

PROBABILITY THEORY FOR DATA SCIENCE

Prof. Ishapathik Das

Department of Mathematics and Statistics

Indian Institute of Technology Tirupati


Week - 03

Lecture - 15

Moments

Now, we will discuss moments. We know that whenever we collect data, we often ask for summary statistics to understand it better, like averages and variability, including variance. For example, when students receive their marks, they usually ask for the average score to gauge how their performance compares to others. However, the average alone is not sufficient to understand the data comprehensively; many other measures are also necessary. So, when collecting data, we might have values like x_1, x_2, \dots, x_n , representing n data points.

Moment




A. Mean:
The mean (or expected value) of a r.v. X , denoted by μ_x or $E(X)$, is defined by

$$\mu_x = E(X) = \begin{cases} \sum_{-\infty}^{\infty} x_i p_X(x_i) & X: \text{discrete} \\ \int_{-\infty}^{\infty} x f_X(x) dx & X: \text{continuous} \end{cases}$$

B. Moment:
The n th moment of a r.v. X is defined by

$$E(X^n) = \begin{cases} \sum_{-\infty}^{\infty} x_i^n p_X(x_i) & X: \text{discrete} \\ \int_{-\infty}^{\infty} x^n f_X(x) dx & X: \text{continuous} \end{cases}$$

Note that the mean of X is the first moment of X .



We assume this data comes from a specific distribution, even though we may not know what that distribution is. It's part of a phenomenon we want to estimate or analyze. There is a certain probability associated with each data point, indicating the likelihood of obtaining each x_i . If the data is discrete, we can assign probabilities to these values. Since we usually do not know these probabilities, we calculate the average, known as the sample mean.

This is computed as $(1/n) \sum_{i=1}^n x_i$, where the sum of all observed values is expressed as $(x_1 + x_2 + \dots + x_n)/n$. Here, $(1/n)$ is the multiplier for the total sum of the observed values. What is the $(1/n)$ here? It seems like we are considering some kind of weight, and these weights are the same for n elements. If the data contains multiple observations of the same value, suppose x_1 occurs m_1 times, x_2 occurs m_2 times, and x_n occurs m_n times, then what will be the average?

The image contains handwritten mathematical work and a video frame of a lecturer. The handwritten work includes:

- A piecewise function: $f_X(x) = \begin{cases} \frac{1}{x^2}, & 1 < x < \infty \\ 0, & \text{otherwise} \end{cases}$
- A probability calculation: $P(2 < X < 3) = \int_2^3 f(x) dx = \int_2^3 \frac{1}{x^2} dx = -\frac{1}{x} \Big|_2^3 = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}$
- A diagram showing a set of values $\{x_1, x_2, \dots, x_n\}$ with arrows pointing to a probability $P(X=x_i)$.
- A formula for the sample average: $\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n) = \frac{1}{n}x_1 + \frac{1}{n}x_2 + \dots + \frac{1}{n}x_n$

Below the handwritten work is a video frame showing a man in a light-colored shirt sitting at a desk, looking down at a book or paper.

If you take this sum, it will include repeated values, so for x_1 , it will appear as $x_1 + x_1$, repeated m_1 times, plus x_2 , repeated m_2 times, and so on until x_n , which will appear m_n times. The total will be $m_1 + m_2 + m_n$. So, the sample average will look like this: $(m_1 * x_1 + m_2 * x_2 + \dots + m_n * x_n) / n$. Now, if you express this in a different way, it is $(m_1/n) * x_1 + (m_2/n) * x_2 + (m_n/n) * x_n$.

Now, what are those weights? They represent some kind of probability. For example, if you think of a box containing balls of different colors, say x_1 appears m_1 times and x_2 appears m_2 times, the probability that you pick a ball of color x_1 would be m_1/n , where n is the total number of balls. This is essentially p_1 . So, the average we calculate here reflects the probability of each observation occurring. When we have unique observations, we usually assume all x values have the same probability. But if you know the exact probabilities according to the distribution, we denote these as $p(x_i)$ for each value x_i . When we talk about the expected value or average in this context, we refer to it as the population mean.

