Exercise Sheet 1

Probability theory and Bayes' theorem

Moodle. If not yet done, register on Moodle at https://moodle.msengineering.ch. Navigate to *Home* → *Region D* → *TSM - Technical Scientific Modules* → *AUT25* → *TSM_ BayMachLe Zurich* and use subscription key: `moodlemsekey`.

Solutions. You do not need to submit your exercise solutions to us. Of course you are free to ask questions during tutorial hours! Exercise solutions will be published on Moodle at the same time with the exercise sheet. Use solutions responsibly, otherwise you risk not being able to solve the exam to a satisfactory level.

Setting up a Python environment

To solve the exercises and run example notebook, you will need a working Python environment with the right packages installed. There are different ways to do this:

a) Remotely: Run your code in a Session on renkulab.io (either on the official repo with no writing permissions or on your own fork).

b) Locally: Download the notebooks from Moodle and set up your own environment (with manual installations or by using the renku CLI).

The different variants of how to run code are explained in the slide deck in the week 1 materials on Moodle (`Introduction_to_Renkulab.pdf`).

Exercise 1 What view of probability is more appropriate?

Decide whether you would rather use a frequentist or a Bayesian understanding of probability to answer the following problems:

a) A medical researcher wants to determine the exact rate of prostate cancer for males between 50 and 60 in Switzerland.

b) A gambler goes into a casino and looks for the game with the highest probability of winning money.

c) A poker player wants to create a mathematical model that predicts the moves of his opponents. He knows the strategies of most of his opponents quite well.

d) A particle physicist has run an interesting experiment. Now she wants to assess whether the collected data puts the standard model of particle physics into question. From her education she knows that a long series of experiments have confirmed the standard model.

e) An insurance company wants to estimate the number of car accidents per year in their portfolio to guess what reserves that they have to put aside.

f) An economist predicts inflation rates in different countries for next year.

g) A doctor is asking himself whether you could have scabies.

h) An engineer is in charge of a production process producing thousands of special screws throughout a day. Her boss asks her to give an estimate of the rate of defective screws.

Exercise 2 The law of large numbers

The law of large numbers states that a relative frequency will stabilize in the limit of large numbers (respectively infinitey). Write a short numpy code that simulates N coin tosses and computes the cumulative relative frequency at each subsequent coin toss. Visualize this cumulative relative frequency. Does your simulation stably reach 0.5 in the limit of large N? How large does N approximately need to be to reach at least three zeros behind the 0.5? (or three nines behind the 0.4)

Exercise 3 Binomially distributed random variables

A manufacturing company produces small pumps that go into ship motors. Your boss claims that they have estimated the rate of defect pumps to be $\pi = 5\%$. The company produces 20 pumps per day. To compute probabilities of different daily outcomes for the number of broken pumps Y, you may use the binomial distribution:

$$P(Y = k|\pi) = \binom{n}{k} \pi^k \, (1-\pi)^{n-k}.$$

You may use Python to give your answers! (e.g. with `scipy.stats.binom`).

a) What is the probability that exactly 5% of the pumps are defective on a production day? Try to compute your answer also in plain Python (using `math.comb()` for the binomial coefficient).

b) Draw the probability mass function for the number of defective pumps per day with a bar plot.

c) What is the probability that more than 3 pumps are defective on a production day? Use the cumulative distribution function (CDF).

d) On one day, 10 pumps are defective. What is the probability for this event and what do you suspect happened?

Exercise 4 Expectation of a PMF

Many count variables follow a Poisson distribution. The PMF of the Poisson distribution is

$$p(k) = P(X = k) = \frac{\lambda^k \, e^{-\lambda}}{k!},$$

where $k \in \mathbb{N}$ (including 0) and $\lambda \in \mathbb{R}$ is a parameter. Show that the expectation of the Poisson distribution is λ.

Hint: Use $E[X] = \sum_{k=0}^{\infty} k \, p(k) = \sum_{k=1}^{\infty} k \, p(k)$ and introduce $l = k - 1$ at an appropriate time. Finally, remember that $e^{\lambda} = \sum_{l=0}^{\infty} \frac{\lambda^l}{l!}$.

Exercise 5 Maximum Likelihood Estimator

This exercise anticipates what we will be studying in next week's lecture and is a repetition of the frequentist maximum likelihood principle.

Let's consider the following problem: A coin has been tossed $n = 10$ times, out of which it came up with heads $k = 8$ times. What is a likely value for the coin's probability of head π ?

