Question 1 (Basic probability theory)
Let $f : \mathbb{R} \to \mathbb{R}$ 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:
(a) The function f is a probability density function (PDF) if

i. $f(x) \geq 0$ for all $x \in \mathbb{R}$, and

ii. $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$.

The function f is supported on the interval $[0,1]$, where it is defined by a downward facing parabola with roots $x = 0, 1$. Hence $f(x) \geq 0$ for all $x \in \mathbb{R}$. On the other hand, we find that

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \int_0^1 (6x - 6x^2)\,\mathrm{d}x = \left[3x^2 - 2x^3\right]_{x=0}^{x=1} = 3 - 2 = 1.$$

Therefore f is a PDF.
(b) We obtain

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x = \int_0^1 (6x^2 - 6x^3)\,\mathrm{d}x = \left[2x^3 - \frac{3}{2}x^4\right]_{x=0}^{x=1} = 2 - \frac{3}{2} = \frac{1}{2}$$

and

$$\begin{aligned} \mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x - \left(\frac{1}{2}\right)^2 &= \int_0^1 (6x^3 - 6x^4)\,\mathrm{d}x - \frac{1}{4} \\ &= \left[\frac{3}{2}x^4 - \frac{6}{5}x^5\right]_{x=0}^{x=1} - \frac{1}{4} = \frac{3}{2} - \frac{6}{5} - \frac{1}{4} = \frac{1}{20}. \end{aligned}$$

(c) Let F denote the cumulative distribution function (CDF) of f and let F^{-1} be the corresponding quantile function. We can use inverse transform sampling to obtain a random sample from this distribution as follows:

1. Draw $U \sim \mathcal{U}(0,1)$.
2. Set $X = F^{-1}(U)$.

In this case, the explicit formula for F^{-1} is somewhat complicated, so in a practical implementation it may be preferable to use the formula for the generalized inverse

$$F^{-1}(q) = \inf\{x \in \mathbb{R} \mid F(x) \geq q\}, \quad q \in (0,1),$$

in step 2 of the inverse transform sampling algorithm.

Alternative methods for drawing a sample from this distribution include, e.g., rejection sampling or MCMC.

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:

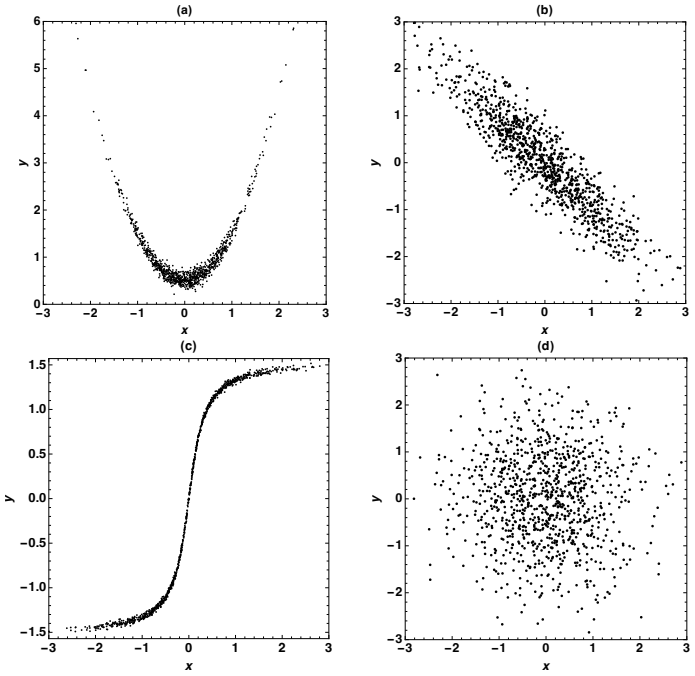


Figure 1: Scatter plots

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

$$0.5, \quad \boxed{0.1}, \quad -0.6, \quad -0.9.$$

Which of the following is the sample Spearman correlation coefficient:

$$0.1, \quad \boxed{0.0}, \quad -0.7, \quad -0.9.$$

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

$$0.1, \quad 0.0, \quad -0.3, \quad \boxed{-0.9}.$$

Which of the following is the sample Spearman correlation coefficient:

$$0.1, \quad 0.0, \quad -0.3, \quad \boxed{-0.9}.$$

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

$$\boxed{0.9}, \quad 0.5, \quad -0.5, \quad -0.9.$$

Which of the following is the sample Spearman correlation coefficient:

$$\boxed{1.0}, \quad 0.5, \quad -0.5, \quad -1.0.$$

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

$$0.8, \quad 0.9, \quad \boxed{0.0}, \quad -0.9.$$

Which of the following is the sample Spearman correlation coefficient:

$$0.8, \quad 0.9, \quad \boxed{0.0}, \quad -0.9.$$

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

Solution: (i) The Pearson correlation coefficient measures the strength and direction of a linear relationship between two variables, while the Spearman rank correlation coefficient measures the strength and direction of a monotonic relationship, which may be nonlinear.

