Statistics for Data Science - 2

Week 5 Practice Assignment Solutions

1. The probability density function of a continuous random variable X is shown in Figure 5.1.P.

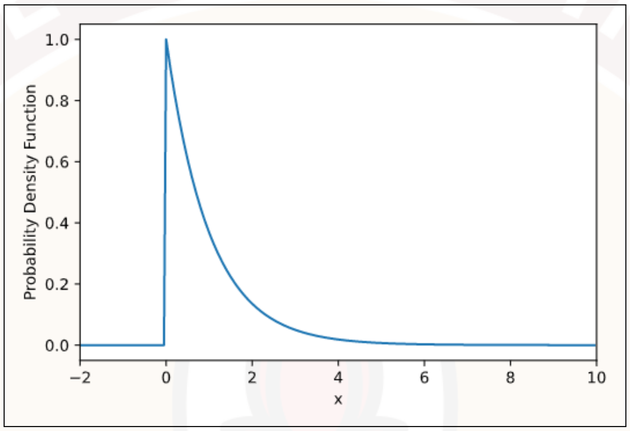

Figure 5.1.P: Probability Density Function graph of X

The PDF is defined as follows:

$$f_X(x) = \begin{cases} e^{-x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Find $P(-\epsilon < X < 0)$, where ϵ is a very small positive number.

(a) e^{ϵ}

(b) 0

(c) $e^{-\epsilon}$

(d) $e^{-2\epsilon}$

Answer: b

Solution:

We know that $P(-\epsilon < X < 0) = \int_{-\epsilon}^{0} f_X(x)dx$

But the value of $f_X(x)$ is zero in the range $-\epsilon$ to zero.

Therefore, $P(-\epsilon < X < 0) = 0$.

Therefore, option b is the correct option.

2. Which of the following statements is/are true for a continuous random variable with PDF $f_X(x)$?

   (a) If $f_X(2) = 2f_X(1)$, then $P(2-\epsilon < X < 2+\epsilon) = 2P(1-\epsilon < X < 1+\epsilon)$ for a small ϵ.

   (b) If $f_X(2) = 2f_X(1)$, then $P(2-\epsilon < X < 2+\epsilon) \approx 2P(1-\epsilon < X < 1+\epsilon)$ for a small ϵ.

   (c) $P(X = x_0) = 0$ for any value of x_0.

   (d) CDF $F_X(x)$ is continuous in the domain $[-\infty, \infty]$.

   Answer: b, c, and d

   Solution:
   Option a: We know that for small ϵ, $P(x-\epsilon < X < x+\epsilon) \propto f_X(x)$.
   Therefore, $P(1-\epsilon < X < 1+\epsilon) \propto f_X(1)$ and $P(2-\epsilon < X < 2+\epsilon) \propto f_X(2)$
   But $P(x-\epsilon < X < x+\epsilon)$ is not exact linear function of $f_X(x)$.
   Therefore when $f_X(2) = 2f_X(1)$, then $P(2-\epsilon < X < 2+\epsilon) \neq 2P(1-\epsilon < X < 1+\epsilon)$ but $P(2-\epsilon < X < 2+\epsilon) \approx 2P(1-\epsilon < X < 1+\epsilon)$
   Hence option a is wrong but option b is correct.
   Option c: The probability at an instant $(P_X(x))$ for a continuous random variable is zero as there is no sudden spike in the CDF function for any value of x. Hence option c is correct.
   Option d: For a continuous random variable CDF is always continuous.

3. If

$$f_X(x) = \begin{cases} \frac{1}{18}(x^2 - 8x + 16) & 1 \le x \le 7 \\ 0 & \text{otherwise} \end{cases}$$

   What is the value of $P(X \le 4)$? Enter the answer correct to one decimal accuracy. $(\int x^a dx = \dfrac{x^{a+1}}{a+1})$
   Answer: 0.5

   Solution:
   $P(X \le 4) = \int_{-\infty}^{4} f_X(x)dx$
   $\Rightarrow P(X \le 4) = \int_{1}^{4} f_X(x)dx$, since $f_X(x) = 0$ for $x < 1$.
   $\Rightarrow P(X \le 4) = \int_{1}^{4} (\frac{1}{18}(x^2 - 8x + 16))dx$
   $\Rightarrow P(X \le 4) = \dfrac{1}{18}(x^3/3 - 8x^2/2 + 16x/1)|_1^4$
   $\Rightarrow P(X \le 4) = \dfrac{1}{18}(4^3/3 - 4*4^2 + 16*4) - \dfrac{1}{18}(1^3/3 - 4*1^2 + 16*1)$
   $\Rightarrow P(X \le 4) = 0.5$

4. If $X \sim \text{Normal}(10, 25)$, what is the value of $E[2X^2]$?

   Answer: 250

   Solutions:

   Given E[X]=10, Var(X)=25

   We know that Var(X)= $E[X^2] - E[X]^2$

   $\Rightarrow E[X^2] = \text{Var}(X) + E[X]^2$

   $\Rightarrow E[X^2] = 25 + 10^2 = 125$

   We know that $E[cX] = cE[X]$, where c is a constant.

   $\Rightarrow E[2X^2] = 2E[X^2]$

   $\Rightarrow E[2X^2] = 2 \times 125 = 250$

5. If $X \sim \text{Normal}(10, 4)$, then what is the value of $P(X \geq 8 | X \leq 9)$? Use the standard normal distribution tables if necessary. Enter the answer up to two decimals accuracy. Use the following CDF values of standard normal distribution.
   $F_Z(-2) = 0.02275, F_Z(-1.5) = 0.06681, F_Z(-1) = 0.15866, F_Z(-0.5) = 0.30854, F_Z(0) = 0.5, F_Z(0.5) = 0.69146, \text{ and } F_Z(1) = 0.84134$

   Answer: 0.485 accepted range 0.48 to 0.49

   Solution:

   Given $\mu = 10, \sigma^2 = 4 \Rightarrow \sigma = 2$

   We need to find $P(X \geq 8 | X \leq 9)$.

   $P(X \geq 8 | X \leq 9) = \dfrac{P(X \geq 8 \cap X \leq 9)}{P(X \leq 9)}$

   $P(X \geq 8 | X \leq 9) = \dfrac{F_X(9) - F_X(8)}{F_X(9)}$

   Converting present normal distribution to standard distribution to get values of $F_X(x)$.

   For $x = 8$, $z = \dfrac{x - \mu}{\sigma} = \dfrac{8 - 10}{2} = -1, \Rightarrow F_X(8) = F_Z(-1)$

   For $x = 9$, $z = \dfrac{x - \mu}{\sigma} = \dfrac{9 - 10}{2} = -0.5, \Rightarrow F_X(9) = F_Z(-0.5)$

   $P(X \geq 8 | X \leq 9) = \dfrac{F_X(9) - F_X(8)}{F_X(9)}$

   $\Rightarrow P(X \geq 8 | X \leq 9) = \dfrac{0.30854 - 0.15866}{0.30854} = 0.485$

6. A random variable X has the following PDF

$$f_X(x) = \begin{cases} 2x & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

   Define $Y = e^X$. What is the PDF $f_Y(y)$ of Y?

(a) $f_Y(y) = \begin{cases} \dfrac{2\log(y)}{y} & 1 \le y \le e \\ 0 & \text{otherwise} \end{cases}$

(b) $f_Y(y) = \begin{cases} \dfrac{\log(y)}{2e^y} & 1 \le y \le e \\ 0 & \text{otherwise} \end{cases}$

(c) $f_Y(y) = \begin{cases} \dfrac{\log(y)}{y} & 1 \le y \le e \\ 0 & \text{otherwise} \end{cases}$

(d) $f_Y(y) = \begin{cases} \dfrac{\log(y)}{e^y} & 1 \le y \le e \\ 0 & \text{otherwise} \end{cases}$

(e) $f_Y(y) = \begin{cases} \dfrac{\log(y)}{2y} & 1 \le y \le e \\ 0 & \text{otherwise} \end{cases}$

Answer: a

Solution:
Given $Y = g(X) = e^X$
$\Rightarrow \log y = x = g^{-1}(y)$
Therefore $g^{-1}(y) = log(y)$
$g(x) = e^x, \Rightarrow g'(x) = e^x$ Since $\dfrac{d(e^x)}{dx} = e^x$
We know that in the range 0 to 1, e^x is monotonic (increasing function).
Therefore, we can use the formula, $f_Y(y) = \dfrac{1}{|g'(g^{-1}(y))|} f_X(g^{-1}(y))$
$g'(g^{-1}(y)) = g'(\log y) = e^{\log y} = y$
$|g'(g^{-1}(y))| = y$ since y is positive in the range $[1, e]$
$f_X(g^{-1}(y)) = f_X(\log y) = 2\log y$
Therefore, $f_Y(y) = \dfrac{1}{y}\log y$
$f_Y(y) = \dfrac{2\log y}{y}$
Hence option a is correct.

