Solutions to 2nd exam NMAI059 Probability and Statistics 1 – June 22, 2021

1. (10 points) (a) Decide, which of the figures show probability density function of some random variable. For Figures 5 and 6 choose an appropriate value of b, c to make the function a pdf, if possible. Do the next two parts only for those figures, that show a pdf.

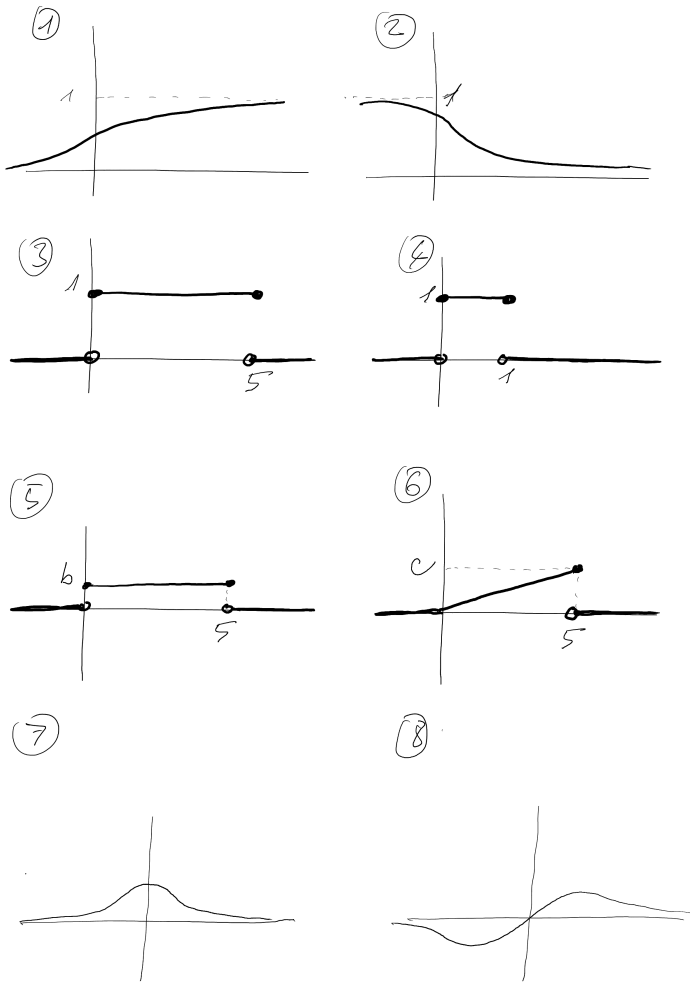
(b) Estimate expectation of the distribution.

(c) Order the distributions by their variance. Do this part only for Figures 3–6, more precisely for those among 3–6 that show a pdf.

(d) Decide which of the images can describe the density of any random variable. For Figures 5 and 6, select an appropriate value of b, c, so that the function is dense, if possible. Only do the other two parts for the images that show the density.

(e) Estimate the mean value of the respective distribution.

(f) Sort the distribution by variance value. Do this section only for images 3–6 or those that show density.



Solution:

(a) We know that a pdf is a non-negative function with the integral across $\mathbb{R}$ equal to 1. Let f_i be the function on the i-th image and X_i the corresponding random variable.

The function f_8 is not non-negative, integrals of $\int_{-\infty}^{\infty} f_1$ and $\int_{-\infty}^{\infty} f_2$ are infinite and $\int_{-\infty}^{\infty} f_3 = 5$. So these four functions cannnot be a pdf.

$\int_{-\infty}^{\infty} f_4 = 1$, so this is a pdf (of the $U(0.1)$ uniform distribution).

$\int_{-\infty}^{\infty} f_5 = 5b$ (we can calculate the integral, or the area of the rectangle) so it is a pdf if we put $b = 1/5$. Clearly X_5 follows the uniform distribution $U(0.5)$.

$\int_{-\infty}^{\infty} f_6 = 5c/2$ (again, we can calculate the integral, or the area of the triangle) so it is a pdf if we put $c = 2/5$.

$\int_{-\infty}^{\infty} f_7$ can't be identified from the image, but it can be one, so it can be a pdf. It is possible that X_7 follows the standard normal distribution.

