FU Berlin • FB Mathematik und Informatik • Vesa Kaarnioja

STATISTICS FOR DATA SCIENCE

Wintersemester 2024/2025

— MOCK EXAM —

Name: .... **Student ID:** ....

Important information:

- *This is a mock exam and it will not be graded.*
- The exam consists of a total of **5 tasks**. The tasks should be completed within **90 minutes**.
- Each task is worth 6 points. You will need **18 points** to pass the exam.
- Please write your solutions directly on the exam paper. If you need additional sheets of paper, please label them with your name and indicate which tasks they belong to. You can ask the invigilator for additional sheets of paper by raising your hand.
- Present your calculations clearly and neatly, providing explanation for all steps.
- All results from the lecture slides and the exercises may be used to answer the questions.
- You may bring one handwritten, double-sided DIN A4 sheet. Only an original, handwritten note is allowed; no photocopies, printouts, or electronic devices are permitted. Other aids (e.g., lecture notes, calculators, cell phone) are **forbidden**.
- Please place your student ID card and your identity card (or passport or driving license) face up on your table.
- Cell phones and other electronic means of communication must be switched off during the exam. Anyone who uses a cell phone during the exam will be removed from the exam and will be deemed to have failed.

Good luck!

Question 1 (Basic probability theory)
Let $f\colon \mathbb{R} \to \mathbb{R}$ be a function defined by

$$f(x) = \begin{cases} 6x - 6x^2 & \text{for } x \in [0,1], \\ 0 & \text{otherwise.} \end{cases}$$

(a) Show that f is a probability density function.

(b) Let X be a random variable with probability density function f. Compute $\mathbb{E}[X]$ and $\mathrm{Var}(X)$.

(c) Explain how you would draw a sample from the distribution with probability density function f. Outline the steps in the algorithm required for sampling from this distribution.

Solution:

Question 2 (Correlation)

(i) Explain briefly (1–3 sentences), what kind of dependence can be measured using the Pearson correlation coefficient and the Spearman rank correlation coefficient.

(ii) Let us consider the Pearson and Spearman sample correlation coefficients for the scatter plots displayed in Figure 1:

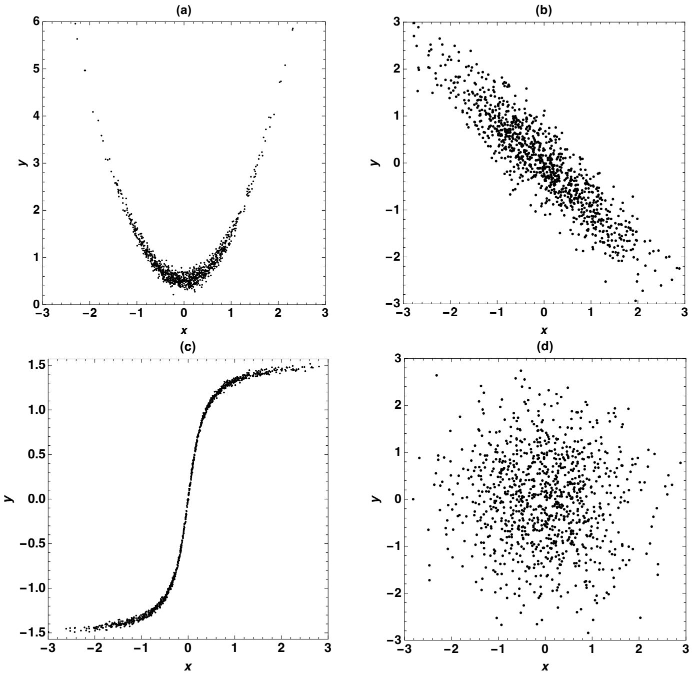

Figure 1: Scatter plots

(a) Figure a: which of the following is the sample Pearson correlation coefficient:

$$0.5, \quad 0.1, \quad -0.6, \quad -0.9.$$

Which of the following is the sample Spearman correlation coefficient:

$$0.1, \quad 0.0, \quad -0.7, \quad -0.9.$$

(b) Figure b: which of the following is the sample Pearson correlation coefficient:

$$0.1, \quad 0.0, \quad -0.3, \quad -0.9.$$

Which of the following is the sample Spearman correlation coefficient:

$$0.1, \quad 0.0, \quad -0.3, \quad -0.9.$$

(c) Figure c: which of the following is the sample Pearson correlation coefficient:

$$0.9, \quad 0.5, \quad -0.5, \quad -0.9.$$

Which of the following is the sample Spearman correlation coefficient:

$$1.0, \quad 0.5, \quad -0.5, \quad -1.0.$$

(d) Figure d: which of the following is the sample Pearson correlation coefficient:

$$0.8, \quad 0.9, \quad 0.0, \quad -0.9.$$

Which of the following is the sample Spearman correlation coefficient:

$$0.8, \quad 0.9, \quad 0.0, \quad -0.9.$$

In tasks (a)–(d), please circle the answer that you think is correct.

Solution:

Question 3 (Hypothesis testing)
A dice is rolled 120 times with the following results: the score one appears 12 times, two 16 times, three 20 times, four 17 times, five 22 times, and six 33 times.

(a) Which statistical test would you use to test the fairness of this dice?

(b) State the null hypothesis and the alternative hypothesis of this test.

(c) What are the statistical assumptions of this test? In your opinion, does the sample satisfy the statistical hypotheses of the test?

Solution:

Question 4 (Linear regression)
Ice cream consumption y (liters/day) is thought to be linearly dependent on the maximum temperature of the day x (in Celsius degrees). Consider the following sample of the variables:

x	10	17	4	7	5	6	11
y	105	163	43	69	48	56	115

(a) Draw a scatter plot of the data.

(b) Estimate both the Pearson sample correlation coefficient and the Spearman sample correlation coefficient using the scatter plot you drew in part (a). (You do not need to compute the exact values of the correlation coefficients, a rough numerical estimate is sufficient.)

(c) Estimate the coefficients $a, b \in \mathbb{R}$ of the l_2 regression line $y = ax + b$ using the scatter plot you drew in part (a). (You do not need to compute the exact values of a and b, a rough numerical estimate is sufficient.)

(d) Draw the l_2 regression line into the scatter plot.

(e) Based on these observations, what can you say about the ice cream consumption if the maximum temperature is $+30$ degrees Celsius?

Solution:

Question 5 (Bayesian inference)
Consider the measurement model

$$y = \frac{1}{2}x + \eta,$$

where $x \in \mathbb{R}$ is the unknown parameter, $y \in \mathbb{R}$ is the measurement, and $\eta \in \mathbb{R}$ is observational noise. Suppose that the unknown x has the prior probability density

$$f(x) = \begin{cases} 2\exp(-2x) & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases}$$

and the observational noise is distributed according to $\eta \sim \mathcal{N}(0, 1)$.

(a) Derive the posterior density $f(x|y)$ up to a constant factor.

(b) Solve the *maximum a posteriori* (MAP) estimator of x when we observe $y = 5$.

Solution: