DS 122 Homework 2

Xiang Fu
xfu@bu.edu
Boston University Faculty of Computing & Data Sciences

Contents

1 Problem A

1.1 Question

Consider a discrete random variable X with the following probability mass function (pmf):

x	P(X = x)
1	0.2
2	0.3
3	0.5

Find $E[X^3]$.

1.2 Answer

The expectation of a function $g(X)$ for a discrete random variable X is given by:

$$E[g(X)] = \sum_x g(x) * P(X = x)$$

In this case, $g(x) = X^3$, so that we want to find:

$$E[X^3] = \sum_x x^3 * P(X = x)$$

Given the provided probability mass function (pmf):

$$x = 1, 2, 3$$

$$P(X = x) = 0.2, 0.3, 0.5$$

We can calculate $E[X]$ by plugging in the values from the table:

$$E(X^3] = 1^3 x 0.2 + 2^3 * 0.3 + 3^3 * 0.5 = 16.1$$

The expected value $E[X^3]$ for the given probability mass function (pmf) is 16.1.

2 Problem B

2.1 Question

Consider discrete random variables X and Y with the following probability mass function (pmf):

x	y	P(X = x)
1	1	0.1
1	2	0.2
2	1	0.3
2	2	0.4

Find $E[X(Y^2)]$.

2.2 Answer

To find the $E(X(Y^2)$, we need to calculate the expected value of the function $g(X, Y) = \mathrm{XY}^2$ for the given joint distribution of X and Y.

The expectation is given by:

$$E[X(Y^2)] = \sum_{x,y} x(y^2) * P(X = x, Y = y)$$

Given the provided joint probability mass function (pmf), we can calculate $E[X(Y^2)]$ by plugging in the values from the table:

$$E(X(Y^3)] = (1 * 1^2 * 0.1) + (1 * 2^2 * 0.2) + (2 * 1^2 * 0.3) + (2 * 2^2 * 0.4) = 4.7$$

The expected value $E[X(Y^2)]$ for the given joint probability mass function (pmf) is 4.7.

3 Problem C - Understanding Central Limit Theorem

3.1 Question

Consider a random variable

X with the following probability distribution:

Outcomes (x)	Probability (P(X = x)
15	0.3

18	0.5
21	0.2

Suppose you draw 100 samples at random based on this distribution.

3.1.1 Question 4.1

What is the expected value (mean) of the random variable X?

3.1.2 Answer

We can use the formula:

$$E[X] = \sum_i x_i * P(X = x_i)$$

Plug in the values we have on the table:

$$E[X] = (15 \times 0.3) + (18 \times 0.5) + (21 \times 0.2) = 17.7$$

Therfore, the expected value (mean) $E[X]$, of the random variable X for the given probability distribution is 17.7.

3.1.3 Question 4.2

Calculate the variance of the random variable.

3.1.4 Answer

We can use the formula:

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2$$

Find $E[X^2]$:

$$E[X^2] = \left(15^2 * 0.3\right) + \left(18^2 * 0.5\right) + \left(21^2 * 0.2\right) = 317.7$$

Then we can calculate the variance $\mathrm{Var}(X)$ with the calculated $E(X)$:

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2 = 317.7 - (17.7)^2 = 4.41$$

The variance $\mathrm{Var}(X)$ of the random variable X for the given probability distribution is approximately 4.41.

3.1.5 Question 4.3

Using the Central Limit Theorem, find the variance for the average of the 100 samples.

3.1.6 Answer

Given:

- n is the sample size (which is 100).
- $\mathrm{Var}(X)$ is the variance of the original distribution (which we just found to be approximately 4.41).

The variance of the sample mean $\overline{X}$ (average of the samples) is given by:

$$\mathrm{Var}\left(\overline{X}\right) = \frac{\mathrm{Var}(X)}{n}$$

We then substitute our values:

$$\mathrm{Var}\left(\overline{X}\right) = \frac{4.41}{100} = 0.0441$$

The variance for the average of the 100 samples, using the Central Limit Theorem, is approximately 0.0441.

4 Problem D - Exploring Relationships in Data

4.1 Quesiton

You're given two sets of data representing the ages and monthly expenses of a group of individuals:

$$\text{Ages} : \{23, 25, 37, 35\}$$

$$\text{Monthly Expenses (in USD)} : \{1000, 1050, 1150, 1400\}$$

4.1.1 Question 5.1

Compute the covariance between age and monthly expenses.

4.1.2 Answer

The formula for the covariance between two variables X and Y for n data points is:

$$\mathrm{Cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) * (y_i - \overline{y})$$

Given:

$$\text{Ages} : \{23, 25, 37, 35\}$$

$$\text{Monthly Expenses (in USD)} : \{1000, 1050, 1150, 1400\}$$

Calculate the mean of both datasets:

$$\overline{x} = \frac{23 + 25 + 37 + 35}{4} = \frac{120}{4} = 20$$

$$\overline{y} = \frac{1000 + 1050 + 1150 + 1400}{4} = \frac{4600}{4} = 1150$$

Calculate the covariance using the covariance formula for two variables X and Y.

For each data point:

$$(i = 1) \to (23\text{—}30) \times (1000\text{—}1150) = -7 * -150 = 1050$$

$$(\mathrm{i} = 2) \to (25 - 30) \times (1050\text{—}1150) = -5 *\text{—}100 = 500$$

$$(\mathrm{i} = 3) \to (37 - 30) \times (1150\text{—}1150) = 7 * 0 = 0$$

$$(\mathrm{i} = 4) \to (35\text{—}30) \times (1400\text{—}1150) = 5 * 250 = 1250$$

Now, we can sum up the individual products and divide by n:

$$\mathrm{Cov}(X, Y) = \frac{1050 + 500 + 0 + 1250}{4} = \frac{2800}{4} = 700$$

Therefore, the covariance between age and monthly expenses is 700, in which this positive covariance suggests that as age increases, the monthly expenses also tend to increase and vice versa.

