Probability and Statistics course

Introduction

Gabor Vigh

TTK Department of Probability Theory and Statistics
ELTE

October 9, 2025

Week 6

Midterm

- There will be a midterm worth of 50 points on October 20 and a exam worth 50 points on December 8 during the class. . Thresholds for the grades:
  - 0 – 34 : 1
  - 35 – 52 : 2
  - 53 – 65: 3
  - 66 – 78: 4
  - 79 – 100: 5.
- Both exams and mid-term are mandatory. To pass them, you need to obtain at least 30% from both of them. If you do not manage, you need to participate in the retake. There is only one possibility for a retake. The only cases where you are allowed not to participate in the midterm/exam is when you are ill, and then I expect you to bring me a proper document from the doctor confirming that you were not able to appear at the faculty due to illness, or when you have another exam at the same time, and then you need to bring a confirmation from another professor.
- If you bring me such a document, I will organize 2 retakes for you if needed (so that everyone has 2 attempts).
- In case you cannot appear on the midterm/exam and you fail to bring the proper documentation justifying your absence, I will organize the so-called "utovizsga" at the end of the semester for you, which is a big exam consisting of the material from the whole semester.

Covariance

Definition (Covariance)

Covariance of two random variables:

$$\text{cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

The joint density function of (X,Y) is $f(x,y)$ if for every appropriate subset $A \subseteq \mathbb{R}^2$ we have $\mathbb{P}(X \in A) = \int_A f(x,y)\,dx\,dy$. Calculation of covariance if the joint density function is $f(x,y)$:

$$\text{cov}(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy \cdot f(x,y)\,dx\,dy - \left(\int_{-\infty}^{\infty} x \cdot f_X(x)\,dx\right) \cdot \left(\int_{-\infty}^{\infty} y \cdot f_Y(y)\,dy\right),$$

where $f_X(x)$ and $f_Y(y)$ are the marginal densities that can be computed as $f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$ and $f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$.

Correlation

Definition (Pearson Correlation)

The (Pearson) correlation coefficient of X and Y (this is always at least -1 and at most 1):

$$R(X, Y) = \frac{\text{cov}(X, Y)}{sd(X)sd(Y)}.$$

$R(X, Y) = 1$ means that $Y = aX + b$ for some a, b, with $a > 0$

Examples

A shop is selling smaller and larger cups of coffee. The price of the small version is 200, the price of the large version is 300. Let X be the number of small coffees sold in a day, and Y the number of large coffees sold in a day. We assume that X and Y are independent, have Poisson distribution, and both have expectation 80. Determine the correlation coefficient of the total number of coffees sold in a day and the income (the total price of coffees sent in a day). What will be different if the price of the large version is not 300, but $350, 400, 450$ etc.?

Corr and Cov examples

Examples

Suppose that the joint density function of (X, Y) is given as follows.

$$f(x,y) = \begin{cases} x+y, & \text{if } 0 \le x \le 1 \text{ and } 0 \le y \le 1; \\ 0 & \text{otherwise.} \end{cases}$$

Determine the correlation coefficient of X and Y.

Examples

1. There are 15 boys and 18 girls in a school class. Suppose that students are absent from school independently of each other, everyone with probability $1/10$ on a given day. Determine the covariance and correlation coefficient of the number of girls who are absent and the total number of absent students.
2. We measured temperature at a given place with two different tools. Suppose that the results are independent, normal variables with expectation 6 and standard deviation 2. Let X be the first result, Y the second (with the second tool). Determine the following quantities: $\mathrm{cov}(X, X+Y), \mathrm{cov}(X, \frac{X+Y}{2})$, $\mathrm{cov}(X-Y, X+Y)$, $R(X, \frac{X+Y}{2})$.

Example Midterm

1. For which $n > 1$ will be independent
   a) the following events: A: there is at least one head and at least one tail, B: there is at most one tail from n tosses of a coin.
   b) the following events: A: there is at least one head and at least one tail and B: the first result is head, from n flips of a coin.
2. A medical examination result shows condition "A" (e.g. high cholesterol level) with probability 0.05 and condition "B" (e.g. low iron level) with probability 0.03. The probability of joint occurrence is 0.01.
   a) Are the two conditions independent? If not, what would be the probability of joint occurrence in case of independence?
   b) What is the probability that a randomly chosen person has neither of the two conditions?
   c) What is the probability of finding condition "B" among those people with condition "A"? Conversely, what is the probability of finding condition "A" among those people with condition "B"?
3. Let us assume that the number of faults against a given player during a basketball game has Poisson distribution with parameter $\lambda = 2$ and that each of these faults is noticed by the referee with probability 0.5 (independently from the others). Compute the distribution of the noticed faults!
4. Let the density function of X $f(x) = 1 - |x|$ for $-1 < x < 1$ (and 0 otherwise). Compute $E(X)$ and $Var(X)$!
5. Let us assume that the strength of a rope is a normally distributed random variable with expectation 1000 kN. What can the standard deviation be if we know that the probability of a rope to be weaker than 980 kN is 2%?
6. Let X and Y be independent random variables with mean 0 and variance 1 Give a value c such that $(3X - cY, 2X + Y) = 0$.

Example Midterm

6 What is shown by the following R-code? What is given by p? What is the theoretical value of p?

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm
temp = np.random.normal(loc=1, scale=2, size=1000)
plt.figure(figsize=(10, 6))
plt.hist(temp, bins=30, color="orange", density=True, edgecolor="black")
xvalues = np.linspace(-6, 8, 100)
yvalues = norm.pdf(xvalues, loc=1, scale=2)
plt.plot(xvalues, yvalues, color="blue", linewidth=3)
plt.title("Normal distribution")
plt.xlabel("values")
plt.ylabel("frequencies (density)")
plt.grid(True)
plt.show()
p = np.sum((-1 < temp) & (temp < 3)) / len(temp)
print(f"The calculated proportion p is:  {p}")
```