PROBABILITY THEORY FOR DATA SCIENCE

Prof. Ishapathik Das

Department of Mathematics and Statistics

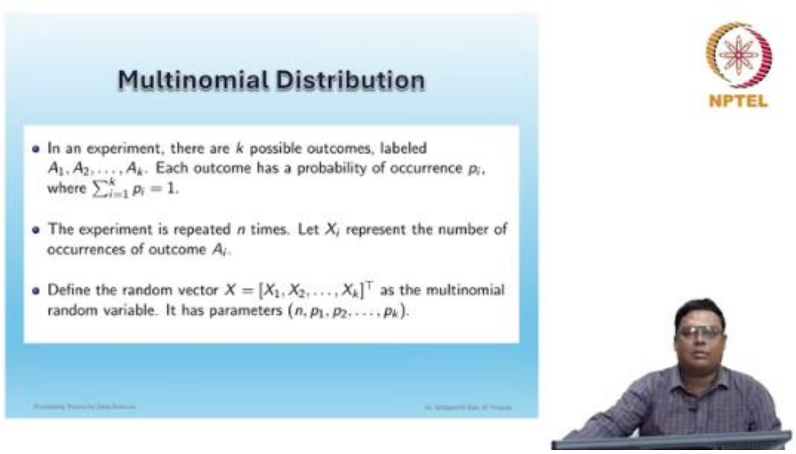
Indian Institute of Technology Tirupati

Week - 10

Lecture - 52

Multinomial Distribution and Multivariate Normal Distribution

Let us discuss one of the important distributions, known as the multinomial distribution. This is a discrete distribution, and discrete multivariate random variables follow a multinomial distribution under certain conditions. Suppose you are collecting responses that fall into more than two categories. For binomial cases, we discussed that the multinomial distribution is an extension of the binomial distribution. When considering responses with more than two categories, for example, different types of fish in a pond, the situation extends beyond binomial responses.



Suppose the fish are categorized by color, such as red, yellow, blue, or white. This hypothetical scenario can also apply to other objects, like colored balls. If there are only two categories, for instance, red and black, the responses are called binomial responses. A simple example of this is tossing a coin, where there are two possible outcomes: heads or tails. This is a classic case of binomial responses.

However, if there are more than two categories for each observation, the responses involve multiple categories. For example, in a pond with four types of fish (red, black, blue, and white), the observations can be classified into these four categories. Suppose you consider a total of "n" fish, and let n_1, n_2, n_3, and n_4 represent the number of red, black, blue, and white fish, respectively. These counts are random variables. Define Y_1, Y_2, Y_3, and Y_4 as random variables corresponding to the categories:

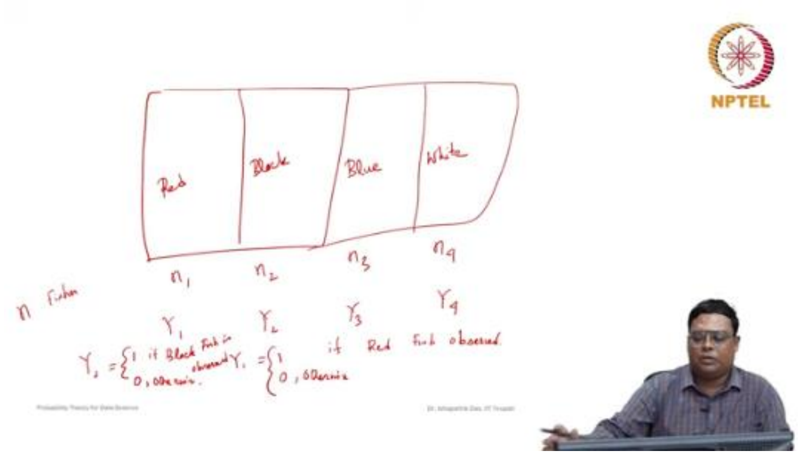
$Y_1 = 1$ if a red fish is observed; otherwise, $Y_1 = 0$.
$Y_2 = 1$ if a black fish is observed; otherwise, $Y_2 = 0$.
Similarly, Y_3 and Y_4 represent blue and white fish, respectively. These variables define the outcomes. Note that they may not be independent because the total number of fish is fixed, making the counts dependent.