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:
(a) The χ^2 goodness-of-fit test can be used to test whether the observations follow a given distribution F.
(b) The null hypothesis is

$$H_0: \text{the die is fair (the score follows the uniform distribution)}$$

and the alternative hypothesis is

$$H_1: \text{the die is not fair (the score does not follow the uniform distribution).}$$

(c) The observations need to correspond to i.i.d. realizations of a categorical random variable, the groups of the categorical variable must be mutually exclusive, and the sample size needs to be sufficiently large. In this case, $n = 120$ should be sufficiently large for the χ^2 test to provide reliable results. (A rule of thumb is that the expected frequencies $E_i = np_i$ should be at least 5 for each group C_i; in the present case, this is satisfied since $E_i = 120 \cdot \frac{1}{6} = 20 \geq 5$ for $i = 1, \ldots, 6$.)

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.

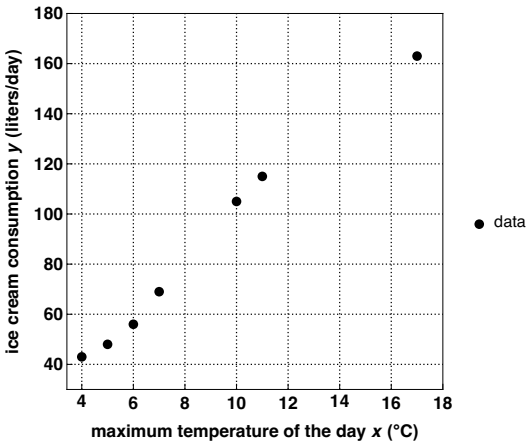
(b) Estimate both the Pearson sample correlation coefficient and the Spearman sample correlation coefficient using the scatter plot you drew in part (a). (You do not need to compute the exact values of the correlation coefficients, a rough numerical estimate is sufficient.)

(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.)

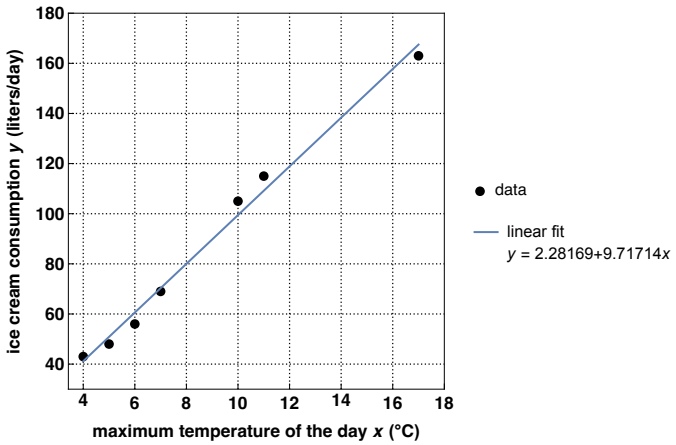
(d) Draw the l_2 regression line into the scatter plot.

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

Solution:
(a)



(b) The data lies approximately on the same line, so we expect the Pearson sample correlation coefficient to be approximately 1. Indeed, $\hat{\rho}_P(x, y) \approx 0.994858$.
The data is strictly monotonically increasing, so $\hat{\rho}_S(x, y) = 1$.
(c) The linear fit is given by $y = 2.28169 + 9.71714x$.
(d)



(e) The value $x = 30$ lies outside the range $[4, 17]$ of the observations, so our linear model cannot be used to make a prediction in this case.

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:
(a) Let $x, y, \eta \in \mathbb{R}$. The PDF of the observational noise is given by

$$\nu(\eta) \propto \exp\left(-\frac{1}{2}\eta^2\right),$$

so the likelihood function is

$$f(y|x) = \nu\left(y - \frac{1}{2}x\right) \propto \exp\left(-\frac{1}{2}\left(y - \frac{1}{2}x\right)^2\right).$$

Bayes' formula yields the posterior density

$$f(x|y) \propto f(y|x)f(x) \propto \exp\left(-\frac{1}{2}\left(y - \frac{1}{2}x\right)^2 - 2x\right)\mathbf{1}_+(x), \quad \text{where } \mathbf{1}_+(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

(b) Clearly, the mode of the posterior cannot occur when $x < 0$, so the MAP estimator can be found by solving

$$\hat{x}_{\text{MAP}} = \arg\max_{x \geq 0} f(x|5).$$

This is equivalent to finding the minimizer of the negative log-posterior

$$F(x) = \frac{1}{2}\left(5 - \frac{1}{2}x\right)^2 + 2x, \quad \text{when } x \geq 0.$$

Setting the derivative of F to 0 yields

$$0 = F'(x) = -\frac{1}{2}\left(5 - \frac{1}{2}x\right) + 2 = \frac{1}{4}x - \frac{1}{2} \quad \Leftrightarrow \quad x = 2,$$

meaning that this point is a local extremum. Since $F''(x) = \frac{1}{4} > 0$ for all $x \geq 0$, F is a convex function, so $x = 2$ is a global minimum of F and thus a global maximum of the posterior $f(x|5)$. We conclude that $\hat{x}_{\text{MAP}} = 2$.