Indian Institute of Technology Madras

Department of Data Science and Artificial Intelligence

DA5000: Mathematical Foundations of Data Science

Tutorial III - Solutions

Problem

1. If the variance of a Dataset is correctly computed with the formula using $(N-1)$ in the denominator, which of the following option is true? Also, explain the reason for using $(N-1)$ in the denominator.

   (a) Dataset is a sample

   (b) Dataset is a population

   (c) Dataset could be either a sample or a population

   Solution:
   Population Variance:
   When calculating the variance for an entire population, the formula used is: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_\text{i} - \mu)^2$
   Here:

   - N is the total number of data points in the population.
   - x_i represents each individual data point.
   - μ is the mean of the entire population.

   Sample Variance:
   When calculating the variance for a sample (a subset of the population), the formula used is: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_\text{i} - \bar{x})^2$
   Here:

   - n is the number of data points in the sample.
   - x_i represents each individual data point in the sample.
   - $\bar{x}$ is the mean of the sample.

   The key difference here is that the denominator is $n-1$, where n is the number of data points in the sample. The use of $n-1$ instead of n is known as Bessel's correction and it corrects the bias in the estimation of the population variance from a sample.

   When you calculate the variance of a sample, you're using the sample mean $\bar{x}$ as an estimate of the population mean μ. Since $\bar{x}$ is derived from the sample itself, it tends to be closer to the individual data points in the sample than the true population mean μ would be. This leads to an underestimation of the variance when you divide by n because the deviations from the sample mean are generally smaller than the deviations from the population mean. To correct for this underestimation, we use $n-1$ as the denominator, which effectively increases the variance estimate slightly, making it an unbiased estimator of the population variance.

   To prove Bessel's correction mathematically, let's consider sample variance s_n^2 without Bessel's correction:
   $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_\text{i} - \bar{x})^2$
   where n is the sample size and $\bar{x}$ is the sample mean. To see why the sample variance underestimates the population variance, let's compute the expected value of s_n^2:

$E[s_n^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]$
We can expand the term $(x_i - \bar{x})^2$ as follows:

$(x_i - \bar{x})^2 = (x_i - \mu + \mu - \bar{x})^2$

Expanding the square:

$(x_i - \bar{x})^2 = (x_i - \mu)^2 + 2(x_i - \mu)(\mu - \bar{x}) + (\mu - \bar{x})^2$

Taking the expectation:

$E[s_n^2] = \frac{1}{n}\sum_{i=1}^{n}\left[E\left[(x_i - \mu)^2\right] + 2E\left[(x_i - \mu)(\mu - \bar{x})\right] + E\left[(\mu - \bar{x})^2\right]\right]$

The middle term, $2E\left[(x_i - \mu)(\mu - \bar{x})\right]$, vanishes because $E\left[(x_i - \mu)(\mu - \bar{x})\right] = 0$, since $E[x_i - \mu] = 0$.

The last term is $E[(\mu - \bar{x})^2]$ (same as $\left[E\left[(x_i - \mu)^2\right]\right.$), which represents the variance of the sample mean $\bar{x}$.

The expected value then simplifies to:

$E[s_n^2] = \frac{1}{n}\sum_{i=1}^{n}\sigma^2 - \mathrm{Var}(\bar{x})$

It turns out that:

$\mathrm{Var}(\bar{x}) = \frac{\sigma^2}{n}$

Therefore, the expected value of s_n^2 becomes:

$E[s_n^2] = \frac{n-1}{n}\sigma^2$

This shows that when we divide by n, we systematically underestimate the true population variance by a factor of $\frac{n-1}{n}$.

To correct for this underestimation, we use Bessel's correction by dividing by $n - 1$ instead of n. The sample variance with Bessel's correction is:

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Now, the expected value of s^2 is:

$E[s^2] = \frac{n-1}{n-1}\sigma^2 = \sigma^2$

Thus, using $n - 1$ in the denominator makes s^2 an unbiased estimator of the population variance σ^2.

2. Consider a continuous random variable X with probability density function $f(x)$ over the interval $[a, b]$. The variance σ^2 of X can be expressed using the following two formulas:
   1. $\sigma^2 = \int_a^b (x - \mu)^2 f(x) dx$
   2. $\sigma^2 = \int_a^b x^2 f(x) dx - \mu^2$

   Explain how the two formulas are equivalent.

   Solution:
   Formula Breakdown:

   First Part:

$$\sigma^2 = \int_a^b (x - \mu)^2 f(x)\, dx$$

   $(x - \mu)^2$: This term represents the squared deviation of the random variable x from the mean μ.
   $f(x)$: The probability density function (PDF) of the random variable X. This function describes how the probability is distributed across different values of x.
   $\int_a^b$: The integral sums up these squared deviations, weighted by the probability density, over the entire range of x from a to b.