The population mean considers all data points and their respective probabilities. For a discrete distribution, the population mean is defined as the sum of each possible value x_k

multiplied by its probability $p(x_k)$, summed over all k . If k is finite, it will be a finite sum; if infinite, it must be absolutely summable; otherwise, the mean does not exist. That is how the population mean is defined. We will discuss here, in general, let x be a discrete random variable with a probability mass function, denoted as $p(x)$.

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

$$P(2 < X < 3) = \int_2^3 f_X(x) dx = \int_2^3 \frac{1}{x^2} dx$$


$$= -\frac{1}{x} \Big|_2^3 = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}$$


$\{x_1, x_2, \dots, x_n\}$
 $\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n)$
 $= \frac{1}{n}x_1 + \frac{1}{n}x_2 + \dots + \frac{1}{n}x_n$

Sample average

$P(X=x_k) = P_x(x_k) = p_k$
 $\bar{x} = \sum_k x_k P_x(x_k)$

Population mean

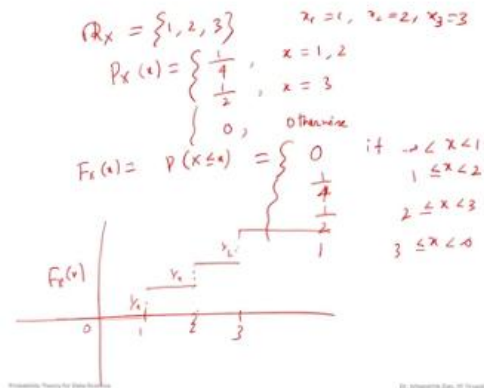




This applies whenever $x \in \text{range}(x)$, which could be finite, such as x_1, x_2, \dots, x_m , or it could be countably infinite. Now, what is the mean of x ? When we refer to the mean, we are talking about the population mean of x , which is defined as the expected value of x . This is denoted by $E(x)$ and sometimes by μ_x . In some books, it may be denoted as μ'_x ; this is not a derivative because there is another concept called central moments. Therefore, we denote it as μ'_x or M_x . The population mean is calculated as the summation $\sum (x_k * p(x_k))$, where k sums over all possible values of x when x is discrete.

Now, if x is a continuous random variable, we know that the probability at a specific point is 0. In this case, the mean is defined analogously to the discrete case, where the summation is replaced by integration. Here, x_k is multiplied by the probability density function, which is defined as $f(x)$. Thus, the mean for continuous random variables is given by the integral \int from $-\infty$ to ∞ of $x * f(x) dx$. This gives us the mean for continuous random variables.

We will discuss some examples, so let us return to the discrete random variables we covered earlier. Suppose we have a discrete random variable, and let's consider this example. The range is $\{-1, 0, 1\}$, with a probability mass function $p(X) = 1/3$ for $X \in \{-1, 0, 1\}$.



In this case, what will be the mean of this random variable? Let X be a discrete random variable with the probability mass function $P_X(X) = 1/3$ whenever $X \in \{-1, 0, 1\}$, and $P_X(X) = 0$ otherwise. The values of X are $X_1 = -1$, $X_2 = 0$, and $X_3 = 1$, which forms a uniform distribution because all the probabilities are the same. To find the expected value $E(X)$, we use the definition, which states that $E(X) = \sum (X_k * P_X(X_k))$. This gives us:

$$E(X) = (-1) * P_X(-1) + 0 * P_X(0) + 1 * P_X(1).$$

Since $P_X(-1) = 1/3$, we have:

$$E(X) = (-1) * (1/3) + 0 + 1 * (1/3).$$

Therefore, $E(X) = -1/3 + 0 + 1/3$, which equals 0. The mean is 0.

Now, let us discuss how to find the mean of a continuous random variable. Suppose X is a continuous random variable with the probability density function given by an example we discussed earlier.



