PROBABILITY THEORY FOR DATA SCIENCE

Prof. Ishapathik Das

Department of Mathematics and Statistics

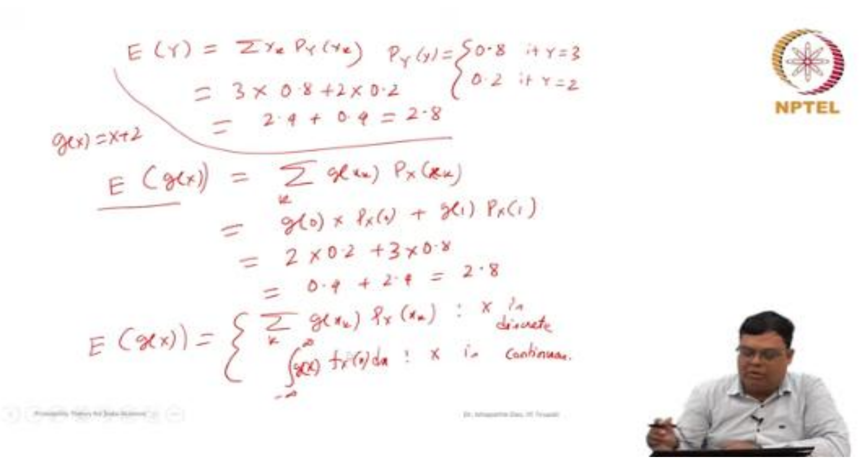
Indian Institute of Technology Tirupati

Week - 04

Lecture - 16

Higher Order Moments and Variance of a Random Variable

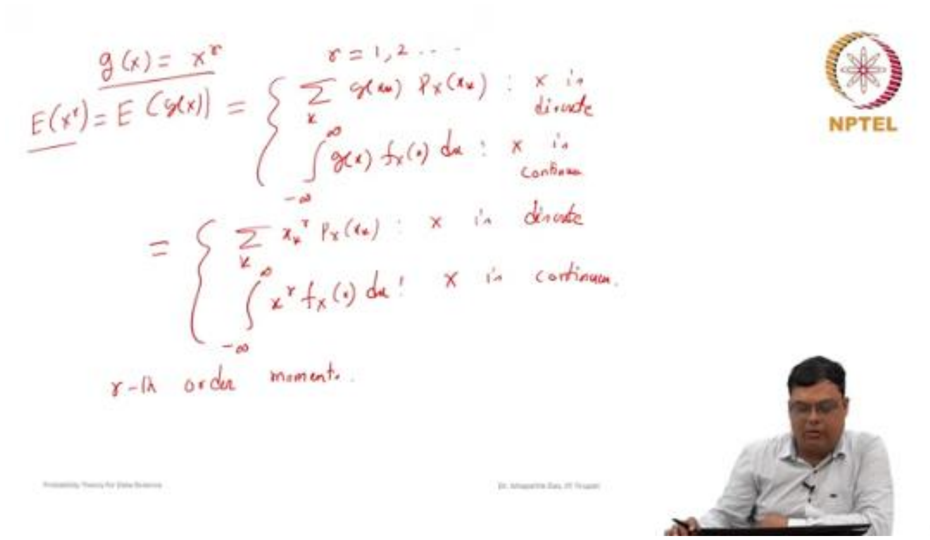
Now, we have this general formula for expected value. The expected value of $g(X)$ is $\Sigma (g(X_k) * p(X_k))$ when X is discrete, and $\int (g(X) * f(X)\, dX)$ when X is continuous. So, this should be $g(X) - 2$ because we're finding the expected value of $g(X)$ with respect to X. That's why it will involve integration over X. So, now, what is the advantage of knowing this distribution and the expected values of $g(X)$?



For example, suppose we want to find $g(X) = X^r$, where r could be any positive integer. Then, what will be the expected value of $g(X)$? By definition, it's basically just the expected value of X^r. So, $g(X)$ is nothing but $\Sigma (X_k^r * P(X = X_k))$ over k when X is discrete, and $\int$ from $-\infty$ to $+\infty$ of $X^r * f(X)\, dX$ when X is continuous. Now, if we set $g(X) = X^r$, then the formula becomes $\Sigma (X_k^r * P(X = X_k))$ over k for discrete X, and $\int$ from $-\infty$ to $+\infty$ of $X^r * f(X)\, dX$ when X is continuous.

