MTH794P Probability and Statistics for Data Analytics — Problem Sheet 2

Kavit Tolia

October 9, 2025

Problem 1.
A continuous random variable X has pdf of the form:

$$f_X(x) = \begin{cases} 0 & \text{if } x < \frac{1}{2}, \\ \frac{C}{x^2} & \text{if } x \geq \frac{1}{2} \end{cases}$$

for some constant C.

(a) What is the value of C?

(b) What is $f_X(\frac{1}{2})$?

(c) Find the cdf of X.

(d) Write down a couple of ways you could check your answer to (c) is plausible. Perform those checks and revisit your answer to (c) if necessary.

(e) How would you calculate $\mathbb{P}(1 < X < 3)$ using the cdf?

(f) How would you calculate $\mathbb{P}(1 < X < 3)$ using the pdf?

Solution.

(a) For $f_X(x)$ to be a pdf, we need:

$$\int_{-\infty}^{\infty} f_X(x)\,dx = 1 \Rightarrow \int_{\frac{1}{2}}^{\infty} \frac{C}{x^2}\,dx = 1 \Rightarrow \left[-\frac{C}{x}\right]_{x=\frac{1}{2}}^{\infty} = 1 \Rightarrow C = \frac{1}{2}$$

(b) $f_X(\frac{1}{2}) = \frac{\frac{1}{2}}{(\frac{1}{2})^2} = 2$

(c) Let us define the cdf of X as $F_X(x)$, then we have:

$$F_x(x) = \int_{-\infty}^{x} f_X(t)\,dt = \int_{-\infty}^{x} \frac{1}{2t^2}\,dt = \left[-\frac{1}{2t}\right]_{t=\frac{1}{2}}^{x} = 1 - \frac{1}{2x}$$

So, we have:

$$F_X(x) = \begin{cases} 0 & \text{if } x < \frac{1}{2}, \\ 1 - \frac{1}{2x} & \text{if } x \geq \frac{1}{2} \end{cases}$$

(d) To check that the cdf is valid, we can do the following checks:

- Non-decreasing: We can see that as $x \geq \frac{1}{2}$ increases, $\frac{1}{2x}$ decreases. This means $F_X(x)$ increases as it is negatively linked to $\frac{1}{2x}$. So the function is indeed non-decreasing.
- Limits at infinity: It's clear that $\lim_{x\to-\infty} F_X(x) = 0$ due to how the function is defined. As for the other side, as x approaches infinity we have $\frac{1}{2x}$ approaching 0, and so $F_X(x)$ approaching 1.

(e) We can calculate $\mathbb{P}(1 < X < 3)$ using the cdf as follows:

$$\mathbb{P}(1 < X < 3) = \mathbb{P}(X \leq 3) - \mathbb{P}(X \leq 1) = F_X(3) - F_X(1) = 1 - \frac{1}{2(3)} - 1 + \frac{1}{2(1)} = \frac{1}{3}$$

(f) If we wanted to do the same thing using a pdf, we can do it by integrating it:

$$\mathbb{P}(1 < X < 3) = \mathbb{P}(1 \leq X \leq 3) = \int_1^3 f_X(x)\,dx = \int_1^3 \frac{1}{2x^2}\,dx = \left[-\frac{1}{2x}\right]_{x=1}^3 = -\frac{1}{6} + \frac{1}{2} = \frac{1}{3}$$

Problem 2.

A continuous random variable X has cdf

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ \frac{x}{6} & \text{if } 0 \leq x < 3, \\ \frac{1}{2} & \text{if } 3 \leq x < 4, \\ \frac{x-2}{4} & \text{if } 4 \leq x < 6, \\ 1 & \text{if } x \geq 6 \end{cases}$$

(a) Find the pdf of X.

(b) Write down the expressions for $\mathbb{E}(X)$ and $\mathrm{Var}(X)$ and compute them.

(c) Describe the distribution this random variable follows in words in a way that would make sense to a non-mathematician.

Solution.