In your frequentist education, you have learnt that $\hat{\pi} = \frac{k}{n}$ is a good estimator. In fact, it is the so-called maximum likelihood estimator for the binomial distribution that we want to derive in this exercise, in order to be better able to distinguish between frequentist and Bayesian inference.

The probability to measure a particular outcome (number of heads) can be computed with the binomial distribution:

$$P(Y = k \mid \pi) = \binom{n}{k} \pi^k \, (1 - \pi)^{n-k}.$$

In the maximum likelihood method, we look for the value of π that is most likely to have produced our observation. In other words, we are looking for the value of π that yields the maximum value for $P(Y = 8|\pi)$. To this end, one typically defines the *likelihood function*

$$L(\pi \mid Y = k) \equiv P(Y = k \mid \pi)$$

in the domain of $\pi \in [0, 1]$. Note that L is not a probability distribution over π (as it does not integrate to 1), for this reason Sir Ronald Fisher called it *likelihood*, as he needed another word than probability. In maximum likelihood, L is maximized.

a) Use the first derivative by π to show that the estimator $\hat{\pi} = \frac{k}{n}$ maximizes L for a given k and n:

$$\frac{\mathrm{d}}{\mathrm{d}\pi} L(\pi \mid Y = k) \stackrel{!}{=} 0$$

For simplicity, assume that $\pi \in (0, 1)$, i.e. $\pi \notin \{0,1\}$.

b) Use Numpy and Matplotlib to plot $L(\pi|Y = 8)$ for different values of π and verify that the maximum is at $\hat{\pi}$.

Hint: Use `scipy.stats.binom.pmf()` to compute $L(\pi|Y = 8)$.

Exercise 6 Fighting terrorism

You work for the Swiss government and are in discussions with a security contractor who offers you a face detection software that could be installed at Zurich mainstation. The following are its performance measures:

- Given that a person is registered as a terrorist, it finds them with a probability of 95%.
- Innocent citizens are only recognized as terrorists with a probability of 0.01%.

Would you buy their system and employ it?

Exercise 7 Prison populations in Switzerland

In 2024, 70.1% of all prison inmates in Switzerland were not of Swiss nationality. One might be inclined to conclude that the majority of Switzerland's inhabitants not holding the Swiss passport are criminal. Using Bayes' theorem, show that this is not a correct interpretation and needs to be put into perspective. In particular, use the the following additional numbers:

- There are 8.74 Million inhabitants in Switzerland.
- About 27% of these 8.74 Million inhabitants are not of Swiss nationality.
- About 6400 of these 8.74 Million inhabitants are currently in prison.

Given events P (person is in prison) and S (person is of Swiss nationality) compute, compare and interpret the probabilities $P(P|S)$ and $P(P|\bar{S})$.

Exercise 8 Non-Invasive Prenatal Testing

Non-invasive prenatal testing (NIPT) is a blood test to determine the risk for a fetus to be born with trisomy 21, 18 or 13. The test has to be conducted between week 5 and 7 of pregnancy and is typically only applied when the first trimester screening (FTS - or Ersttrimestertest ETT in German) is conspicuous.

In this exercise, we focus on NIPT's abilities to detect the presence of the down syndrome (trisomy 21). The literature indicates a sensitity of 99.2% and a specificity of 99.91%[1]. The prevalence of trisomy 21 depends on age: at age 20 it's around 0.1% or less, at age 40 it increases significantly towards 1.4%.

You have just administered such a test and want to be prepared what it means to have a positive or negative test result.

a) Given the events D: "*Down syndrome present*" and T: "*test result is positive*", formulate and compute the following probabilities of interest for a 40 year old woman:

1. that the Down syndrome is present given that the test turns out positive,
2. that the Down syndrome is present given that the test turns out negative,

b) Answer the same questions for a 20 year old woman and explain the principal reason behind why NIPT is typically not used on 20 year old women.

c) Using the same argument as in b), argue that the probability that you actually have Covid-19 given a positive Covid-19 self test was a time varying function between 2020-2022. What causes the variability in time?

[1] source: aerzteblatt.de

Exercise 9 Magic coin

You bought a magic coin from a wizard shop. Unfortunately, you have thrown away the instruction manual and cannot remember whether the coin is biased towards heads or towards tails. This is important, since you have to perform a magic trick in a few minutes. You call the company and they tell you that the coin comes in two models:

- model 1 with $\pi = 0.2$ (probability to show heads is 20% in the long run) is sold in 10% of the cases
- model 2 with $\pi = 0.8$ is sold in 90% of the cases

a) You throw the coin 5 times and it shows heads twice. What model do you think it is with your limited amount of data?