For instance, observing one type of fish reduces the probability of observing other types in the same sample. In general, consider an experiment with K mutually exclusive outcomes labeled $A_1, A_2, ..., A_k$. For example, the four outcomes in the fish scenario are red, black, blue, and white. Mutually exclusive means that if one outcome occurs, such as observing a red fish, none of the others can occur simultaneously. Each outcome has a probability of occurrence denoted by $p_1, p_2, ..., p_k$, with the condition that the sum of all probabilities equals 1:

$p_1 + p_2 + ... + p_k = 1$

For instance:
The probability that Y_1 equals 1 (observing a red fish) is p_1.
The probability that Y_2 equals 1 (observing a black fish) is p_2, and so on.

These probabilities are either known or estimated from data.



Provided that $Y_4 = 1$, this corresponds to p_4, while p_3 corresponds to $Y_3 = 1$. These probabilities cover the four possibilities: red, black, blue, and white fish. Since any fish caught must belong to one of these categories, the probabilities are mutually exclusive and exhaustive. Thus, the sum of these probabilities satisfies:

$p_1 + p_2 + p_3 + p_4 = 1$

If the experiment is repeated n times, for example, drawing n fish sequentially, let X_i represent the number of occurrences of outcome A_i (e.g., the number of fish of color i).

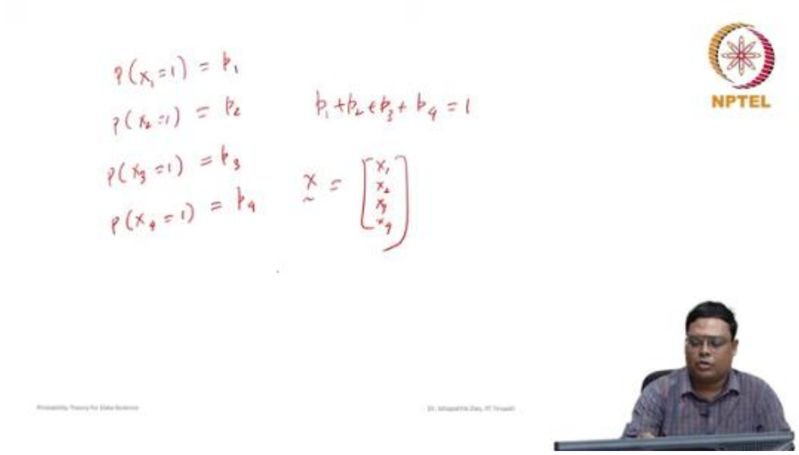
Previously, we used Y_i, but here we use X_1, X_2, X_3, X_4 for consistency. The choice of notation does not affect the results, but clarity is important. The sum of probabilities remains:

$p_1 + p_2 + p_3 + p_4 = 1$

The random variable X, comprising the components X_1, X_2, X_3, X_4, is referred to as a multinomial random variable. It is characterized by the parameters n (the number of trials) and p_1, p_2, p_3, p_4 (the probabilities of each outcome).

Thus, $X = (X_1, X_2, X_3, X_4)$ is a four-dimensional random vector following a multinomial distribution with parameters n, p_1, p_2, p_3, p_4. The task is to determine its probability mass function (PMF).

Recall that for the binomial distribution, the PMF specifies the probabilities of each possible outcome based on the given parameters. Similarly, the multinomial distribution has a specific PMF, which generalizes the binomial case to multiple categories. In the binomial distribution, each trial is a random experiment repeated n times, with only two possible categories for each trial. For example, we may have a "yes" or "no" outcome, or categories like "blue" or "black."



Let's assume that Y_1 represents the first time you draw, and for simplicity, Y is the random variable. The probability of Y being "yes" (1) is p, and the probability of Y being "no" (0) is 1 - p. If you repeat this experiment n times, what would be the probability of observing a certain number of "yes" outcomes? For each trial, Y is a Bernoulli random variable. We define the random variable Y to represent how many times "yes" occurs.

This follows a binomial distribution with parameters n (the number of trials) and p (the probability of a "yes"). The probability mass function (PMF) for this binomial distribution is given by:

$P(Y = y) = nCy * p^y * (1 - p)^{(n - y)}$

Here, y can take values from 0 to n, where y represents the number of "yes" outcomes, and n - y represents the number of "no" outcomes. Let's now denote the random variable for "yes" outcomes by Y_1 and for "no" outcomes by Y_2, such that $Y_1 + Y_2 = n$.