(a) The pdf of X, $f_X(x)$ is defined as the derivative of its cdf. So, we have:

$$f_X(x) = \begin{cases} \frac{1}{6} & \text{if } 0 \leq x < 3, \\ \frac{1}{4} & \text{if } 4 \leq x < 6, \\ 0 & \text{if } x < 0,\ 3 \leq x < 4, \text{ or } x \geq 6 \end{cases}$$

(b) The expectation of X is defined as follows:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_0^3 \frac{x}{6}\,dx + \int_4^6 \frac{x}{4}\,dx = \left[\frac{x^2}{12}\right]_{x=0}^3 + \left[\frac{x^2}{8}\right]_{x=4}^6 = \frac{9}{12} + \frac{36}{8} - \frac{16}{8} = 3.25$$

The variance of X is defined as $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$. We already have $\mathbb{E}(X)$, so we just need to work out $\mathbb{E}(X)^2$:

$$\mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\,dx = \int_0^3 \frac{x^2}{6}\,dx + \int_4^6 \frac{x^2}{4}\,dx = \left[\frac{x^3}{18}\right]_{x=0}^3 + \left[\frac{x^3}{12}\right]_{x=4}^6 = \frac{27}{18} + \frac{216}{12} - \frac{64}{12} = 14.17$$

This means we have the following for $\mathrm{Var}(X)$:

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = 14.17 - (3.25)^2 = 3.60$$

(c) Let's assume X is the waiting time (in minutes) for a bus.
The bus never arrives before 0 minutes or after 6 minutes, and there is a gap between 3 and 4 minutes where it never comes too.

- The bus arrives half the time in the interval between 0 to 3 minutes, with all times in that range being equally likely
- The other half of the time, the bus arrives between 4 and 6 minutes, where all times are equally likely too

Problem 3.

The number $m \in \mathbb{R}$ is a *median* for the random variable X if $F_X(m) = \frac{1}{2}$.

(a) Indicate why every continuous random variable has a median. [This can be an informal explanation rather than a rigorous mathematical proof, although if you know some analysis you could try to write a proof.]

(b) Can you say anything about $f_X(m)$ (i.e. the pdf evaluated at m)?

Solution.

(a) The intuitive explanation here can be as follows:
Given the cdf of X is a function which tells us the probability of X being less than a certain value, there will always be a value at which point the probability of being less than that value is exactly equal to $\frac{1}{2}$. If this was not the case, then the X is not a continuous random variable.

In terms of a mathematical proof, $F_X(x)$ is a continuous function in the interval [0, 1] (as it is a cdf). We can then use the *Intermediate Value Theorem*, which states that if a function f is continuous in the interval [a, b], then for any value k between $f(a)$ and $f(b)$, there much be at least one value c in the interval (a, b) such that $f(c) = k$. In our case, we have $F_X(x)$ which is continuous everywhere, so there will be a number $m \in \mathbb{R}$ such that $F_X(m) = k$, where $k = \frac{1}{2}$.

(b) Given any continuous random variable will have a median, that means it is agnostic of the type of distribution followed by X. It could be a uniform distribution, normal distribution, or any other type. All this means we cannot really say much about the value of the pdf at m.

Problem 4.

Let T be the random variable giving the time (in minutes) between consecutive customers arriving in a shop. Suppose that $T \sim \text{Exp}(0.5)$. Each customer spends 5 minutes in the shop and then leaves. The first customer of the day has just entered the shop.

(a) What is the probability that the next customer does not arrive until after the first customer has left?

(b) What is the expectation of the time before the next customer arrives?

(c) Is the median of the time before the next customer arrives smaller, larger or the same as the expectation?

(d) What is the probability that the time before the next customer arrives is greater than twice its expectation?

(e) How do your answers to (c) and (d) change if the parameter of T changes?

Solution.