(b) The median value of $U(a, b)$ is $(a + b)/2$. So $\mathbb{E}(X_4) = 1/2$ and $\mathbb{E}(X_5) = 5/2$. The f_7 function appears to be even, thus $\mathbb{E}(X_7) = 0$ (or possibly $\mathbb{E}(X_7)$ does not exist if it follows e.g. the Cauchy distribution). For $\mathbb{E}(X_6)$, we use a formula from the definition

$$\mathbb{E}(X_6) = \int_{-\infty}^{\infty} x \cdot f_6(x)dx = \int_0^5 x \cdot \frac{2}{25}xdx = \Big[\frac{2}{25}\frac{x^3}{3}\Big]_0^5 = \frac{10}{3}.$$

We could also use geometric knowledge about the centre of gravity of a triangle (the centre of gravity is at a third of the median), with the same result.

(c) If we recall the formula, we know straight away that $\mathrm{var}(X_4) = (1-0)^2/12 = 1/12$ and $\mathrm{var}(X_5) = (5-0)^2/12 = 25/12$. (Even without the formula it should be clear that $\mathrm{var}(X_4)$ is 25-times smaller than $\mathrm{var}(X_5)$.) X_6 seems to be a bit more concentrated than X_5, let us compute it precisely:

$$\mathbb{E}(X_6^2) = \int_{-\infty}^{\infty} x^2 \cdot f_6(x)dx = \int_0^5 x^2 \cdot \frac{2}{25}xdx = \Big[\frac{2}{25}\frac{x^4}{4}\Big]_0^5 = \frac{25}{2}.$$

So the variance is $\mathrm{var}(X_6) = \mathbb{E}(X_6^2) - \mathbb{E}(X_6)^2 = 25/2 - 100/9$. Which is clearly a bit over 1 (exactly $25/18 = 1.388\ldots$). Thus the desired order is $\mathrm{var}(X_4) < \mathrm{var}(X_6) < \mathrm{var}(X_5)$.

---

2. (10 points) There are one hundred balls in the box with numbers 1, 2, ..., 100. We pull out three of them (we do not return them back).

(a) What is the probability that they all have a number at most equal to 40?

(b) What is the expected value of the sum of the numbers on the drawn balls?

(c) What is the expected value of the number of balls drawn whose number is at most equal to 40?

Solution:

(a) We use the chain rule for conditional probability. Let A_i be the event "the i-th ball has number at most 40". Then

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) = \frac{40}{100} \cdot \frac{39}{99} \cdot \frac{38}{98}.$$

(b) Let X_i be the number on the i-th ball. When we ignore the other balls (we simply don't look at the numbers), we see that X_i is a uniformly random number in 1, 2, ..., 100; thus $\mathbb{E}(X_i) = (1 + 100)/2$. The variables X_1, X_2, X_3 *are not* independent: if we get a ball with 1, then we cannot get it another time, which slightly increases the expected number on the other balls. However, this is not a problem for the linearity of expectation. Thus

$$\mathbb{E}(X_1 + X_2 + X_3) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \mathbb{E}(X_3) = 303/2.$$

We might count with the dependency of X_2 on X_1, that is to determine conditional expectation $\mathbb{E}(X_2 \mid X_1 = 1)$, etc. This leads to calculating the average of all sums of 3-element subsets of $\{1, \ldots, 100\}$ using a triple sum. But is is unnecessary and cumbersome.

(c) For simplicity let us color red the balls with numbers 1, ..., 40, So we are interested in R, the number of drawn red balls. This is a textbook example of hypergeometric distribution, by the formula from class we have $\mathbb{E}(R) = 3 \cdot \frac{40}{100} = 1.2$.

If we forget about hypergeometric distribution, we can compute directly by definition

$$\mathbb{E}(R) = 3 \cdot P(R = 3) + 2 \cdot P(R = 2) + 1 \cdot P(R = 1) + 0 \cdot P(R = 0).$$

Part (a) implies that $P(R = 3) = \frac{40}{100} \cdot \frac{39}{99} \cdot \frac{38}{98}$. The other two terms take a bit more work. If $R = 2$, there are three possibilities, which ball was not red, each of them has the same probability. Thus $P(R = 2) = 3 \cdot \frac{40}{100} \cdot \frac{39}{99} \cdot \frac{60}{98}$. Similarly, $P(R = 1) = 3 \cdot \frac{40}{100} \cdot \frac{60}{99} \cdot \frac{59}{98}$. Plugging in the formula for $\mathbb{E}(R)$ gives us the results. There is no need to get a numerical answer during exam, but the result is, of course, the same 1.2.