4.1.3 Question 5.2

Calculate the correlation coefficient between age and monthly expenses.

4.1.4 Answer

We can use the formula for the correlation coefficient between two variables X and Y, which is given by:

$$r = \frac{\mathrm{Cov}(X, Y)}{s_X * s_Y}$$

Given:

- X (Ages): [23, 25, 37, 35]
- Y (Monthly Expenses in USD): [1000, 1050, 1150, 1400]
- $\mathrm{Cov}(X, Y) = 700$

We need to first calculate the standard deviation of both datasets.

For Ages (X):

$$s_X = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Using the mean of ages $\overline{x}$ = 30, we can have:

$$s_X = \sqrt{\frac{(23 - 20)^2 + (25 - 30)^2 + (37 - 30)^2 + (35 - 30)^2}{4}}$$

$$s_X \approx 6.1644$$

For Monthly Expenses (Y):

$$s_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}$$

Using the mean of ages $\overline{x}$ = 1150, we can have:

$$s_X = \sqrt{\frac{\left(1000-1150\right)^2 + \left(1050-1150\right)^2 + \left(1150-1150\right)^2 + \left(1400-1150\right)^2}{4}}$$

$$s_X \approx 152.7535$$

After all of this, we can calculate the covariance using the formula:

$$r = \frac{\mathrm{Cov}(X, Y)}{s_X * s_Y}$$

$$r = (\frac{700}{6.1644 * 152.7535}$$

$$r \approx 0.7467$$

The correlation coefficient between age and monthly expenses is approximately 0.7467. This is a value is close to 1, indicating a strong positive linear relationship between age and monthly expenses, which means as age increases, monthly expenses also tend to increase, and the relationship is fairly linear.

4.1.5 Question 5.3

Interpret the sign and magnitude of the correlation coefficient. What does it tell you about the relationship between age and monthly expenses for this group?

4.1.6 Answer

1. Signs of Correlation Coefficient: The positive sign of the correlation coefficient $r \approx 0.7467$ indicates that there is a positive linear relationship between age and monthly expenses. In other words, as one variable increases, the other also tends to increase. In this context, as age increases, monthly expenses also tend to increase.
2. Magnitude of the Correlation Coefficient: Given our calculated $r \approx 0.7467$, it falls in the "strong linear relationship" category $(0.7 \leq r < 1)$. This means that age and monthly expenses have a strong positive linear relationship in the provided dataset.
3. Interpretation: For the given group, as individuals get older, their monthly expenses tend to increase. The relationship between age and monthly expenses is not just a casual association but is relatively strong and linear. This could suggest that factors associated with aging, such as lifestyle changes, family responsibilities, or career advancements, might be leading to higher monthly expenses for this group of individuals.

5 Question 6 - Computational Section

See the Jupyter Notebook.

6 Question 7 - Optional Section

6.1.1 Question 7.1

Consider a continuous random variable (X) with the following probability density function (pdf):

$$f(x)\begin{cases} x, \ \text{if} \, 0 \le x \le 1 \\ x-1, \ \text{if} \, 1 \le x \le 2 \\ 0\,, \ \text{if otherwise} \end{cases}$$

Find $E[X^3]$.

6.1.2 Answer

For this, we need to use an integral formula:

$$E[X^3] = \int_{-\infty}^{\infty} x^3 f(x) d(x)$$

Given the pdf:

$$f(x)\begin{cases} x, \ \text{if} \, 0 \le x \le 1 \\ x-1, \ \text{if} \, 1 \le x \le 2 \\ 0\,, \ \text{if otherwise} \end{cases}$$

To evaluate $E[X^3]$, we'll split the integral into two parts based on the ranges of x:

For $0 \le x \le 1 : f(x) = x$:

$$\int_0^1 x^3 * x d(x)$$

For $1 \le x \le 2 : f(x) = x - 1$:

$$\int_1^2 x^3 * (x-1) d(x)$$

Now, we'll evaluate these integrals and sum them up.

The expected value $E[X^3]$ is 53/20.

6.1.3 Question 7.2

Let $((X, Y))$ be a random vector with joint probability density function (pdf) given by:

$$f_{X,Y}(x, y)\begin{cases} 4xy\,, \text{if}\, 0 \leq x \leq 1 \,\text{and}\, 0 \leq y \leq 1 \\ 0\,,\ \text{if otherwise} \end{cases}$$

Find $E[X(Y^2)]$.

6.1.4 Answer

For this, we need to use the double integral formula:

$$E[X(Y^2)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} ((x)y^2) * 4(x)(y)d(x)d(y)$$

Given the joint pdf $f_{X,Y}(x, y)$:

$$f_{X,Y}(x, y)\begin{cases} 4xy\,, \text{if}\, 0 \leq x \leq 1 \,\text{and}\, 0 \leq y \leq 1 \\ 0\,,\ \text{if otherwise} \end{cases}$$

The non-zero part of the function is defined over the region [0, 1] × [0, 1]. Thus, the integration limits will be from 0 to 1 for both x and y.

Substituting in the given pdf, we can get:

$$E[X(Y^2)] = \int_0^1 \int_0^1 ((x)y^2) * 4(x)(y)d(x)d(y) = \frac{1}{3}$$