This is called the expectation of X^r, or the r-th order moment. So, we could define the r-th order moments directly with this formula. On this slide, you can see that the r-th order moment has the mean and the moment itself. The n-th order moment of a random variable X is defined as the expectation of X^n, with $\Sigma (X_k^n * P(X = X_k))$ when X is discrete, and $\int$ from $-\infty$ to $+\infty$ of $X^n * f(X)\, dX$ when X is continuous. Now, I find it helpful to represent this with g(X) because, next, we'll define variance. For finding variance, we won't need to worry about this setup.

So, this is the r-th order moment. We will discuss some of the r-th order moments, along with some numerical examples. Now, this moment is called a raw moment. There is another type, called a central moment. First, we define the r-th order moment around a point, say, point a, where a is a real number.

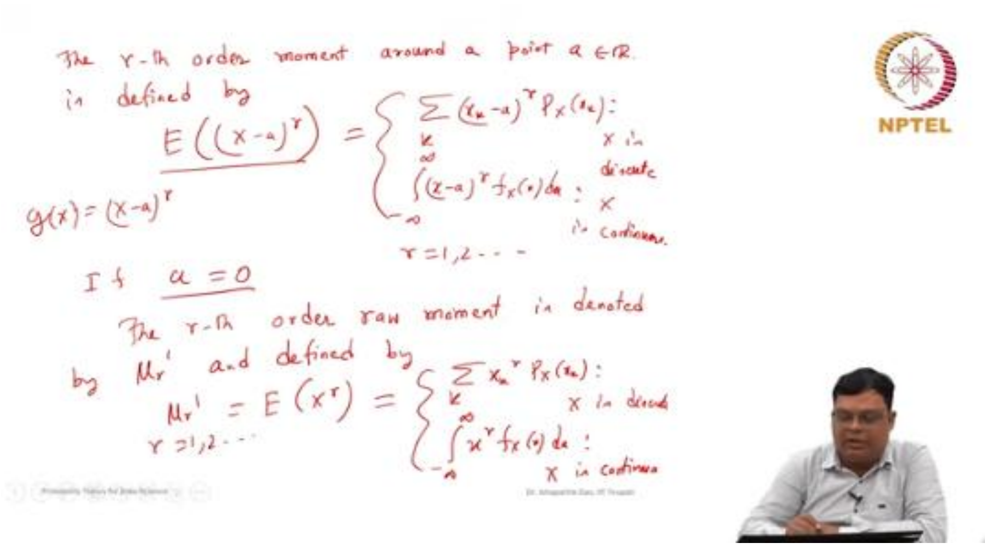


The r-th order moment around a point a is defined by the expected value $E[(X - a)^r]$. For this, if we take $g(x) = (x - a)^r$, the general form for $E[g(X)]$ when X is a discrete random variable is $E[g(X)] = \Sigma\, g(x_k) * P(X = x_k)$. Therefore, in this case, $g(x_k) = (x_k - a)^r * P(X = x_k)$.

When X is a continuous random variable, $E[g(X)]$ is given by the integral $\int g(x) * f(x)\, dx$ from minus infinity to plus infinity. Specifically, with $g(x) = (x - a)^r$, this becomes $E[(X - a)^r] = \int (x - a)^r * f(x)\, dx$ from minus infinity to plus infinity. This defines the r-th order moment, where r can take values such as 1, 2, and so on. For example, when r = 1, this is the first-order moment.

If we set a = 0, the moment is called the r-th order raw moment, often denoted by μ_r' and is defined as $E[X^r]$. With a = 0, this simplifies to X^r, giving $E[X^r]$ as the formula.

In summary, the r-th order moment around a point a follows this general form. When $a = 0$, we get the r-th order raw moment, represented as $E[X^r]$. For discrete X, this raw moment is calculated by $\Sigma x_k^r * P(X = x_k)$, and for continuous X, it is $\int x^r * f(x)\,dx$ from minus infinity to plus infinity. This approach applies for any $r = 1, 2$, and so on. For example, μ_1', or the first raw moment, is simply the expected value of X.

