01. Bayes to the Rescue

To be completed Friday 4th of November, by 12 noon sharp. You are encouraged to submit in pairs, but are also allowed to submit alone. You must submit your solution as **a single pdf file**, to the folder on Ilias, by naming *exercise_01_name1_matriculationnumber1_name2_matriculationnumber2.pdf*, with the obvious replacements in the strings. **Not adhering to the formatting requirements might result in the submission not being graded.**
In particular, solutions to coding questions must be submitted by exporting the output of your jupyter notebook as a pdf. This drastically simplifies evaluation by the tutors (who would otherwise have to run your exercise sheet, which can take significant time). The cleanest way to make a submission is to type the answers to the written exercises in markdown cells at the top of your jupyter notebook, using jupyter's LATEX abilities. You can also type up the solution in normal LATEX and add the resulting pages to the front of your pdf by any method you prefer. It is also OK to write your solutions by hand and scan it, adding the scan to the pdf.
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 — Prosecutor's Fallacy**
   Oh no, you are wrongly accused of stealing a bike! The whole thing is taken to court, and the sole piece of evidence presented against you is a positive DNA match. The prosecutor claims that this is definitive evidence indicating you are guilty of the accused, as the probability of a DNA match, given one is guilty is 99%.
   Luckily you have some knowledge of statistics up your sleeve. The probability of a DNA match, given one is not guilty is 0.001, and the percentage of bike thieves in the population of your city is 0.01%. What is the actual probability of you being guilty, given the presented evidence?

2. **Theory Question — Pooled Testing**
   Suppose you have a Covid test with given sensitivity $P(T|C)$ and specificity $P(\neg T|\neg C)$. The incidence $P(C)$, is also known. To save money, you want to start doing pooled testing: k independent samples are thrown into the same test, and tested as one. You can assume that the probability of the test being positive given that at least one person in the pool of k samples is infected, is equal to the original sensitivity. Similarly, the probability of the test being negative given that all the people in the pool are not infected is equal to the original specificity. Phrase all your answers in terms of the three known probabilities.

   A. What is the probability of at least one person in a sample of k people being infected?

   B (optional). What is the specificity of the pooled test for one person (i.e., the probability of a negative result of the pooled test, given that the person is not infected)?

   Hint: you will likely have to introduce two variables, one for the probability of you being (not) infected, and one for everyone else in in the pool, besides you

3. **Practical Question — Bayesian Spam Filter**
   In this week's practical question, we will practice handling data with pandas and applying Bayesian statistics. More concretely, you will implement a Bayesian spam detector, using an SMS Spam Collection Data Set. You will find the required information in notebook `Exercise_01.ipynb`, which can be solved without solving the questions above.