Then there are other, more tricky ways (with much less computation needed) to compute $\mathbb{E}(R)$, along the lines of how we derived the general formula in class. These express R as a sum of three terms (indicating whether first, second, third ball are red) and use the linearity of expectation.

---

3. (10 points) Oral exam takes time that follows exponential distribution with expected value 20 minutes. Two students are scheduled: one for 10:00, another for 10:20. If the first student examination takes longer than 20 minutes, the second one starts just after the first one finishes. Otherwise, the second one starts at 10:20 sharp.

What is the expected time when the second student finishes?

Solution: Let X be the time it took to examine the first student, Y the second student. We are told that $X, Y \sim Exp(\lambda)$ for $\lambda = 1/20$ (as the expectation is $1/\lambda$). We will treat X, Y as independent.

We divide the calculation depending on whether the second student needs to wait, or not. If not, which happens with probability $P(X \leq 20) = 1 - e^{-\lambda \cdot 20} = 1 - e^{-1}$, then the total time is $S = 20 + Y$ and thus $\mathbb{E}(S \mid X \leq 20) = 20 + \mathbb{E}(Y) = 40$.

If the second student has to wait, then we know that $X > 20$, which occurs with probability e^{-1}. Let us use the fact that exponential distribution is memoryless, thus $X - 20 | X > 20$ follows the same, exponential distribution $Exp(1/20)$. In this case we have

$$\mathbb{E}(S \mid X > 20) = 20 + \mathbb{E}(X \mid X > 20) + \mathbb{E}(Y) = 20 + 20 + 20 = 60.$$

By the law of total expectation we have

$$\mathbb{E}(S) = P(X \leq 20) \cdot \mathbb{E}(S \mid X \leq 20) + P(X > 20) \cdot \mathbb{E}(S \mid X > 20) = (1 - e^{-1}) \cdot 40 + e^{-1} \cdot 60 = 40 + 20/e.$$

This is approximately 47.3. (No need to evaluate this during the exams.)

---

4. (10 points) (a) Define independent events.
(b) Define conditional expectation of a discrete random variable.

Solution:

(a) We say that events A_i, $i \in I$ are independent if for each finite set $S \subseteq I$ we have

$$P(\bigcap_{i \in S} A_i) = \prod_{i \in S} P(A_i).$$

(b) If $X : \Omega \to \mathbb{R}$ is a random variable on probability space $(\Omega, \mathcal{F}, P)$ and $A \in \mathcal{F}$, we define

$$\mathbb{E}(X \mid A) = \sum_{x \in ImX} x \cdot \Pr(X = x | A).$$

---

5. (10 points) Explain hypothesis testing. (In particular explain what Type I and Type II errors are.)

Solution: We want to test validity of a statement, e.g., whether a coin is fair. We let H_0 be the claim we consider unsurprising, "a default". It is called the null hypothesis. The alternative hypothesis is denoted H_1 (in this class we only considered the case of two hypotheses, where H_1 is the negation of H_0).

We create a statistical model of the situation: if we toss a coin n times, we typically assume that the individual tosses are independent and the corresponding model is the binomial distribution $Bin(n, \vartheta)$. Here ϑ is unknown parameter and the null hypothesis says that $\vartheta = 1/2$.

Before measuring the data we decide, what will be the critical region, i.e., the set W of possible measurements for which we reject H_0. We can choose this set in several ways, but we wish to ensure, that the probability of the Type I error (false rejection of the null hypothesis) is not too large – we do not want to claim exciting news whenever the data randomly fluctuate. Our bound on this probability is denoted by α.

At the same time we want to minimize β, the probability of Type II error (false admission, we do not reject H_0 even when it is not valid). Probability that we reject false null hypothesis is called the strength of the test.

---

6. (10 points) State and prove the theorem about convolution formula for the sum of independent random variables, the case of discrete random variables.

Solution:

Theorem:

When X, Y are independent discrete random variables, then for $Z = X + Y$ we have the following formula:

$$P(Z = z) = \sum_{x \in ImX} P(X = x)P(Y = z - x).$$

Proof: By the law of total probability we have

$$P(Z = z) = \sum_{x \in Im(X)} P(X = x) \cdot P(Z = z \mid X = x).$$

From the definition of Z it follows that $P(Z = z \mid X = x) = P(Y = z - x \mid X = x)$. Independence of X and Y implies $P(Y = z - x \mid X = x) = P(Y = z - x)$. Putting all together finishes the proof.