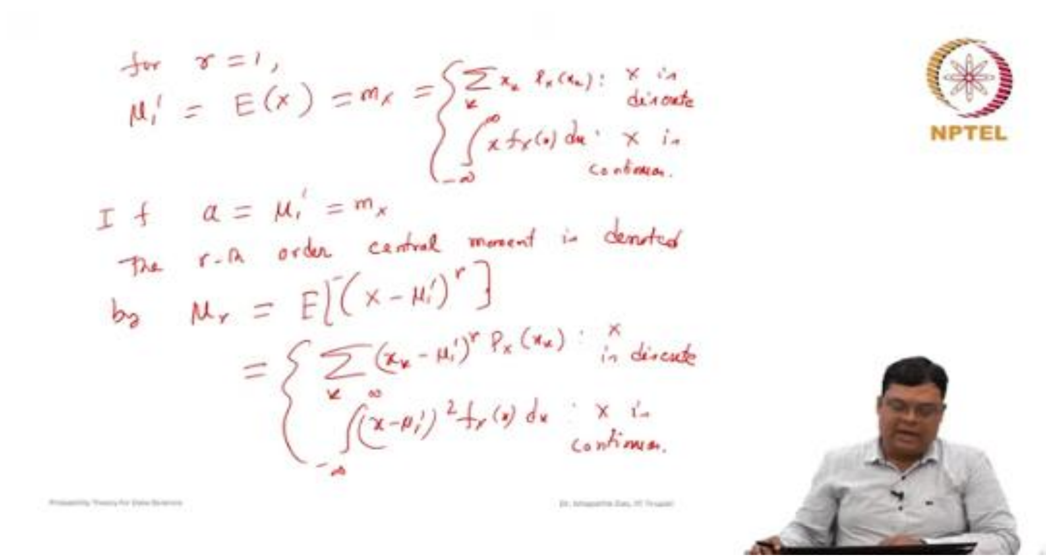


The notation μ_1' represents the first-order raw moment, specifically for the random variable X, and is calculated as the expected value $E(X)$, also denoted as m_X, which is simply the mean of X by definition. When $r = 1$, this expectation simplifies to $\sum x_k \cdot p(x_k)$ for discrete X and to $\int$ from $-\infty$ to ∞ of $x \cdot f(x)\,dx$ for continuous X. Thus, μ_1' represents the mean of X.

Now, if we interpret μ_1' as m_X, the general form becomes the r-th order moment around a point a. Setting a to the mean of X, we obtain the r-th order central moment, denoted μ_r. For clarity, we use μ_r' for r-th order raw moments and μ_r for r-th order central moments.

The central moment is defined as $E[(X - \mu_1')^r]$, or equivalently $E[(X - m_X)^r]$ since μ_1' is also denoted as m_X. If $a = \mu_1'$, this translates to $\sum (x_k - \mu_1')^r \cdot p(x_k)$ in the discrete case, and $\int$ from $-\infty$ to ∞ of $(x - \mu_1')^r \cdot f(x)\,dx$ in the continuous case. These definitions illustrate that for $a = 0$, we retrieve the r-th order raw moment, and for $r = 1$, we obtain the formula for the mean.

If a = $\mu1'$, then we calculate the r-th order central moment. For r = 1, the central moment $\mu1$ equals E[(X − $\mu1'$)^1]. Demonstrating this expectation involves calculating E[X − $\mu1'$], where $\mu1'$ is a constant. Since E(X) = $\mu1'$ and the expected value of a constant remains constant, the expression simplifies according to these properties of expectation.