This notation simplifies understanding, as it clearly separates the outcomes of "yes" (Y_1) and "no" (Y_2). The probability that Y equals y is equivalent to the probability that Y_1 equals y_1 and Y_2 equals y_2. For one trial, if Y_1 represents "yes," the probability is p_1, and for Y_2 representing "no," the probability is p_2. This is analogous to the binomial case, where p_1 is p and p_2 is 1 - p.

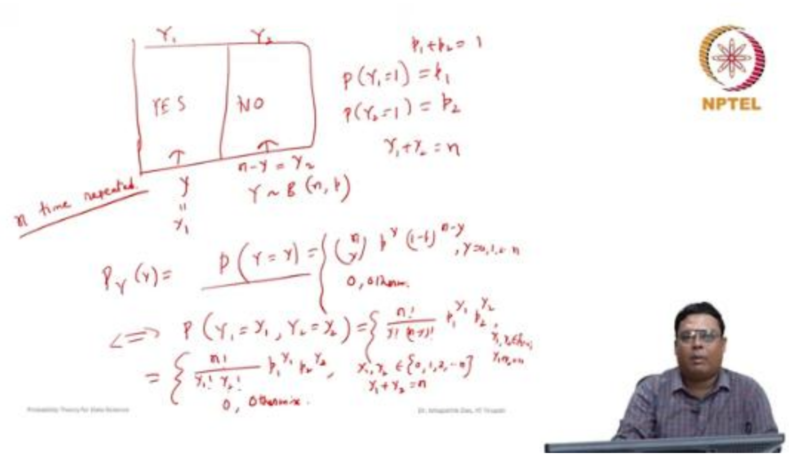
Now, for n trials, the probability that there are y_1 "yes" outcomes ($Y_1 = y_1$) and y_2 "no" outcomes ($Y_2 = y_2$) is represented by:

$P(Y_1 = y_1, Y_2 = y_2) = nCy_1 * p_1^{y_1} * p_2^{y_2}$

Where y_1 and y_2 are values between 0 and n such that $y_1 + y_2 = n$. This is the same PMF as in the binomial case, just expressed with more specific notation for clarity.

Now, we generalize this for the multinomial distribution. The binomial case is a special case where there are only two categories. However, when dealing with more than two categories, we need to generalize the binomial distribution to the multinomial distribution. This involves adjusting the probability mass function to account for multiple categories, where each category has its own probability and each trial can result in one of the multiple categories. Let's now discuss how the multinomial distribution's probability mass function is derived.

In the case of multinomial distributions, we have more than two categories.



For example, suppose there are q categories. Let's consider the notation for k categories, labeled 1, 2, ..., k. For each category, there is a corresponding random variable: Y_1 for category 1, Y_2 for category 2, and so on, up to Yk for category k. The probability for each category is denoted as p_1 for category 1, p_2 for category 2, and pk for category k. The sum of all these probabilities must equal 1:

$p_1 + p_2 + ... + pk = 1$

Now, suppose we repeat the trial n times. We want to find the probability that the number of occurrences of each category is y_1 for category 1, y_2 for category 2, and so on, up to yk for category k. This means, if we draw n fishes from the pond, we want to know how many will belong to each category, such as how many fishes are red, blue, or black. The total number of occurrences must add up to n:

$y_1 + y_2 + ... + yk = n$

The probability of this outcome can be obtained by extending the binomial distribution to the multinomial case. The formula for the probability is:

$P(Y_1 = y_1, Y_2 = y_2, ..., Yk = yk) = n! / (y_1! * y_2! * ... * yk!) * p_1\text{^}y_1 * p_2\text{^}y_2 * ... * pk\text{^}yk$

Where: n! is the factorial of n, representing the total number of ways the trials can be arranged. $y_1!$, $y_2!$, ..., yk! are the factorials of y_1, y_2, ..., yk, representing the number of ways to arrange the outcomes for each category. $p_1\text{^}y_1$, $p_2\text{^}y_2$, ..., pk^yk are the probabilities raised to the power of the occurrences for each category.