Let X be a discrete random variable with PMF $P_X(x)$, $x \in \mathcal{R}_X = \{x_1, x_2, \dots\}$. The mean of X is defined by

$$\mu_X = E(X) = m_X = \begin{cases} \sum_k x_k P_X(x_k) & : X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f_X(x) dx & : X \text{ is continuous} \end{cases}$$

Let X be a discrete random variable with PMF

$$f_X(x) = \begin{cases} \frac{1}{3} & ; x \in \{-1, 0, 1\} \\ 0 & , \text{ otherwise} \end{cases}$$

$$m_X = E(X) = \sum x_k P_X(x_k) = (-1) \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3} = (-1) \times \frac{1}{3} + 0 + 1 \times \frac{1}{3} = 0$$


The function is $f_X(x) = (3x^2)/28$ whenever $x \in [-1, 3]$, and $f_X(x) = 0$ otherwise. To find the expected value $E(X)$, we will use the definition. This means we need to evaluate the integral from $-\infty$ to ∞ of $x * f_X(x) dx$. Since the density is 0 outside of the interval $[-1, 3]$, we can limit our integral to that range. Thus, we have:

$$E(X) = \int_{-1}^3 x * (3x^2/28) dx.$$

Now we calculate:

$$E(X) = (3/28) * \int_{-1}^3 x^3 dx.$$

This integral can be computed as follows:

$$E(X) = (3/28) * [x^4/4] \text{ from } -1 \text{ to } 3.$$

Evaluating this gives:

$$E(X) = (3/28) * [(3^4/4) - ((-1)^4/4)] = (3/28) * [(81/4) - (1/4)] = (3/28) * (80/4) = (3 * 20)/28 = 60/28 = 15/7.$$

Thus, the mean of this continuous random variable is $15/7$.

Now, let us consider two data sets, two random variables. For example, suppose we have a uniform random variable. We have already considered one random variable with the probability mass function $P_X(x) = 1/3$ for $X \in \{-1, 0, 1\}$. Let us take another similar type of random variable but with a different range. Here, let Y be a random variable, and suppose the range of Y is $\{-10, 0, 10\}$. The probability mass function $P_Y(Y) = 1/3$ whenever $Y \in \{-10, 0, 10\}$, and $P_Y(Y) = 0$ otherwise.



Let x be a discrete random variable with PMF $P_X(x)$, $x \in \mathcal{R}_X = \{x_1, x_2, \dots\}$. The mean of x is defined by

$$\mu_X = E(x) = m_X = \begin{cases} \sum_k x_k P_X(x_k) & : x \text{ is discrete} \\ \int_{-\infty}^{\infty} x f_X(x) dx & : x \text{ is continuous} \end{cases}$$

Let x be a discrete random variable with PMF

$$P_X(x) = \begin{cases} \frac{1}{3} & ; x \in \{-1, 0, 1\} \\ 0 & , \text{ otherwise} \end{cases}$$

$$m_X = E(x) = \sum x_k P_X(x_k) = (-1) \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3} = (-1) \times \frac{1}{3} + 0 + 1 \times \frac{1}{3} = 0$$



Note that this random variable takes different values compared to the random variable X we discussed earlier, which had a range of $\{-1, 0, 1\}$. Now, if we calculate the mean of this random variable, the expected value of Y , denoted as $E(Y)$ or μ_Y , is defined by the formula:

$$E(Y) = \sum (y_k * P(Y = y_k)) \text{ over } k.$$

Here, Y takes three values: $-10, 0$, and 10 . So, we can compute the expected value as follows:

$$E(Y) = (-10 * P(Y = -10)) + (0 * P(Y = 0)) + (10 * P(Y = 10)).$$

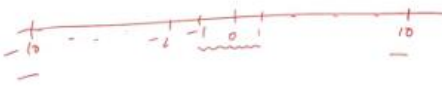
Since $P(Y = -10)$, $P(Y = 0)$, and $P(Y = 10)$ are all equal to $1/3$, we get:

$$E(Y) = (-10 * 1/3) + 0 + (10 * 1/3) = -10/3 + 0 + 10/3 = 0.$$

Therefore, both random variables, X and Y , have the same mean of 0 . This demonstrates that you can have two data sets—one with values $\{-1, 0, 1\}$ and another with values $\{-10, 0, 10\}$ —that can have the same mean despite being situated differently on the number line. For example, one class of students might have marks that are very close to zero, while the other class has marks ranging from -10 to 10 .