   This expression directly calculates the variance by integrating the squared deviations from the mean across the entire range of x.

Second Part:

$$\sigma^2 = \int_a^b x^2 f(x)\,dx - \mu^2$$

$\int_a^b x^2 f(x)\,dx$: This term represents the expected value of x^2, which is the mean of the squared values of x weighted by the probability density $f(x)$. This is also known as the second moment of X about the origin.
μ^2: This is the square of the mean of X.

This form expresses the variance as the difference between the expected value of the square of X and the square of the expected value of X. It leverages the property of variance:

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

Transition from the First to the Second Formula:

The two formulas are mathematically equivalent. The transition from the first to the second can be understood as follows:

Start with the definition of variance:

$$\sigma^2 = \int_a^b (x-\mu)^2 f(x)\,dx$$

Expand the squared term inside the integral:

$$(x-\mu)^2 = x^2 - 2x\mu + \mu^2$$

Substitute this into the integral:

$$\sigma^2 = \int_a^b (x^2 - 2x\mu + \mu^2) f(x)\,dx$$

Distribute the integral across each term:

$$\sigma^2 = \int_a^b x^2 f(x)\,dx - 2\mu \int_a^b x f(x)\,dx + \mu^2 \int_a^b f(x)\,dx$$

Recognize that:

$$\int_a^b x f(x)\,dx = \mu$$

(this is the definition of the mean, $E[X]$)

$$\int_a^b f(x)\,dx = 1$$

(this is the total probability for a continuous distribution)

So, the equation simplifies to:

$$\sigma^2 = \int_a^b x^2 f(x)\,dx - 2\mu \cdot \mu + \mu^2 \cdot 1$$

Combine like terms:

$$\sigma^2 = \int_a^b x^2 f(x)\,dx - \mu^2$$

This final expression shows that the variance is the difference between the expected value of x^2 and the square of the mean μ.

3. Let X and Y be continuous random variables with joint probability density function $f_{X,Y}(x, y)$. Prove that:

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

, where a and b are constants, and

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

Solution:
Given the continuous random variables X and Y, the expectation of the linear combination $Z = aX + bY$ is given by:

$$\mathbb{E}[Z] = \mathbb{E}[aX + bY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (ax + by) \cdot f_{X,Y}(x, y)\,dx\,dy$$

Distributing the terms inside the integral, we get:

$$\mathbb{E}[aX + bY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} ax \cdot f_{X,Y}(x, y)\,dx\,dy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} by \cdot f_{X,Y}(x, y)\,dx\,dy$$

Factoring out the constants a and b:

$$\mathbb{E}[aX + bY] = a\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \cdot f_{X,Y}(x, y)\,dx\,dy + b\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y \cdot f_{X,Y}(x, y)\,dx\,dy$$

The marginal distribution of X is:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\,dy$$

Similarly, the marginal distribution of Y is:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\,dx$$

Therefore, the expectation simplifies to:

$$\mathbb{E}[aX + bY] = a\int_{-\infty}^{\infty} x \cdot f_X(x)\,dx + b\int_{-\infty}^{\infty} y \cdot f_Y(y)\,dy$$

This is equivalent to:

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

Similarly,

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$$

The expectation of the product XY is:

$$\mathbb{E}[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy \cdot f_{X,Y}(x,y)\, dx\, dy$$

Substituting the joint PDF:

$$\mathbb{E}[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy \cdot f_X(x) \cdot f_Y(y)\, dx\, dy$$

Since the integrals are independent of each other, we can separate them:

$$\mathbb{E}[XY] = \left(\int_{-\infty}^{\infty} x \cdot f_X(x)\, dx \right) \cdot \left(\int_{-\infty}^{\infty} y \cdot f_Y(y)\, dy \right)$$

This gives us:

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

4. Let (X, Y) be continuous random variables with joint PDF:

$$f_{X,Y}(x,y) = \begin{cases} \frac{2}{x^2} & \text{if } 1 < y < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the marginal densities $f_X(x)$ and $f_Y(y)$.

(b) Determine if X and Y are independent.

(c) Calculate the expected value $E[X \cdot Y]$.