(a) We are looking for the probability that $T > 5$:

$$\mathbb{P}(T > 5) = 1 - \mathbb{P}(T \leq 5) = 1 - F_T(5) = 1 - (1 - e^{-\lambda(5)}) = e^{(-2.5)} = 0.08$$

(b) $\mathbb{E}(T) = 1/\lambda = 1/0.5 = 2$

(c) Let m be the median of T, then we have:

$$F_T(m) = \frac{1}{2} \Rightarrow 1 - e^{-\lambda m} = \frac{1}{2} \Rightarrow e^{-\lambda m} = \frac{1}{2} \Rightarrow -\lambda m = \ln\frac{1}{2} \Rightarrow m = \frac{\ln 2}{\lambda} = \frac{\ln 2}{0.5} = 1.39$$

So, the median is less than the expectation, which makes sense given the shape of the cdf. This result is true for all exponential distributions as $\ln 2$ is 0.69. This means the median of an exponential distribution is always 0.69 times the expectation.

(d) Let's work this out for a generic exponential distribution first:

$$\mathbb{P}(T > 2\mathbb{E}(T)) = 1 - \mathbb{P}(T \leq 2\mathbb{E}(T)) = 1 - F_T(2\mathbb{E}(T)) = 1 - (1 - e^{-2\lambda\mathbb{E}(T)}) = e^{-2\lambda(\frac{1}{\lambda})} = \frac{1}{e^2} = 0.14$$

Given this derivation does not depend on λ, it is always true!

(e) No changes to the answers.

Problem 5.

Let Ω be the right-angled triangle with vertices (0,0), (1,0) and (1,1). Let a be a point chosen randomly from within Ω with the probability that a is in any fixed region being proportional to the area of the region. Let T be the random variable "The angle of the line from (0,0) to a makes with the x-axis". Find the cdf and pdf of T.

Solution.
Given it is a right-angled triangle with the provided vertices, we know that the angle between the hypotenuse and the x-axis is $\frac{\pi}{4}$. This means the cdf of the random variable T is 0 if $T < 0$ and 1 if $T > \frac{\pi}{4}$. Now, let's say we want to find the probability that $T \leq \theta$, where $0 \leq \theta \leq \frac{\pi}{4}$. This probability, $\mathbb{P}(T \leq \theta)$, is the same as the probability that a is in the right-angled triangle where the angle between the hypotenuse and the x-axis is θ, which is given to be proportional to the area of the right-angled triangle. The area of this right-angled triangle is $\frac{1}{2}bh$, where $b = 1$ and h can be written as $\tan(\theta)$. So, the area of that sub-region becomes $\frac{1}{2}\tan(\theta)$.

$$\mathbb{P}(T \leq \theta) = \frac{\text{Area of sub-region}}{\text{Area of } \Omega} = \frac{\frac{1}{2}\tan(\theta)}{\frac{1}{2}} = \tan(\theta)$$

This means the cdf can be defined as:

$$F_T(t) = \begin{cases} 0 & \text{if } t < 0, \\ \tan(t) & \text{if } 0 \leq t \leq \frac{\pi}{4}, \\ 1 & \text{if } t > \frac{\pi}{4} \end{cases}$$

Given that the pdf is just the derivative of the cdf, we have:

$$f_T(t) = \begin{cases} 0 & \text{if } t < 0 \text{ or } t > \frac{\pi}{4}, \\ \sec^2(t) & \text{if } 0 \leq t \leq \frac{\pi}{4}, \end{cases}$$

Problem 6.

Decide whether each of the following statements is true or false. Give a reason in either case:

(a) For any continuous random variable X, the values of the pdf are probabilities so $0 \leq f_X(x) \leq 1$ for all $x \in \mathbb{R}$.

(b) For any continuous random variable, the median is always unique (i.e. there is only one $m \in \mathbb{R}$ which satisfies the definition of median).

Solution.

(a) False. The pdf of a random variable does not give us probabilities. It provides us the density (can be thought of as probability per unit of x), which can take any non-negative value. We can see this in (a) of Problem 1.

(b) False. You can have a constant $F_X(x) = \frac{1}{2}$ between two values, which means every value in that interval is a median. The median is only unique if the cdf $F_X(x)$ is strictly increasing (or the pdf $f_X(x) > 0$ everywhere).