So, the key point here is that knowing the mean of a random variable is not enough to understand the data. The mean gives a measure of central tendency, representing the average behavior of the data. However, different distributions can have the same mean, which may not fully capture the underlying characteristics of the data. We need to learn some summarized measures of data.

In general, if we consider another concept here, it's called a function of a random variable. Suppose $Y = g(X)$, where g is a function. We have already been using f for the density function, so let's use g for this example. For instance, g could be $\sin(X)$ or simply $X + 2$.

$$\begin{aligned}
 \mathcal{R}_Y &= \{-10, 0, 10\} \\
 P_Y(y) = P(Y=y) &= \begin{cases} \frac{1}{3}, & y \in \{-10, 0, 10\} \\ 0, & \text{otherwise.} \end{cases} \\
 \mu_Y^1 = E(Y) = \sum_k y_k P_Y(y_k) &= (-10) \times \frac{1}{3} + 0 \times \frac{1}{3} + 10 \times \frac{1}{3} \\
 &= (-10) \frac{1}{3} + 0 + 10 \times \frac{1}{3} = 0
 \end{aligned}$$




Now, if we define this function g as a mapping from \mathbb{R} to \mathbb{R} and we make X a random variable, we can examine how this impacts our understanding of $g(X)$. Since X is a measurable function from a sample space S to \mathbb{R} , we can apply another measurable function g from \mathbb{R} to \mathbb{R} . When we consider the composite function $g(X)$, for any point s in our sample space, $X(s)$ will yield a real number. Thus, $g(X(s))$ will also be a random variable, denoted as Y , which is defined by $g(X)$. Therefore, Y is a measurable function, as long as g is a defined function from \mathbb{R} to \mathbb{R} .

Next, if we want to determine the distribution of Y given the distribution of X , we will not delve into the exact distribution of Y here. Instead, we will focus on finding the mean of Y through this transformation. The expected value of Y , denoted $E(Y)$, can be defined in two cases: for a discrete random variable, it is the summation of y_k for all k , weighted by their probabilities. By definition, this expected value is

$$E(Y) = \sum (y_k * P(Y = y_k)).$$

Thus, the expected value of $g(X)$ will be defined for discrete cases as

$$E[g(X)] = \sum (g(X_k) * P(X = X_k)).$$

For continuous cases, it is defined as

$$E[g(X)] = \int g(X) * f_X(X) dX,$$

where $f_X(X)$ is the density function. Why define this? This is a general definition. The expected value of Y , which we will revisit later, will align with the definitions we established. We are not focusing on finding the probability mass function in the discrete case or the probability density function in the continuous case.

Instead, we are establishing that when we take a transformation like $Y = g(X)$, we can directly find the expected value of $g(X)$. To illustrate this intuitively, consider an experiment involving flipping a coin, where the probability of heads is 0.8 and tails is 0.2. Let's define a random variable X such that $X = 1$ if heads are observed and $X = 0$ if tails are observed. In this case, the probability mass function of X is

$$P(X = 1) = 0.8 \text{ and } P(X = 0) = 0.2.$$

Now, if we take a transformation where $Y = g(X) = X + 2$, we find that if $X = 1$, then $Y = 1 + 2 = 3$, and if $X = 0$, then $Y = 0 + 2 = 2$.

Thus, the probability that $Y = 3$ is the same as the probability that $X = 1$, which is 0.8. Conversely, the probability that $Y = 2$ corresponds to the probability that $X = 0$, which is 0.2. Now, if you want to find the expected value of Y directly from the distribution of Y , you can see that the probability mass function of Y , $P(Y = y)$, is based on the probabilities assigned to the values of Y . For example, if $Y = 3$, the probability is 0.8, which corresponds to observing heads. So, when $Y = 3$, it indicates that heads were observed.