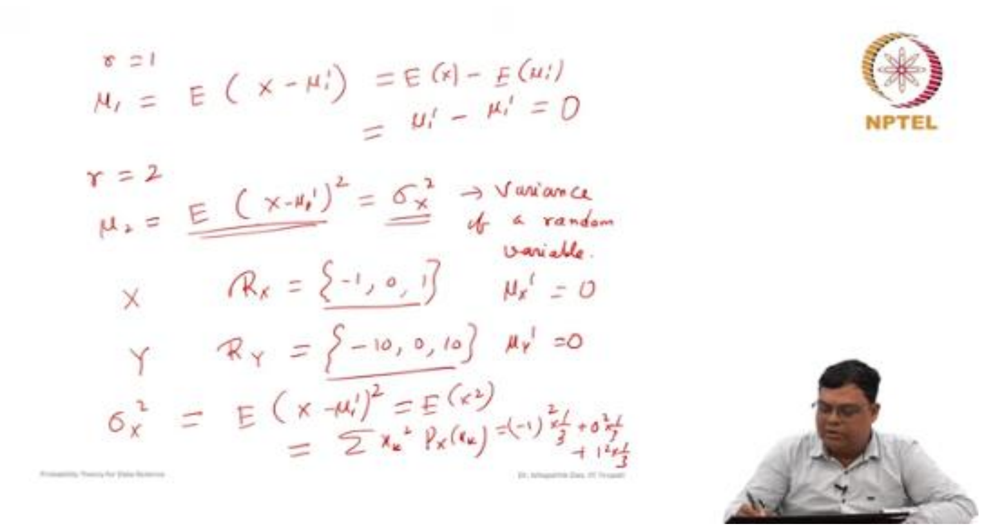


Therefore, this leads to the conclusion that this is equal to 0. Now, when r = 2, what is μ_2? The second central moment is defined as $(X - \mu_1')^2$. This is known as σ_x^2, a notation for the variance of a random variable. We have already shown two examples of discrete random variables: one random variable X takes values -1, 0, and 1, while another random variable Y takes values -10, 0, and 10, each with equal probabilities of 1/3.

Both have the same mean, with $\mu_x' = 0$ and $\mu_y' = 0$. But how do we distinguish between these two datasets? The key difference is in how the values are distributed. In one case, all the values are close to 0, while in the other case, most of the values are far from 0—two values are away from 0, and one value is exactly equal to the mean. This measure, μ_2, provides the expected value of the distance from the mean.

That's why μ_1' represents the average distance from the mean of this random variable. This is also why we find the variance. To calculate the variance of these two random variables, you need to find σ_x^2, which is the expected value of $(X - \mu_1')^2$. Since $\mu_1' = 0$ for this random variable, we find that $\mu_x' = 0$. Thus, the expected value of X^2 is determined using the formula for the r-th order raw moments.

For this random variable, it will be calculated as: $(-1)^2 * (1/3) + (0)^2 * (1/3) + (1)^2 * (1/3)$. So, your variance, σ^2, will be calculated as follows: this is 0, 1, 1/3, 1, and 1/3. Therefore, this is nothing but 2/3. Now, what will be the variance of Y, σ_Y^2? σ^2 will be the expected value of $(Y - \mu_Y')^2$.