   Hint: Compute the ratio of

$$\frac{P(\pi = 0.8 \mid d)}{P(\pi = 0.2 \mid d)},$$

   where d stands for your collected coin toss data, and use Bayes' theorem and the binomial distribution.

b) Luckily, your magic trick worked and the coin showed heads as you predicted. At home, you have time for more flips. Interestingly, you flip the same ratio, just with ten times higher numbers: Out of 50 flips, you get heads 20 times. How much do you believe now that you have model 2?

c) Argue why a frequentist magician would have failed at this task.

d) Under what circumstance would you (the Bayesian magician) and the Frequentist magician have come to the same conclusion?

   Hint: The frequentist cannot use the above ratio for decision making, but only

$$\frac{P(d \mid \pi = 0.8)}{P(d \mid \pi = 0.2)}.$$

   When are these two ratios the same?

Exercise 10 Covid symptoms

In 2021, the Israeli Ministry of Health has publicly released data of all individuals who were tested for SARS-CoV-2 via PCR nasopharyngeal swab (see '`covid_tests.csv`' in the exercise material on Moodle). The dataset lists the result of the covid test and the symptoms of the test subject. You are a data scientist (from 2021) and want to create a simple rule how individuals are to be prioritized for a test according to their symptoms (anticipating a shortage of the available covid tests).

a) Read the data from the CSV with pandas and perform the following additional actions:
   - select only the columns '`corona_result`' (result of the test) and the symptoms '`cough`', '`fever`', '`sore_throat`', '`shortness_of_breath`' and '`head_ache`',
   - filter out only 'positive' and 'negative' test results
   - replace 'positive' test results with 1 and 'negative' with 0

b) Using the dataset, compute the conditional probabilities $P(\text{covid}|\text{cough})$ and $P(\text{cough}|\text{covid})$.

c) Use your pandas skills to compute all values for $P(\text{symptom}|\text{covid})$ and $P(\text{covid}|\text{symptom})$.

   Hint: $P(\text{symptom}|\text{covid})$ is easy to compute in a vectorized way using `groupby()`. The inverse probability is more difficult to do in a vectorized way, but it works if you compute the marginal probabilities ($P(\text{symptom}), P(\text{covid})$) and multiply / divide them with $P(\text{symptom}|\text{covid})$ according to Bayes' theorem.

d) Which one would you rather use to prioritize individuals for covid tests, $P(\text{symptom}|\text{covid})$ or $P(\text{covid}|\text{symptom})$? List the symptoms from highest to lowest priority and argue that it makes a big difference what conditional probability you use.

Exercise 11 Sick trees

This is an exercise borrowed from the book *Bayes rules!*[2]:

A local arboretum contains a variety of tree species, including elms, maples, and others. Unfortunately, 18% of all trees in the arboretum are infected with mold. Among the infected trees, 15% are elms, 80% are maples, and 5% are other species. Among the uninfected trees, 20% are elms, 10% are maples, and 70% are other species. In monitoring the spread of mold, an arboretum employee randomly selects a tree to test.

a) What's the prior probability that the selected tree has mold?

b) The tree happens to be a maple. What's the probability that the employee would have selected a maple?

c) What's the posterior probability that the selected maple tree has mold?

d) Compare the prior and posterior probability of the tree having mold. How did your understanding change in light of the fact that the tree is a maple?

e) There is another way to compute the posterior probability using a *simulation*! Proceed in the following way:

   1. Simulate 10'000 random trees and whether they have mold with `np.random.choice()` according to the prior probability $P(\text{mold})$ and store your result in a data frame.
   2. Assign a maple probability ($P(\text{maple}|\text{mold})$ or $P(\text{maple}|\overline{\text{mold}})$) from the description above to each tree according to whether the tree has mold or not.
   3. For each tree, randomly sample whether it's a maple using `np.random.choice()` and $P(\text{maple}|\text{mold})$.
   4. Compute $P(\text{mold}|\text{maple})$ by filtering out only maple trees and computing the ratio of them that has mold.

   How close is the simulated posterior probability to the one computed in d) with Bayes' theorem? Take some time and ponder why this simulation works. In the literature, this type of simulation is often called *ancestral sampling*, since it samples from two hierarchical levels.

[2] https://www.bayesrulesbook.com