04. Maximum Likelihood Estimation and Bootstrapping

To be completed Friday 25th of November, by 12 noon sharp. You are encouraged to submit in pairs, but are also allowed to submit alone. **One person** must submit the solution as **a single pdf file**, to the folder on Ilias, by naming *exercise_03_name1_matriculationnumber1_name2_matriculationnumber2.pdf*, with the obvious replacements in the strings. **Not adhering to the formatting requirements may result in the submission not being graded.**
Exercises are only graded in a binary fashion as sufficient or insufficient. To be graded as sufficient, you do not necessarily have to have correct solutions to every sub-question, but you must have made a clear and earnest effort to solve the entire exercise. Ultimately, what constitutes sufficient is at the discretion of the tutors. To be admitted to the exam, you must have submitted sufficient answers to at least 5 of the 6 (maybe 7) exercise sheets.

1. **EXAMple Question — iPhone sales**
   A common manufacturing practice is assigning ascending sequences of serial numbers in produced goods. In 2008, a London investor started asking for people to post the serial number of their iPhone and the date they bought it so that he could decipher how many phones Apple sold that year[1]. In this exercise, you will encounter the shortcomings of using maximum likelihood estimation for such a purpose.

   Suppose people posted serial numbers $X = \{x_1, x_2, \dots, x_n\}$. From them, we want to estimate the total number N of iPhones sold this year. We assume that the posted iPhone serial numbers are i.i.d. and their distribution is uniform in the range {1, 2, .., N}, and that the serial numbers are assigned sequentially and are integers that restart from 1 every year.

   (a) What is the likelihood of a single serial number? What is the joint likelihood of the observed serial numbers?
   (b) Given $X = \{4, 1860, 972029, 3872915, 65\}$, estimate N using MLE.
   (c) Critically assess your answer. What are the shortcomings of MLE in this example?

2. **Theory Question — Exam room booking**
   You are the lecturer of the Data Literacy course and you want to book a room for the final exam. Naturally, not all students will turn up, so if you booked a room that has the capacity of the number of enrolled students, then this may cause unnecessary problems to other courses that might actually need one of the very few big rooms at the University of Tübingen. Thus, you wish to estimate the expected number of students that will sit the exam from the number of students attending the lectures.

   Assume that all students have independent probability of attending p, the number of enrolled students is n, and the probability distribution which describes k out of n enrolled students showing up to the exam ($k \leq n$) is the binomial distribution[2]. Let's say we collect attendance numbers from the lectures $K = \{k_1, k_2, \dots, k_l\}$, with l being the number of the lectures.

   (a) What is the likelihood of a single observation k? What is the joint likelihood of the observations?
   (b) Given $K = \{97, 88, 85, 72, 43, 54, 55, 82, 65, 52, 73, 42, 75\}$ and $n = 100$, estimate p using MLE.
   (c) Now that we know p, if you book a room with 75 seats, what is the probability it can accommodate all the students that will show up in the final exam? *In case you were not successful in solving 2b, assume a value for p, but please state this clearly.*

3. **Practical Question — Bootstrapping**
   In this exercise, you will use *bootstrapping* to compute *confidence intervals* for an estimator. See **exercise_04.ipynb**.

[1] Guardian article: `https://www.theguardian.com/technology/blog/2008/oct/08/iphone.apple`
[2] PMF (probability mass function) on `https://en.wikipedia.org/wiki/Binomial_distribution`