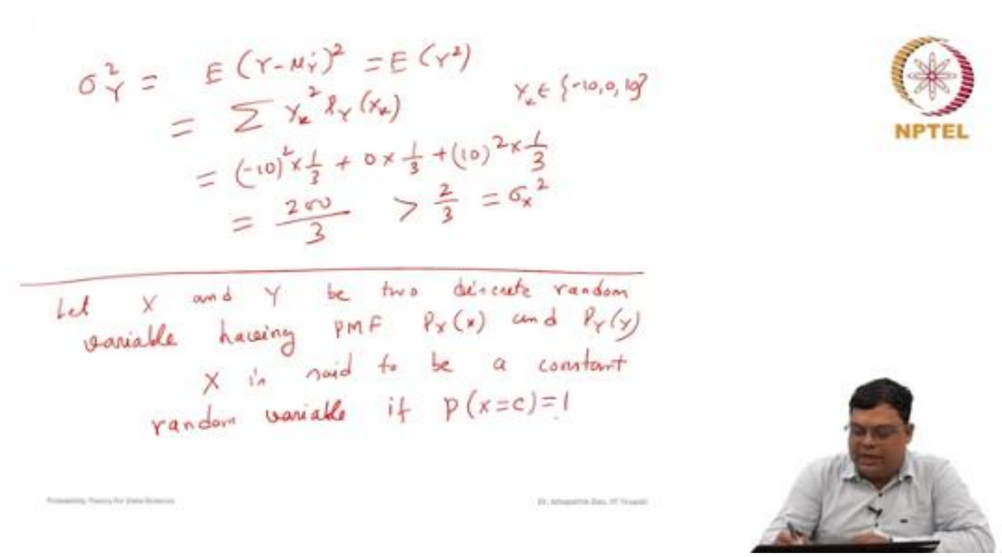


Here, the mean is also 0, so this is the expected value of Y^2. Now, the expected value of Y^2 is the summation of y_k multiplied by the probability of y_k. For this random variable y_k, we know that Y takes the values -10, 0, and 10, each with a probability of 1/3. So, we calculate this as follows: $(-10)^2 * (1/3) + 0^2 * (1/3) + 10^2 * (1/3)$. This gives us: $100 * (1/3) + 0 + 100 * (1/3)$.

So, we have 100/3 + 100/3, which equals 200/3. You can see that this variance is greater than the variance for X, which we found to be 2/3. Essentially, the variance provides a measure of how the data are scattered around the center. You can find the mean of a random variable, but if you want to understand how the students performed—whether their scores are concentrated around the average or if the highest and lowest scores are far from the average—you need to calculate the variance. That is why variance is important; it measures the dispersion of the dataset and indicates whether the values are near the center or scattered away from it.

Now, let's consider another example. We have already discussed uniform random variables. This is an example with a probability density function or cumulative distribution function in the discrete case. Regarding uniform distribution, we have previously found the mean. I just want to find some additional examples based on what we discussed.

Suppose we have two random variables, X and Y, with probability mass functions $p_x(x)$ and $p_Y(y)$. If you consider the expected values of a constant, say c, for any random variable, what will be the expected value of a constant random variable? If we define a constant random variable, X is said to be a constant random variable if it takes only one value with a probability of 1. $P(X = c) = 1$ for some $c \in \mathbb{R}$. Then, what will be the expected value of c?



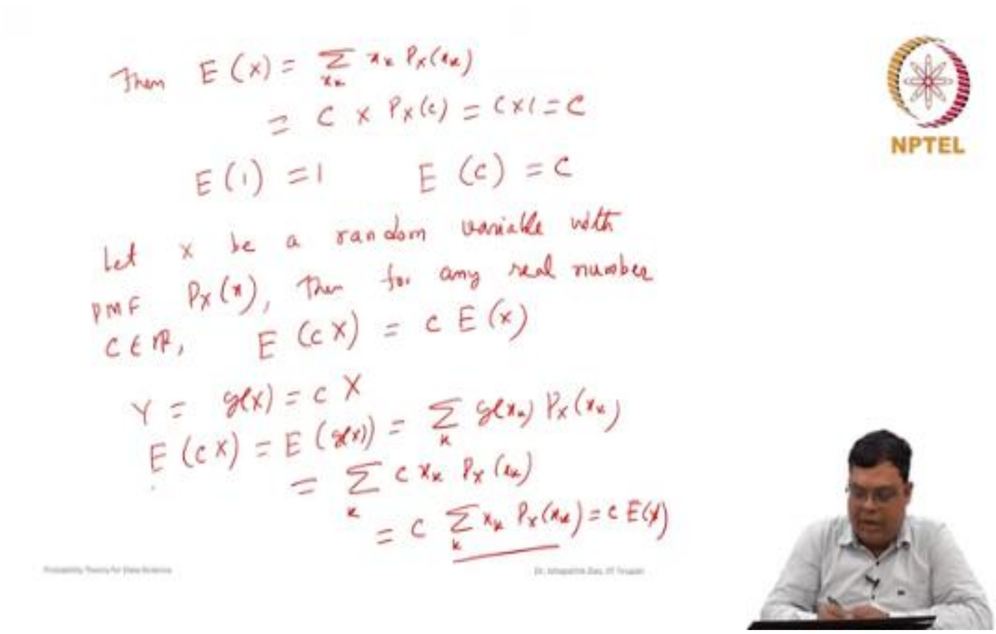
So, basically, expected values we say for X. It is nothing but the summation by definition of $x_k * p_x(x_k)$, which is the summation of $x_k * p_x(x_k)$. So, x_k is what values X is taking. Now, X is a constant random variable taking only the value c with probability $p_x(c) = 1$. So, c * 1 is nothing but c.