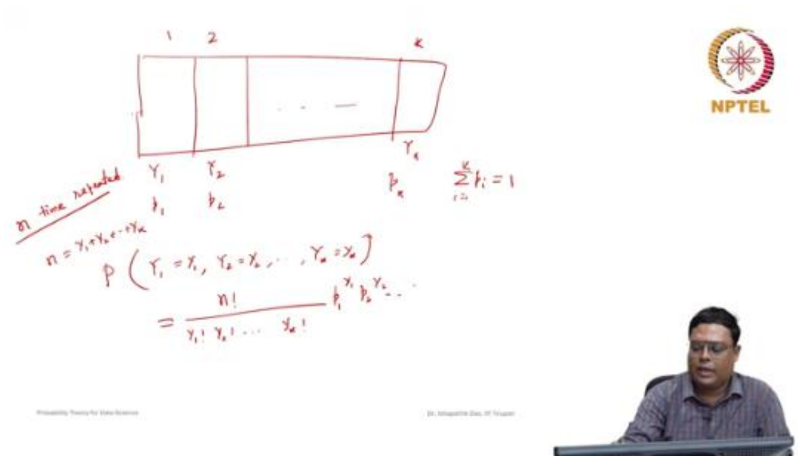
The values of y_1, y_2, ..., yk must satisfy the equation:

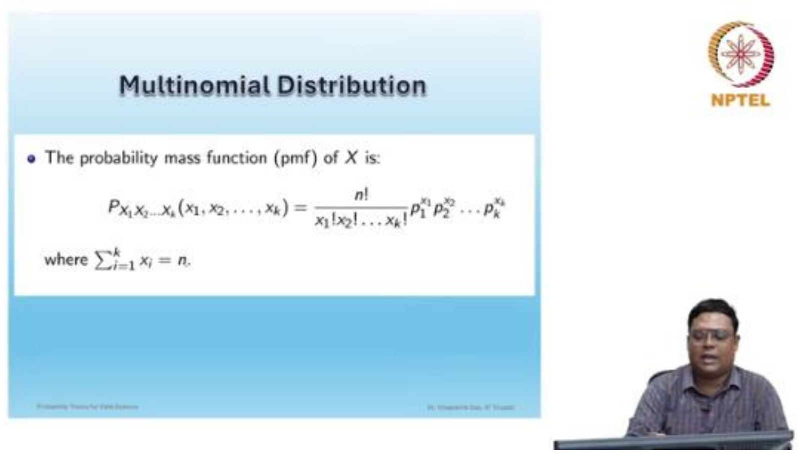
$y_1 + y_2 + ... + yk = n$

If this condition is not met, the probability is 0. This is the probability mass function for a multinomial distribution. It generalizes the binomial distribution, which applies only to two categories, to cases with more than two categories. We extended to the k categories, which is known as a multinomial random variable.

The probability mass function of the multinomial random variable X is given by:

$P(x_1, x_2, ..., xk) = (n! / (x_1! * x_2! * ... * xk!)) * p_1\text{^}x_1 * p_2\text{^}x_2 * ... * pk\text{^}xk$.



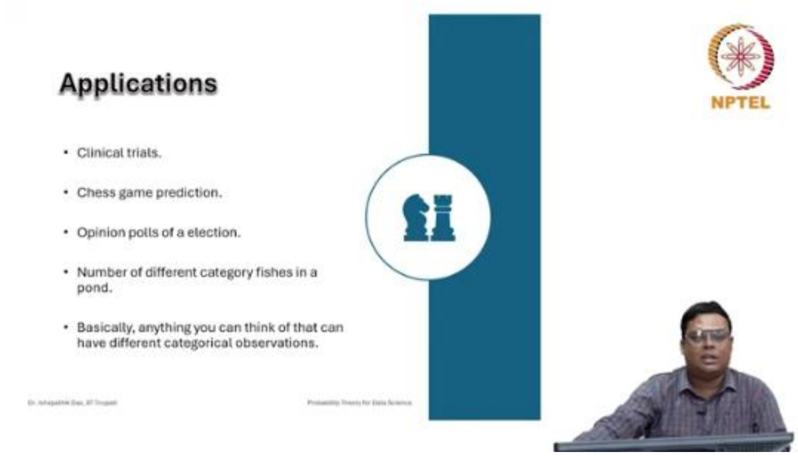


This represents the probability of observing x_1 outcomes in category 1, x_2 outcomes in category 2, and so on, out of n total trials. Such that the summation of x_i should be equal to n:

$x_1 + x_2 + \ldots + x_k = n$

$x_1, x_2, \ldots, x_k$ can take any value between 0 and n, but the summation of all x_i must equal n. This is known as the multinomial distribution function. Since this is a multivariate case, it can be a bit more complicated. We won't go into further details right now, but we will discuss it more in the future whenever it's needed.

Now, let's discuss the applications of multinomial random variables. One of the applications of multinomial random variables is in clinical trials. When there is more than one category in the responses, for example, when a disease can be classified based on its severity, multinomial distributions are used to analyze and model the different levels or categories of severity. The presence or absence of a disease can be categorized as 0 or 1. However, diseases like cancer can be classified into stages: no cancer, stage 1, stage 2, or stage 3.

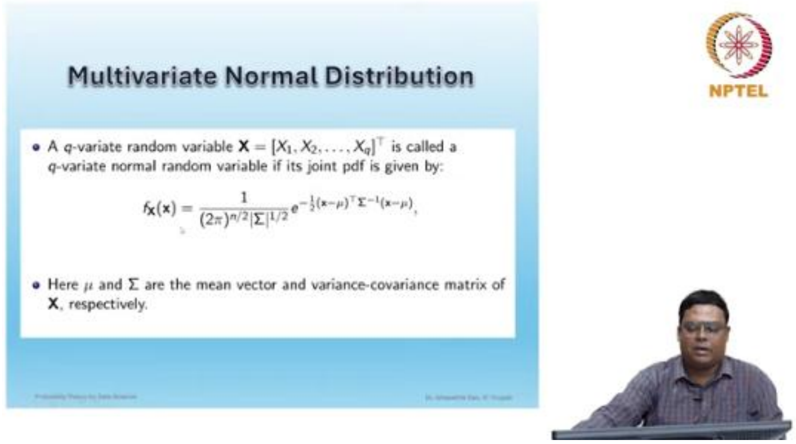


In this way, multiple categories can be defined. For example, when administering medicine, there may be adverse effects or toxicity. These can be categorized as either no adverse effect or a side effect, but they can also be classified into more than two categories, such as mild, severe, or even death. In this case, we would use the multinomial distribution. Multinomial distribution is also used in predicting game outcomes or in opinion polls during elections.

In opinion polls, there are more than two candidates or different posts to consider. If there are only two categories, it is a binomial distribution. However, if there are more than two, we apply the multinomial distribution, as it involves multiple categories. In summary, the multinomial distribution is an extension of the binomial distribution, used when dealing with more than two categories.

Next, we will discuss multivariate random variables in the case of continuous data, specifically the multivariate normal distribution.

A q-variate random variable X is called a q-variate normal random variable if its joint probability density function can be represented in a certain way. Although it may look complicated, let's first review the univariate normal distribution to help understand the concept. Let X be a univariate random variable. For a univariate normal distribution, X is a normally distributed random variable with mean (μ) and variance (σ^2), where $\sigma^2 > 0$. The mean (μ) can be any real number.



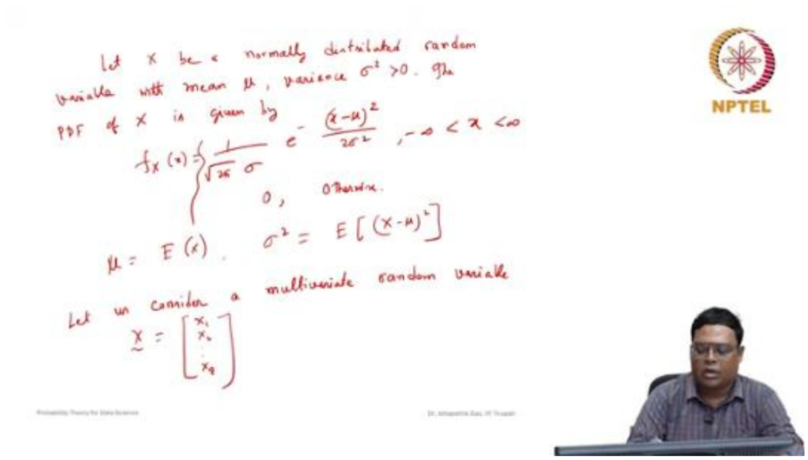
The probability density function (PDF) of X is given by: $f\_X(x) = 1 / (\sqrt{(2\pi\sigma^2)}) * e^{(-(x - \mu)^2 / (2\sigma^2))}$,

