Data Science I

Overall Points: / 100

Exercise 2: Mathematical Preliminaries

<u>**Submission Deadline: November 03 2025, 07:00 UTC**</u>
University of Oldenburg
Winter 2025/2026
Instructors: Jannik Schröder, Wolfram "Wolle" Wingerath

Submitted by: <your names here>

Part 1: Probabilities

/ 25

1.) Suppose that 80% of people like peanut butter, 89% like jelly, and 78% like both. Given that a randomly sampled person likes peanut butter, what is the probability that she also likes jelly?

Solution:

In [ ]:
```
<just add code cells as required>
```

By the way, you can also add markdown cells where you can use LaTeX, e.g. in equations:

$$
\begin{aligned}
P(A \cap B) &\neq P(A \cup B) && \mid \text{blah blah blah} && (1)\\
&= \ldots && && (2)\\
&= \underline{\underline{< result >}} && && (3)
\end{aligned}
$$

2.) Suppose that P(A) = 0.3 and P(B) = 0.7.
(a) Can you compute P(A and B) if you only know P(A) and P(B)?

Solution:

XXX

(b) Assuming that events A and B arise from independent random processes:
(1) What is P(A and B)?

Solution:

XXX

(2) Again assuming that A and B are independent, what is P(A or B)?

Solution:

XXX

(3) Assuming again the independence of A and B, what is P(A|B)?

Solution:

XXX

3.) Consider a game where your score is the maximum value from two dice throws. Write a small python function that outputs the probability of each event from {1, ..., 6}.

Solution:

XXX

4.) If two binary random variables X and Y are independent, is the complement of X and Y also independent? Give a proof or a counterexample.

Solution:

XXX

Part 2: Statistics

/ 25

1.) Consider the following pair of distributions:

In [ ]:
```
dist1 = [3, 5, 5, 5, 8, 11, 11, 11, 13]
dist2 = [3, 5, 5, 5, 8, 11, 11, 11, 20]
```

(a) Decide which one has the greater mean and the greater standard deviation without computing either of those.

Solution:

XXX

(b) Compute mean and standard deviation of these distributions using built-in functions of the pandas library.

Solution:

XXX

(c) Now define a function yourself to compute mean and standard deviation. Verify that the functions work correctly by comparing the output with the output of the built-in functions.

Solution:

XXX

2.) Consider the following distribution:
Distribution 1: [1, 1, 1, 1, 1, 1, 1, 1, 1]
How do the arithmetic and geometric mean compare?

Solution:

XXX

3.) How do the arithmetic and geometric mean compare on random integers that are not identical? Draw a sample distribution of size 10 a few times and write down your findings.

Note:
Please implement your own version of the geometric mean and use it in this task.

Solution:

XXX

Part 3: Correlation Analysis

/ 25

1.) A correlation coefficient of –0.9 indicates a stronger linear relationship than a correlation coefficient of 0.5 – true or false? Explain why.

Solution:

XXX

2.) Compute the Pearson and Spearman Rank correlations for uniformly drawn samples of points (x, x^k) and answer the questions below.

Note:
You can use scipy.stats.spearmanr and scipy.stats.pearsonr to compute the ranks.

Solution:

XXX

(a) How do these values change as a function of increasing k?

Solution:

XXX

(b) Does it matter whether x can be positive and negative or positive only?

Solution:

XXX

Part 4: Logarithms

/ 25

Recall the definition of the geometric mean:

$$\left(\prod_{i=1}^{n} a_i\right)^{1/n} = \sqrt[n]{a_1 \cdot a_2 \cdots a_n}$$

For large sample sets or small sample sets with very large numbers, the computed product will become very large. Chances are that your implementation of the geometric mean does not work with large sample sizes, either – go ahead and try it with a sample size of, let's say, 1000!

Assuming that your implementation does not work for a sample size of 1000, create one that does!

Solution:

XXX

Finally: Submission

Save your notebook and submit it (as both **notebook and PDF file**). And please don't forget to ...

- ... choose a **file name** according to convention (see Exercise Sheet 1) and to
- ... include the **execution output** in your submission!