That is why we denote a constant random variable by a constant. Whenever we say expected value of 1, what is 1? You can think of it as a constant random variable taking the value 1 with probability 1. Therefore, for any real number, the expected value of c will be c. Let X be a random variable with probability mass function p_x.

For any real number $c \in \mathbb{R}$, the expected value of c * X is nothing but c * E[X]. How will we prove that? Let us find out what it is. Here your transformation is y = c * X, which we denote as g(X). This is nothing but c * X.

Now, how will we find the expected value of cX? This is nothing but the summation of expected value of g(X). We have the formula as $\sum g(x_k) * p_x(x_k)$. Now, what are the possible values of k? This is $\sum k$, where $g(x_k)$ is nothing but $g(X) = c * x_k * p_x(x_k)$.

Now, since c is a constant, we can factor it out, resulting in $c * \sum k$, where $x_k * p_x(x_k)$. Now, by definition, this is nothing but the expected value. So, this is c * E[X]. Therefore, for any constant random variable, the expected value of c * X is nothing but c * E[X]. Now, in the case of variability, let's consider X as a continuous random variable.

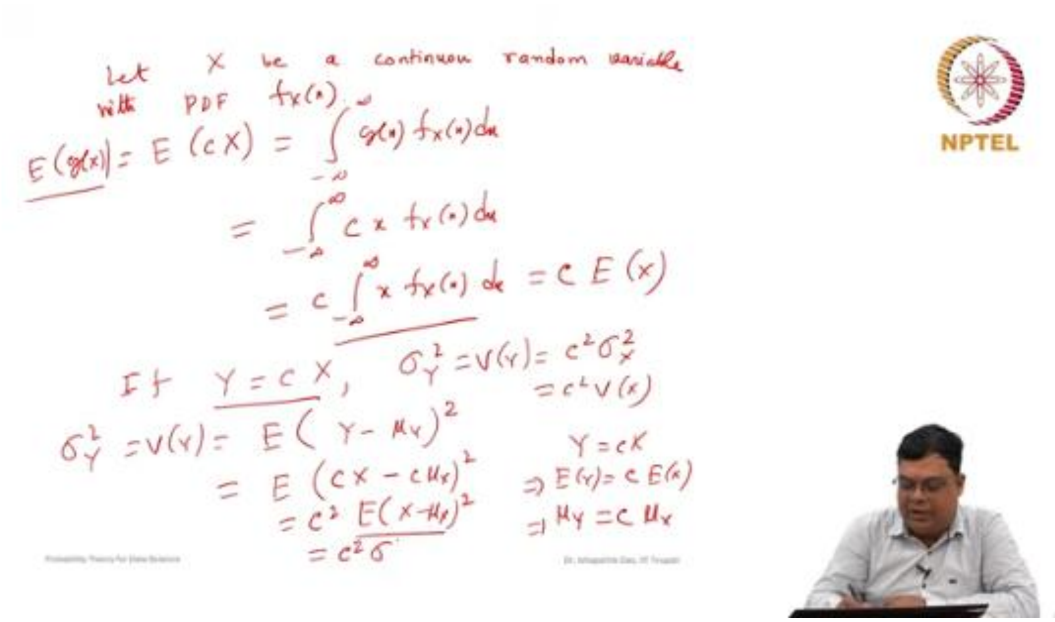


Suppose X has a probability density function fX. Then, what will be the expected value of cX? The expected value of cX is the integration of gX times fX dx. By definition, the expected value of gX is the integral from minus infinity to plus infinity of gX times fX dx. So, if we set gX equal to cX, we have the integral from minus infinity to plus infinity of cX times fX dx.