where x ranges from negative infinity to positive infinity, and the PDF is 0 otherwise. This is the probability density function of a univariate random variable. Here, μ is the mean or expected value of X, and σ^2 is the variance, which is the expected value of $(X - \mu)^2$. These are the two key parameters of the univariate normal distribution.

Now, consider a multivariate random variable. Let X be a q-dimensional random variable, represented as $X = (X_1, X_2, ..., Xq)$. To extend this to q dimensions, we need to define the parameters for a q-dimensional random variable. The expected value of X in q dimensions is defined as the expected values of the vector $X = (X_1, X_2, ..., Xq)$.

Since each X_i is a random variable, this defines the expected value for each component of the vector. We can define the expected values of $X_1, X_2, ..., Xq$. These are all real numbers, and we denote the expected values of $X_1, X_2, ..., Xq$ as $\mu_1, \mu_2, ..., \mu q$. This can be represented as a vector, denoted as μ, which belongs to $\mathbb{R}^p$.

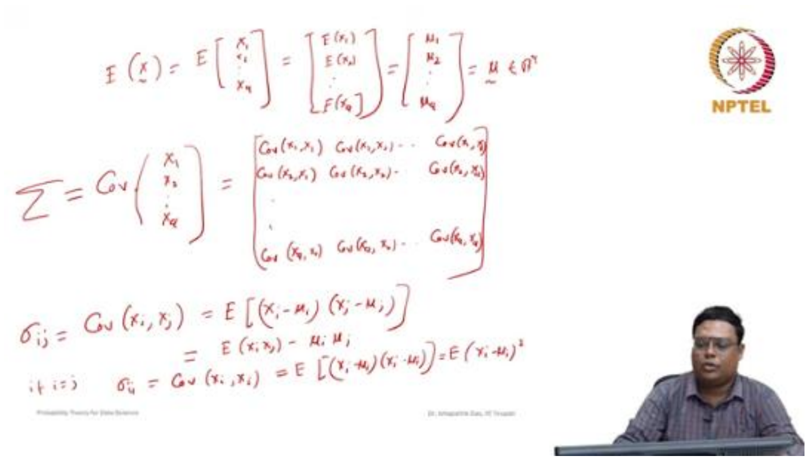
Next, we consider the variance. In the univariate case, variance is just a single number: $(X - \mu)^2$.



However, in the multivariate case, we need to consider covariance. The covariance of the random variables X_1, X_2, ..., X_q is defined by considering all possible combinations of covariances between pairs of random variables. This results in a matrix, where each element represents the covariance between a pair of random variables. The covariance matrix will contain the covariance between X_1 and X_1, which is the variance of X_1 ($Var(X_1)$), the covariance between X_1 and X_2, X_1 and X_q, and so on. Similarly, the covariance between X_2 and X_1, X_2 and X_2, X_2 and X_q, and so on. The matrix will be symmetric.

We denote this covariance matrix as Σ. For any covariance between X_i and X_j, we define it as the expected value of the product of $(X_i - \mu_i)(X_j - \mu_j)$, which simplifies to $Cov(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)] = E[X_i * X_j] - \mu_i * \mu_j$. If i equals j, the covariance between X_i and X_i is simply the variance of X_i, $Cov(X_i, X_i) = Var(X_i)$.

Thus, the covariance matrix Σ is a square matrix with dimensions equal to the number of random variables (q). The diagonal elements represent the variances of each random variable, and the off-diagonal elements represent the covariances between pairs of random variables.