$$E(Y) = \sum_k x_k P_Y(x_k) \quad Y = g(x)$$

$$E(Y) = E(g(x)) = \begin{cases} \sum_k g(x_k) P_X(x_k) & : x \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x) f_X(x) dx & : x \text{ is continuous} \end{cases}$$

$$P(X=1) = P(H) = 0.8$$

$$P(X=0) = P(T) = 0.2$$

$$Y = g(x) = x + 2$$

$$\text{if } x=1, Y = 2+1 = 3$$

$$\text{if } x=0, Y = 0+2 = 2$$

$$\Rightarrow P(Y=3) = P(X=1) = 0.8$$

$$P(Y=2) = P(X=0) = 0.2$$



On the other hand, if tails are observed, $Y = 2$. Therefore, $Y = 3$ means heads were observed, while $Y = 2$ means tails were observed. So, if you want to find the expected values of y directly from the distribution of y , you can see that the distribution of y , which is the probability mass function, is defined as $p(y) = 0.8$ when $y = 3$. This corresponds to the probability of heads, which is 0.8. So, in this case, when $y = 3$, it means that a head has been observed, while when $y = 2$, it represents the tail observed.

$$E(y) = \sum_k y_k P_y(y_k) \quad Y = g(x)$$

$$E(y) = E(g(x)) = \begin{cases} \sum_k g(x_k) P_x(x_k) & : x \text{ is discrete} \\ \int g(x) f_x(x) dx & : x \text{ is continuous} \end{cases}$$

$P(x=1) = P(H) = 0.8$
 $P(x=0) = P(T) = 0.2$
 $\Rightarrow P(y=3) = P(x=1) = 0.8$
 $P(y=2) = P(x=0) = 0.2$

$Y = g(x) = x + 2$
 if $x=1$, $Y = 2+1 = 3$
 if $x=0$, $Y = 0+2 = 2$

© Copyright 2004 by NPTEL. All rights reserved. NPTEL

Thus, y of head is 3 and y of tail is 2. This means that when $y = 3$, the probability remains the same. Next, what will be the expected values of y ? The expected value of y will be calculated as the summation of $y_k * p(y_k)$. In this case, the summation involves the values of y_k , which are 3 and 2.

Therefore, the calculation is as follows: $E(y) = (3 * 0.8) + (2 * 0.2) = 2.4 + 0.4 = 2.8$. Now, if you find the expected values using the formula for expected values of $g(x)$, this is defined as the summation of $g(x_k) * p(x_k)$. Here, we keep the probability mass function of x in consideration. The values that x can take are 0 and 1. So, we have: $E[g(x)] = g(0) * p(x=0) + g(1) * p(x=1)$. Now, $g(0)$ is defined as $x + 2$.

Therefore, when $g(0) = 0 + 2 = 2$, we multiply this by the probability of $x = 0$, which is 0.2. For $g(1)$, we have: $g(1) = 1 + 2 = 3$. The probability $p(x = 1)$, when $x = 1$ (which corresponds to observing a head), is 0.8. Putting this all together: $E[g(x)] = (2 * 0.2) + (3 * 0.8) = 0.4 + 2.4 = 2.8$. Thus, the formula we can use here is quite effective instead of finding the details of the distribution. Although we are focused on finding the mean in this case, if we need to determine the actual distribution function of y , we will need to investigate the distribution function in detail.

We will cover this in more detail later, but for now, it is a straightforward example involving heads and tails. That is why it is easy to find the expected values, and we can verify this.

$$\begin{aligned} E(Y) &= \sum y_k P_Y(y_k) & P_Y(y) &= \begin{cases} 0.8 & \text{if } Y=3 \\ 0.2 & \text{if } Y=2 \end{cases} \\ &= 3 \times 0.8 + 2 \times 0.2 \\ &= 2.4 + 0.4 = 2.8 \end{aligned}$$

$g(x) = x+2$

$$\begin{aligned} E(g(x)) &= \sum_k g(x_k) P_X(x_k) \\ &= g(0) \times P_X(0) + g(1) \times P_X(1) \\ &= 2 \times 0.2 + 3 \times 0.8 \\ &= 0.4 + 2.4 = 2.8 \end{aligned}$$