Since c is a constant, we can factor it out, resulting in c times the integral from minus infinity to plus infinity of X times fX dx. By definition, this is nothing but c times the expected value of X. Now, let's discuss the variance. If Y is equal to c times X, then the variance of Y, denoted as sigma squared Y, will be equal to c squared times the variance of X. How do we prove this?

The variance of Y is defined as the expected value of (Y minus muY) squared. Now, if we take Y equal to cX, we have already shown that muY becomes c times muX. Therefore, the expected value of Y is equal to c times the expected value of X, which implies muY equals c times muX. Consequently, we can write the expression as (cX minus c times muX) squared. Factoring out the constant gives us c squared times the expected value of (X minus muX) squared.

Thus, we conclude that the variance of Y is equal to c squared times the variance of X. This is nothing but the variance of X. So, it is c squared times the variance of X. These are some of the results we require. More generally, as an exercise, you can prove this.

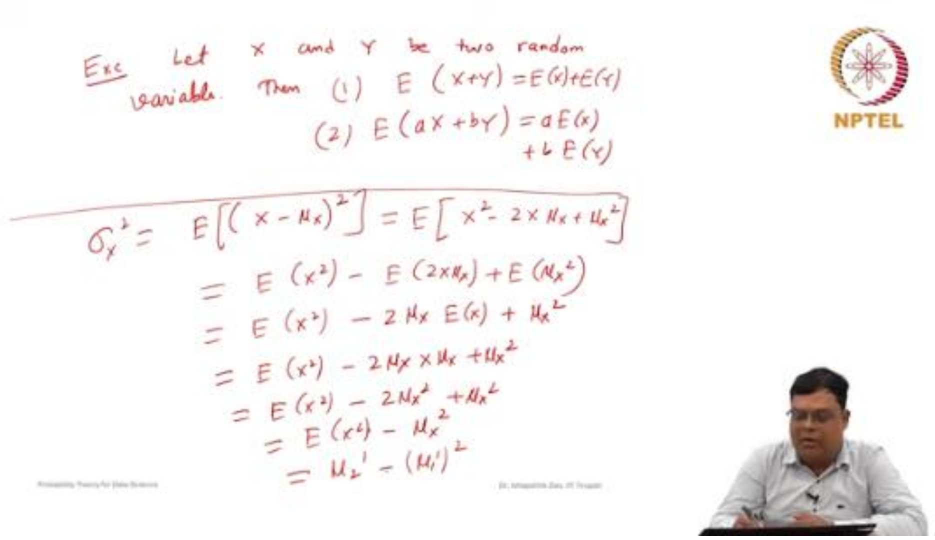


Suppose X and Y are two random variables. You can consider both as discrete random variables or both as continuous random variables. Then, you can show that the expected value of X + Y is equal to the expected value of X plus the expected value of Y. If we write it as E(aX + bY), this is equal to a * E(X) + b * E(Y). This resembles a linear transformation.

In linear algebra, the definition of a linear transformation states that it preserves operations like addition and scalar multiplication between vector spaces. Similar principles apply here; hence, this is a linear transformation. You can try this exercise later, and we will discuss it again. Now, using this concept, we can simplify the variance formula. When computing variance, you need to subtract all values from the mean and then compute the square of that difference.

Thus, the variance formula, σ^2_x, is $E[(X - \mu_x)^2]$. Now, this can be expressed as $E[(X - \mu_x)^2]$, which expands to $E[X^2] - 2X * \mu_x + \mu_x^2$. Note that μ_x is a real number, and after performing the expected value operation, it is still a real number. If we assume this is true, we can express it as $E[X^2] - E[2X * \mu_x] + E[\mu_x^2]$. Since μ_x is constant, the expected value of a constant times X is the constant times the expected value of X.

Therefore, we have $2 * \mu_x * E[X] + \mu_x^2$. This simplifies to $E[X^2] - 2 * \mu_x * E[X] + \mu_x^2$. Further simplifying, we find that the result is $E[X^2] - \mu_x^2$. If you remember the notation for raw moments, this can be expressed as $\mu_2' - (\mu_1')^2$. Whenever you are using data and computing the expected values, it will be easy to find because you just need to calculate the raw moments—first-order raw moments, second-order raw moments—and then you subtract them to find the variance.



So, let us do one example.