Solution:

(a) Marginal densities $f_X(x)$ and $f_Y(y)$:

1. Marginal density of X, $f_X(x)$:

The marginal density $f_X(x)$ is found by integrating the joint PDF over y:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy.$$

From the support of $f_{X,Y}(x,y)$, we know that the joint PDF is non-zero only when $1 < y < x < 2$. So, for a fixed x, y ranges from 1 to x. Thus,

$$f_X(x) = \int_1^x \frac{2}{x^2}\, dy = \frac{2}{x^2}(x-1).$$

The support of X is from 1 to 2, so we have:

$$f_X(x) = \begin{cases} \frac{2(x-1)}{x^2}, & 1 < x < 2, \\ 0, & \text{otherwise.} \end{cases}$$

2. Marginal density of Y, $f_Y(y)$:

The marginal density $f_Y(y)$ is found by integrating the joint PDF over x:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx.$$

From the support of $f_{X,Y}(x,y)$, we know that $1 < y < x < 2$. So, for a fixed y, x ranges from y to 2. Thus,

$$f_Y(y) = \int_y^2 \frac{2}{x^2}\,dx = \left[-\frac{2}{x}\right]_y^2 = \frac{2}{y} - 1.$$

The support of Y is from 1 to 2, so we have:

$$f_Y(y) = \begin{cases} \frac{2}{y} - 1, & 1 < y < 2, \\ 0, & \text{otherwise.} \end{cases}$$

(b) Independence of X and Y: Two random variables are independent if the joint PDF factors into the product of the marginal PDFs, i.e.,

$$f_{X,Y}(x,y) = f_X(x) f_Y(y).$$

In this case, the joint PDF $f_{X,Y}(x,y) = \frac{2}{x^2}$ is non-factorizable into the form $f_X(x)f_Y(y)$, as it does not equal the product of the marginal densities $f_X(x)$ and $f_Y(y)$. Therefore, X and Y are **not independent**.

(c) Expected value $E[X \cdot Y]$:
The expected value $E[X \cdot Y]$ is computed as:

$$E[X \cdot Y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{X,Y}(x,y)\,dx\,dy.$$

Given the support of $f_{X,Y}(x,y)$, this integral simplifies to:

$$E[X \cdot Y] = \int_1^2 \int_1^x xy \frac{2}{x^2}\,dy\,dx.$$

Simplifying the integrand:

$$E[X \cdot Y] = \int_1^2 \int_1^x \frac{2y}{x}\,dy\,dx.$$

First, integrate with respect to y:

$$\int_1^x 2\frac{y}{x}\,dy = \frac{2}{x} \cdot \frac{y^2}{2}\Big|_1^x = \frac{x^2 - 1}{x}.$$

Now, integrate with respect to x:

$$E[X \cdot Y] = \int_1^2 \frac{x^2 - 1}{x}\,dx = \int_1^2 \left(x - \frac{1}{x}\right)\,dx.$$

We compute the two integrals separately:

$$\int_1^2 x\,dx = \frac{x^2}{2}\Big|_1^2 = \frac{4}{2} - \frac{1}{2} = \frac{3}{2},$$

$$\int_1^2 \frac{1}{x}\,dx = \ln(x)\Big|_1^2 = \ln(2).$$

Thus,

$$E[X \cdot Y] = \frac{3}{2} - \ln(2).$$

5. Suppose a company is analyzing the relationship between the number of hours a person spends exercising weekly (denoted by E) and their weight (denoted by W). The joint probability density function (PDF) of the number of hours exercised and weight is given by $f_{E,W}(e, w)$.

$$f_{E,W}(e,w) = \begin{cases} c_1 \cdot e \cdot w & \text{if } 0 \le e \le 5 \text{ and } 30 \le w \le 50 \\ c_2 \cdot e \cdot w^2 & \text{if } 0 \le e \le 5 \text{ and } 50 \le w \le 100 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

(a) Find the values of c_1 and c_2

(b) Determine the PDF which describes the distribution of exercise hours among the population.

(c) Calculate the conditional PDF which represents the weight distribution for people who exercise a given number of hours e.

(d) Find the expected number of hours a person in the range 75-80 kgs exercises.

Solution:

(a) **Find the values of c_1 and c_2:**
To ensure that $f_{E,W}(e, w)$ integrates to 1 over its entire range, we need to solve:

$$\int_0^5 \int_{30}^{50} c_1 \cdot e \cdot w \, dw \, de + \int_0^5 \int_{50}^{100} c_2 \cdot e \cdot w^2 \, dw \, de = 1$$

$$10000 \cdot c_1 + \frac{10937500}{3} \cdot c_2$$

Continuity Condition at $w = 50$:

$$f_{E,W}(e, 50^-) = f_{E,W}(e, 50^+)$$

$$c_1 \cdot e \cdot 50 = c_2 \cdot e \cdot 50^2$$

$$c_1 = 50 \cdot c_2 = 50 \cdot c_2$$

Solving for c_1 and c_2:

$$10000 \cdot (50 \cdot c_2) + \frac{10937500}{3} \cdot c_2 = 1$$

$$c_2 = \frac{3}{12437500}$$

$$c_1 = 50 \cdot \frac{3}{12437500} = \frac{150}{12437500} = \frac{3}{248750}$$

(b) **Determine the PDF which describes the distribution of exercise hours:**
To find the marginal PDF $f_E(e)$, integrate the joint PDF $f_{E,W}(e, w)$ with respect to w:

$$f_E(e) = \int_{30}^{50} c_1 \cdot e \cdot w \, dw + \int_{50}^{100} c_2 \cdot e \cdot w^2 \, dw$$

$$f_E(e) = e \left(800 \cdot c_1 + \frac{875000}{3} \cdot c_2 \right) = 0.08e$$

(c) **Calculate the conditional PDF which represents the weight distribution for people who exercise a given number of hours e:**
The conditional PDF is:

$$f_{W|E}(w|e) = \frac{f_{E,W}(e,w)}{f_E(e)}$$

For $30 \leq w \leq 50$:

$$f_{W|E}(w|e) = \frac{c_1 \cdot e \cdot w}{e \cdot (800 \cdot c_1 + \frac{875000}{3} \cdot c_2)} = \frac{3w}{19900}$$

For $50 \leq w \leq 100$:

$$f_{W|E}(w|e) = \frac{c_2 \cdot e \cdot w^2}{e \cdot (800 \cdot c_1 + \frac{875000}{3} \cdot c_2)} = \frac{3w^2}{995000}$$

(d) **Find the expected number of hours a person in the range 75-80 kg exercises:**
To find the expected value of E for people with W in the range 75-80 kg, we use:

$$\mathbb{E}[E|75 \leq W \leq 80] = \frac{\int_{75}^{80}\int_{0}^{5} e \cdot f_{E,W}(e,w)\,de\,dw}{\int_{75}^{80}\int_{0}^{5} f_{E,W}(e,w)\,de\,dw}$$

$$\mathbb{E}[E|75 \leq W \leq 80] = \frac{\frac{125}{3} \cdot c_2 \cdot w^2}{\frac{25}{2} \cdot c_2 \cdot w^2} = \frac{10}{3} hours$$

6. Suppose two persons, A and B have a meeting at a given time and each will arrive at the meeting place with a delay between 0 and 2 hours, with all pairs of delays being equally likely. The first to arrive will wait for 20 minutes and leave after that if the other hasnt arrived. What is the probability that they will meet?

Solution:

Let X and Y be the arrival times of persons A and B, respectively, with both X and Y uniformly distributed over the interval $[0, 2]$. The total area of possible outcomes is the area of the square in the XY-plane with side length 2, so the total area is:

$$\text{Total area} = 2 \times 2 = 4.$$

The two people will meet if the absolute difference in their arrival times is less than or equal to 20 minutes, or $\frac{1}{3}$ hours. This condition is represented by the inequality:

$$|X - Y| \leq \frac{1}{3}.$$

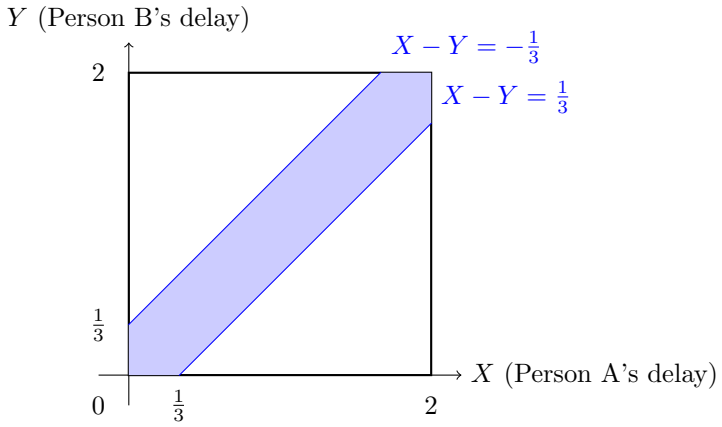

Figure 1: The shaded region represents the area where A and B will meet.

The area of the favorable region (blue shaded region in 1) is the area between the two lines $X - Y = 1/3$ and $Y - X = 1/3$. Thus, the area of the favorable region is:

$$\text{Favorable area} = 4 - (\frac{1}{2} \times \frac{5}{3} \times \frac{5}{3}) \times 2 = \frac{11}{9}$$

Therefore, the probability that the two people will meet is the ratio of the favorable area to the total area:

$$P(\text{meet}) = \frac{\text{Favorable area}}{\text{Total area}} = \frac{\frac{11}{9}}{4} = \frac{11}{36}.$$

Thus, the probability that they will meet is $\frac{11}{36}$.