (2 points. (from 27.11.) Maximum Likelihood). Let $X_1, \ldots, X_n$ be i.i.d. with geometric distribution $\mathcal{G}eo(\theta)$ with PMF $p_\theta(x) = (1-\theta)^{x-1}\theta$. Compute the maximum likelihood estimator (MLE) for θ. (Hint: sometimes it is easier to maximize the log of a function than a function itself.)

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 $(2,2)$. He then wants to infer the boundary of the window by maximum likelihood. Assume that the points $(X_1, Y_1), \ldots, (X_4, Y_4)$ are i.i.d. from a uniform distribution with parameters $\theta := (x_{\min}, x_{\max}, y_{\min}, y_{\max})$. That is, the pdf equals

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



a) Find the correct value for $c \in \mathbb{R}$.

b) Compute the likelihood for the values $\theta_1 = (-1, 4, -1, 3)$, see right, and $\theta_2 = (-2, 5, -3, 6)$.

c) Compute the MLE $\hat{\theta}^{\mathrm{ML}} = (\hat{x}_{\min}, \hat{x}_{\max}, \hat{y}_{\min}, \hat{y}_{\max})$.