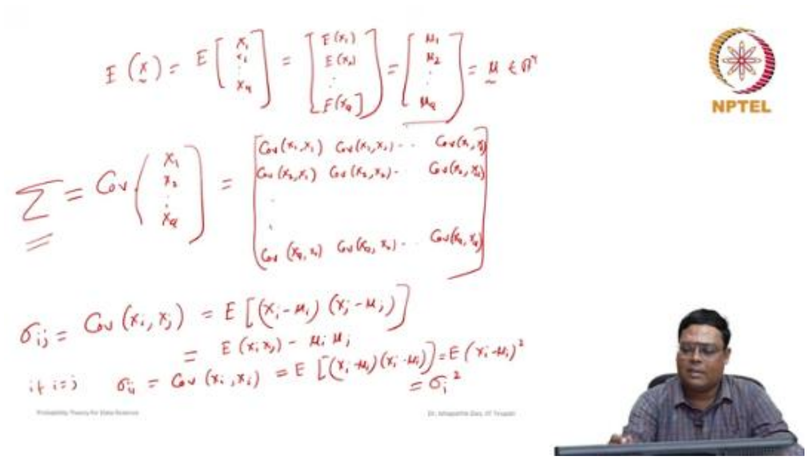


Certainly, here is the text with all mathematical words converted into symbols or equations, as per your request:

So, you can write the covariance matrix as follows: the first row will be the variance of X_1 (denoted as σ_1^2), followed by the covariances between X_1 and X_2, X_1 and Xq, and so on. The second row will start with the covariance between X_2 and X_1, then the variance of X_2 (σ_2^2), followed by the covariance between X_2 and Xq, and so on. Similarly, the last row will contain the covariances between Xq and each of the other variables, ending with the variance of Xq (σq^2). This forms the variance-covariance matrix, which is denoted as Σ. The variance-covariance matrix of the multivariate random variable X, which consists of X_1, X_2, ..., Xq, is represented by Σ.

This matrix captures both the variances of the individual variables and the covariances between them. In the univariate case, we have the mean and variance. For a multivariate random variable, the mean is represented as a vector μ, and the covariance is represented by the matrix Σ. The covariance matrix is positive definite, meaning that its eigenvalues are positive. The vector of means μ can be any real number or vector in $\mathbb{R}^{\varphi}$.

The covariance matrix Σ is assumed to be positive definite. In the univariate case, the mean μ can be any real number, ranging from -∞ to +∞, and the variance σ^2 must be greater than 0. However, in the multivariate case, the covariance matrix Σ needs to be positive definite for the distribution to be well-defined. This ensures that the random variable has meaningful relationships (covariances) between its components. The covariance matrix, Σ, must be a positive definite matrix, which is essential for the variance-covariance matrix.