Use the following information to answer the questions 7 and 8.

The CDF of random variable X is given below:

$$F_X(x) = \begin{cases} 0 & x \leq 0 \\ 2x^2 & 0 \leq x \leq \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \leq x \leq 1 \\ \frac{x}{2} & 1 \leq x \leq 2 \\ 1 & x \geq 2 \end{cases}$$

Use the following derivative formula:

$$\frac{d(x^a)}{dx} = ax^{a-1}$$

7. Which of the following statements is/are correct?

(a) X is a continuous random variable.

(b) X is a discrete random variable.

(c) The PDF of X is not defined as X is discrete random variable.

(d) The PDF of random variable X is $f_X(x) = \begin{cases} 0 & x \leq 0 \\ 4x & 0 \leq x \leq \frac{1}{2} \\ 0 & \frac{1}{2} \leq x \leq 1 \\ \dfrac{x}{2} & 1 \leq x \leq 2 \\ 0 & x > 2 \end{cases}$

(e) The PDF of random variable X is $f_X(x) = \begin{cases} 0 & x < 0 \\ 2x^2 & 0 \leq x \leq \frac{1}{2} \\ 0 & \frac{1}{2} < x < 1 \\ \frac{x}{4} & 1 \leq x \leq 2 \\ 0 & x > 2 \end{cases}$

Answer: a, d

Solution:

We know that $f_X(x) = \dfrac{d(F_X(x))}{dx}$

Given

$$F_X(x) = \begin{cases} 0 & x \leq 0 \\ 2x^2 & 0 \leq x \leq \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \leq x \leq 1 \\ \frac{x}{2} & 1 \leq x \leq 2 \\ 1 & x \geq 2 \end{cases}$$

$$\Rightarrow f_X(x) = \begin{cases} \dfrac{d(0}{dx} = 0 & x \leq 0 \\ \\ \dfrac{d(2x^2)}{dx} = 4x & 0 \leq x \leq \frac{1}{2} \\ \\ \dfrac{d(\frac{1}{2})}{dx} = 0 & \frac{1}{2} < x \leq 1 \\ \\ \dfrac{d(\frac{x}{2})}{dx} = \frac{1}{2} & 1 < x \leq 2 \\ \\ \dfrac{d(1)}{dx} = 0 & x > 2 \end{cases}$$

Therefore, $f_X(x) = \begin{cases} 0 & x \leq 0 \\ 4x & 0 \leq x \leq \frac{1}{2} \\ 0 & \frac{1}{2} < x \leq 1 \\ \dfrac{1}{2} & 1 < x \leq 2 \\ 0 & x > 2 \end{cases}$

Since, $F_X(x)$ is continuous in the given domain, hence X is a continuous random variable.

8. What is the value of $P(X \geq 1 | X \leq 1.5)$? Enter the answer correct to two decimals accuracy.

   Answer: 0.33, accepted range 0.31 to 0.35

   Solution:

   $P(X \geq 1 | X \leq 1.5) = \dfrac{F_X(1.5) - F_X(1)}{F_X(1.5)} = \dfrac{1.5/2 - 1/2}{1.5/2} = 1/3$

9. The time taken by Rohith to complete a race follows the exponential distribution with expected time of completion of 10 minutes. What is the probability that Rohith takes less than 20 minutes but more than 10 minutes to complete the race? Enter the answer correct to 2 decimals accuracy. ($\int e^{-ax} dx = \dfrac{e^{-ax}}{-a}$)

   Answer: 0.2325, accepted range: 0.23 to 0.235

   Solution:

   Given $E[X] = 10$ minutes.

