Universität Bielefeld
Tanja Pasurek
Ehsan Abedi WiSe 2023/24

Foundations of Statistics

Homework 5

Lecture material: Chapters 1.4–1.6

Exercise 1. (Normal distribution). Let X be a $\mathcal{N}(\mu, \sigma^2)$ distributed random variable with probability density function (PDF) $\phi_{\mu,\sigma^2}(x)$ and distribution function (CDF) $\Phi_{\mu,\sigma^2}(x) := \mathbb{P}(X \leq x)$.

a Show that the distribution of the standardized random variable $Z := \frac{X-\mu}{\sigma}$ is $\mathcal{N}(0, 1)$ (= *standard normal distribution*).

b Show $\mathbb{P}(X \leq b) = \Phi_{0,1}\left(\frac{b-\mu}{\sigma}\right)$ and deduce a formula for $\mathbb{P}(a \leq X \leq b)$.

c Show $\Phi_{0,1}(x) + \Phi_{0,1}(-x) = 1$.

d For $\mu = 0$ and $\sigma^2 = 1$ find b with `R` such that $\mathbb{P}(-b \leq X \leq b) = 0.8$.

e Use parts b and c to show that the value of $\mathbb{P}(\mu - \sigma \leq X \leq \mu + \sigma)$ does not depend on μ, σ. Calculate its value in `R`.

f Generate 10000 random samples of X with arbitrary numeric values μ and σ and verify the result of (e) with a suitable simulation in `R`.

Exercise 2. (Sum of two independent random variables). The goal of this exercise is to study the distribution of sum of two independent random variables.

a Let X, Y be two independent discrete random variables with PMF f_X and f_Y. Prove that the PMF of $Z = X + Y$ is given by

$$f_Z(z) = \sum_u f_X(z-u) f_Y(u). \tag{1}$$

b Now let X, Y be two independent continuous random variables with PDF f_X and f_Y. Prove that the PDF of $Z = X + Y$ is given by

$$f_Z(z) = \int_{u=-\infty}^{u=\infty} f_X(z-u) f_Y(u)\,\mathrm{d}u. \tag{2}$$

which is the convolution of their respective PDFs.
Hint: First derive the CDF of Z,

$$F_Z(z) := \mathbb{P}(Z \leq z).$$

c Use formula (2) to find how $Z = X + Y$ is distributed if $X \sim Exp(\lambda)$ and $Y \sim Exp(\lambda)$ are independent. To illustrate the result, pick some particular $\lambda > 0$. Use `rexp()` in `R` to generate random samples. Create two plots: one with histogram of samples of X and density function of exponential distribution, and the other with histogram of samples Z and the density function that you have found.

d (optional*) Use formula (2) to find how $Z = X + Y$ is distributed if $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent and normally distributed. Compare the result with computation of $\mathbb{E}(Z)$ and $\mathrm{Var}(Z)$.

Exercise 3. Let X be the number of network breakdowns that occur randomly and independently of each other on an average rate of 3 per month.

a Which model would you use to describe the phenomenon? Find the mean and variance of X.

b What is the probability that there will be just 1 network breakdown in a month?

c What is the probability that there will be at least 6 network breakdowns in a month? Use `R` for this computation.

d In part a, you have found the mean and variance of X. Using only this information, apply *Chebyshev's inequality* to obtain a bound for $\mathbb{P}(X \geq 6)$ and compare the result with what you have found in part c.

Exercise 4. The yearly number of car accidents (denoted by X) in a city can be modeled by a Poisson distribution. In a given accident, the probability of a casualty is p. In this exercise, we want to find the distribution of the number of car accidents with casualties (denoted by Y). Let us consider $X \sim Pois(\lambda)$ and $Y|X \sim Binom(X; p)$ conditional upon X.

a Find the joint distribution of X and Y.

b Prove that the marginal distribution of Y is given by $Y \sim Pois(p\lambda)$. (That is, the number of car accidents with casualties is again Poisson but with a smaller parameter.)

c Let $X' \sim Pois(\mu)$ be the yearly number of bicycle accidents, and assume that it is independent of X. Find the distribution of the total number of accidents $X + X'$. *Hint*: use formula (1).

d What is the distribution of the number of bicycle accidents if we know that the total number accidents in a year is k?

Exercise 5. The exponential distribution $Exp(\lambda)$ with rate parameter $\lambda > 0$ is typically used to model the waiting time $X \geq 0$ until the occurrence of a certain event. Then $\mathbb{E}(X) = 1/\lambda$ is the average time until the occurrence of the event of interest (measured in some given unit of time).

A crucial property of the exponential distribution is that it is "*memoryless*": No matter how long you have been waiting already, the probability of waiting for an additional amount of time $s > 0$ only depends on s, and not on your past waiting time $t > 0$. This can be written as

$$\mathbb{P}(X > t + s | X > t) = \mathbb{P}(X > s). \tag{3}$$

Prove identity (3) using the CDF of $X \sim Exp(\lambda)$.

Exercise 6. The **Pearson correlation coefficient** (cf. Def. 6 in Ch. 1.5) of two random variables X and Y (with $\mathbb{E}(X^2), \mathbb{E}(Y^2) < \infty$) is defined to be 0 if $\mathrm{Var}(X) = 0$ or $\mathrm{Var}(Y) = 0$, and otherwise

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)}}.$$

Prove that the Pearson coefficient always satisfies

$$-1 \leq \rho(X, Y) \leq 1,$$

with the equality if and only if there is a *linear relationship* between X and Y. Namely,

$$|\rho(X, Y)| = 1 \iff Y = cX + d,$$

where

$$c = \begin{cases} \sqrt{\frac{\mathrm{Var}(Y)}{\mathrm{Var}(X)}}, & \rho(X, Y) = 1, \\ -\sqrt{\frac{\mathrm{Var}(Y)}{\mathrm{Var}(X)}}, & \rho(X, Y) = -1, \end{cases} \quad , \quad d = \mathbb{E}(Y) - c\mathbb{E}(X).$$

Hint: use the **Cauchy–Schwarz inequality** (cf. Corollary (2) in Ch. 1.4)

$$|\mathbb{E}(XY)| \le \sqrt{\mathbb{E}(X^2)} \cdot \sqrt{\mathbb{E}(Y^2)}$$

for any $X, Y : \Omega \to \mathbb{R}$ (with $\mathbb{E}(X^2)$, $\mathbb{E}(Y^2) < \infty$), whereas the equality holds if and only if $X = aY$ for some constant $a \in \mathbb{R}$.