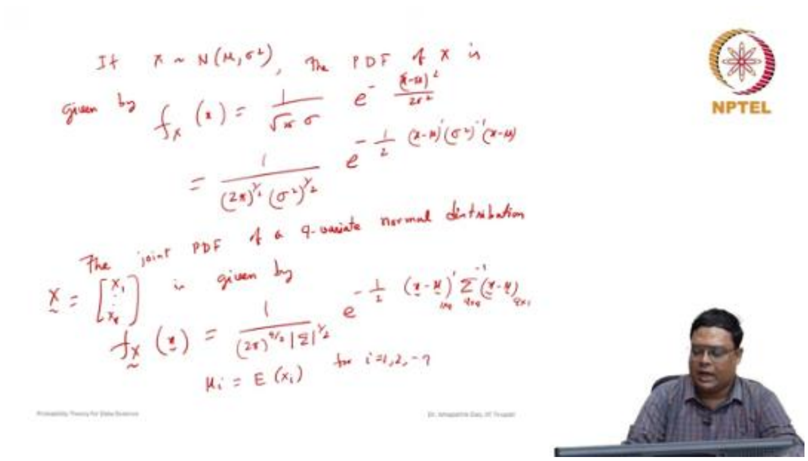


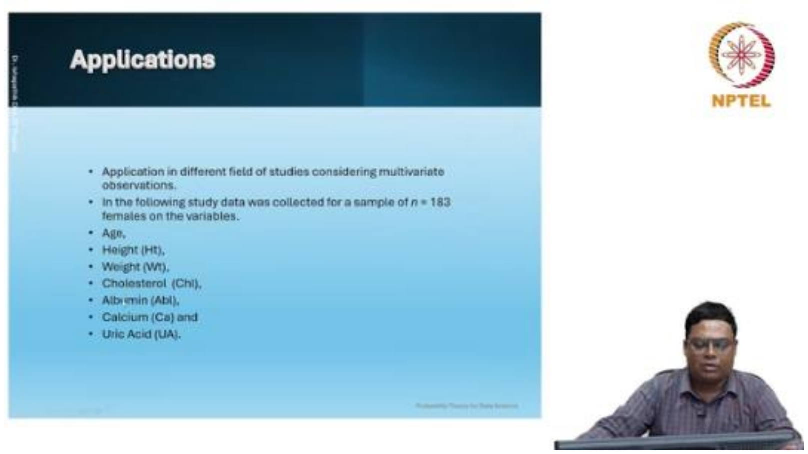
Now, let's return to the probability density function (PDF) for univariate cases. When X has a normal distribution with mean μ and variance σ^2, the PDF of X is given by: $f(x) = (1 / \sqrt{2\pi\sigma^2}) * e^{-(x - \mu)^2 / 2\sigma^2}$. This can be represented as: $f(x) = (1 / \sqrt{2\pi}) * (1 / \sqrt{\sigma^2}) * e^{-(x - \mu)^2 / 2\sigma^2}$.

We can further simplify this, as $(x - \mu)$ is a real number and can be treated as a 1x1 matrix, which allows us to write it in matrix form. Now, let's generalize this for multivariate random variables. The joint probability density function (PDF) of a multivariate normal distribution, or a q-variate normal distribution, is represented as follows: $f(X) = (1 / (\sqrt{2\pi})^q) * (1 / (|\Sigma|^{1/2})) * e^{-(1/2) * (X - \mu)^T \Sigma^{-1} (X - \mu)}$. Here, X is the vector of random variables $(X_1, X_2, ..., X_q)$, μ is the mean vector, Σ is the variance-covariance matrix, and $|\Sigma|$ denotes the determinant of Σ.

Since Σ is positive definite, its determinant is always greater than 0. In this formula, we take the determinant of Σ (which is always positive), and the matrix $(X - \mu)$ represents the difference between the observed value and the mean. The term Σ^{-1} represents the inverse of the covariance matrix, and $(X - \mu)^T$ denotes the transpose of the vector $(X - \mu)$. The result of this operation is a real number. Thus, this is the joint probability density function for a multivariate normal distribution.

The significance of the parameters is that μ represents the expected values of the random variables X_i (for i = 1 to q), and Σ is the variance-covariance matrix of X. We have already discussed random variables like $X_1, X_2, ..., X_q$. We won't go into more detail about the multivariate normal distribution here because it is more complicated, given that we are considering multivariate random variables. It's not as simple as univariate cases. However, just remember that for univariate distributions, the density function looks like this.





For multivariate cases, it is not as simple and can be represented in a more complex way. To understand this concept, imagine a hill or a peak, where the distribution increases towards the peak and then decreases. At the peak, you will find the mean (μ), mode, and median of the normal distribution. This distribution is symmetric on both sides. Similar properties apply in multivariate cases as well.

The applications of multivariate random variables are numerous. For instance, when you consider multiple parameters or variables and multiple responses, multivariate random variables are used. In a study, data was collected from a sample of 183 females, measuring variables like height (X_1), weight (X_2), cholesterol (X_3), albumin (X_4), calcium (X_5), and uric acid (X_6). These are seven variables, so $q = 7$. Each variable could be considered as a univariate normal distribution, with most being normally distributed random variables.

However, if we consider them together as a vector, they may not be independent. For example, age (X_1) and height (X_2) might be correlated, as well as weight (X_3) and height (X_2), or cholesterol (X_4) and weight (X_5). Since these variables are not independent, we cannot treat them as univariate random variables and simply multiply them to find the multivariate distribution function or density function. Instead, we need to account for their correlations, and that's where the multivariate normal distribution comes in.

In summary, we have discussed some examples of multivariate random variables, including the multinomial distribution in discrete cases and the multivariate normal distribution in continuous cases. Hopefully, you have followed the explanation. Next, we will discuss a different topic: the transformation of random variables, which is an important topic to cover.