   We know for a exponential distribution $E[X] = \frac{1}{\lambda}$

   $\Rightarrow \frac{1}{\lambda} = 10$, $\lambda = 0.1$

   For exponential distribution $F_X(x) = 1 - e^{-\lambda x}$

   The probability that athlete takes more than 10 minutes is,

   $F_X(10) = 1 - e^{-0.1 \times 10} = 1 - e^{-1}$

   The probability that athlete takes more than 20 minutes is,

$F_X(20) = 1 - e^{-0.1\times 20} = 1 - e^{-2}$
The probability that athlete takes more than 10 minutes but less than 20 minutes to complete race is $F_X(20) - F_X(10) = e^{-1} - e^{-2} = 0.232$ approximately.

10. The PDFs of random variables $X1, X2, X3, X4$, and $X5$ are shown in Figure 5.2.P. Based on the information, choose the correct option(s) from below.

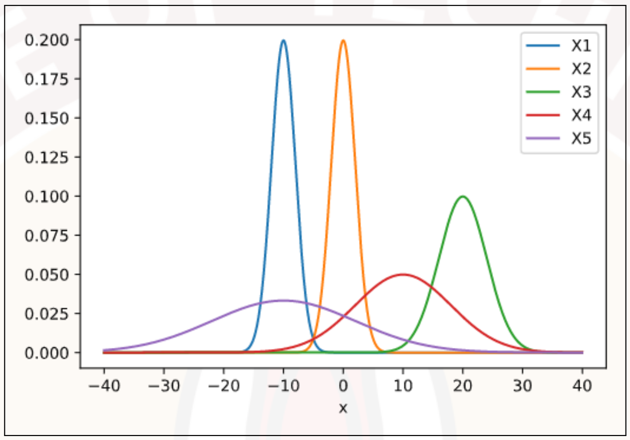

Figure 5.2.P: PDF of Normal Distributions for different variables.

(a) $E(X1) \approx E(X5) < E(X2) < E(X4) < E(X3)$
(b) $E(X1) < E(X5) < E(X2) < E(X4) < E(X3)$
(c) $E(X1) < E(X5) = E(X2) < E(X4) < E(X3)$
(d) $\text{Var}(X1) < \text{Var}(X3) < \text{Var}(X4) < \text{Var}(X5)$
(e) $\text{Var}(X1) \approx \text{Var}(X2) < \text{Var}(X3) < \text{Var}(X4) < \text{Var}(X5)$

Answer: a, d, and e
Solution:
We know that in the PDF of normal distribution, the peak value occurs at mean.
$E[X] = \mu$(mean)
Also, the value of PDF at mean is inversely proportional to standard deviation
Since, $f_X(\mu) = \dfrac{1}{\sqrt{2\pi}\sigma}$.
The peak value, which is mean or $E[X]$, of PDF occurs approximately for $X1, X2, X3, X4$, and $X5$ at -10, 0, 20, 10, and -10 respectively.
Therefore, $E(X1) \approx E(X5) < E(X2) < E(X4) < E(X3)$
The peak value $(f_X(\mu))$ for variables $X1, X2, X3, X4$, and $X5$ are such that $f_{X1}(\mu) \approx f_{X2}(\mu) > f_{X3}(\mu) > f_{X4}(\mu) > f_{X5}(\mu)$.
Therefore, $\text{Var}(X1) \approx \text{Var}(X2) < \text{Var}(X3) < \text{Var}(X4) < \text{Var}(X5)$
Hence, options a, d, and e correct.

11. The PDF of a continuous random variable is given as

$$f_X(x) = \begin{cases} 4x^3 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

What is the value of Var(X)? ($\int x^a dx = \dfrac{x^{a+1}}{a+1}$)

(a) $\dfrac{1}{75}$

(b) $\dfrac{2}{75}$

(c) $\dfrac{3}{75}$

(d) $\dfrac{4}{75}$

Answer: b

We know that $\text{Var}(X) = E[X^2] - E[X]^2$
$E[X] = \int x f_X(x)dx$
$E[X] = \int_0^1 x * 4x^3 dx$
$\Rightarrow E[X] = \int_0^1 4x^4 dx$

$\Rightarrow E[X] = \frac{4x^5}{5}|_0^1$

$\Rightarrow E[X] = \frac{4}{5} - 0 = \frac{4}{5}$

$E[X^2] = \int x^2 f_X(x)dx$
$E[X] = \int_0^1 x * 4x^4 dx$
$\Rightarrow E[X] = \frac{4x^6}{6}|_0^1$

$\Rightarrow E[X] = \frac{4}{6} - 0 = \frac{2}{3}$

Therefore, $\text{Var}(X) = E[X^2] - E[X]^2$

$\text{Var}(X) = \frac{2}{3} - (\frac{4}{5})^2$

$\text{Var}(X) = \frac{2}{3} - \frac{16}{25}$

$\text{Var}(X) = \frac{2}{75}$

12. Let $X \sim \text{Uniform}(a_1, b_1)$ and $Y \sim \text{Uniform}(a_2, b_2)$. Based on this information, choose the correct option(s) from below.

(a) If $b_2 - a_2 = b_1 - a_1$, then $\text{Var}(X) = \text{Var}(Y)$.

(b) If $b_2 + a_2 = b_1 + a_1$, then $\text{Var}(X) = \text{Var}(Y)$.

(c) If $b_2 - a_2 = b_1 - a_1$, then $E(X) = E(Y)$.

(d) If $b_2 - b_1 = a_1 - a_2$, then $E(X) = E(Y)$.

Answer: a and d

Solution:

We know that mean $(E(X))$ and Variance $(\text{Var}(X))$ of uniform random variable ($X \sim \text{Uniform}(a, b)$ is $\frac{a+b}{2}$ and $\frac{(b-a)^2}{12}$ respectively.

Given $X \sim \text{Uniform}(a_1, b_1)$ and $Y \sim \text{Uniform}(a_2, b_2)$, $E(X) = \frac{a_1 + b_1}{2}$, $E(Y) = \frac{a_2 + b_2}{2}$. So, for $E(X)$ to be equal to $E(Y)$, $a_1 + b_1 = a_2 + b_2$ or $b_2 - b_1 = a_1 - a_2$. Hence option d is correct and option c is incorrect.

Similarly for $\text{Var}(X)$ to be equal to $\text{Var}(Y)$, $\frac{(b_1 - a_1)^2}{12} = \frac{(b_2 - a_2)^2}{12}$ or $b_1 - a_1 = b_2 - a_2$, hence option a is correct and option b is incorrect.

13. The CDF of a random variable X is given as:

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \dfrac{x}{\ln 4} & 0 \le x \le \ln 2 \\ 1 - e^{-x} & \ln 2 \le x < \infty \end{cases}$$

Derivative formulas required to solve the problem:

$$\frac{d(ax)}{dx} = a$$

$$\frac{d(e^{-ax})}{dx} = -ae^{-ax}$$

The PDF of the random variable X is:

(a) $f_X(x) = \begin{cases} 0 & x < 0 \\ \dfrac{1}{\ln 4} & 0 \le x < \ln 2 \\ e^{-x} & \ln 2 \le x < \infty \end{cases}$

(b) $f_X(x) = \begin{cases} 0 & x < 0 \\ 1 & 0 \le x < \ln 2 \\ e^{-x} & \ln 2 \le x < \infty \end{cases}$

(c) $f_X(x) = \begin{cases} 0 & x < 0 \\ \dfrac{1}{\ln 2} & 0 \le x \le \ln 2 \\ e^{-x} & \ln 2 < x < \infty \end{cases}$

(d) $f_X(x) = \begin{cases} 0 & x < 0 \\ \dfrac{1}{\ln 2} & 0 \le x < \ln 2 \\ e^x & \ln 2 \le x < \infty \end{cases}$

Answer: a

Solution:

We know that $f_X(x) = \dfrac{d(F_X(x))}{dx}$

Given,

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \dfrac{x}{\ln 4} & 0 \le x \le \ln 2 \\ 1 - e^{-x} & \ln 2 \le x < \infty \end{cases}$$

Therefore,

$$f_X(x) = \begin{cases} \dfrac{d(0)}{dx} = 0 & x < 0 \\ \\ \dfrac{d(\frac{x}{\ln 4})}{dx} = \dfrac{1}{\ln 4} & 0 \le x \le \ln 2 \\ \\ \dfrac{d(1 - e^{-x})}{dx} = e^{-x} & \ln 2 \le x < \infty \end{cases}$$

Hence option a is correct.