PROBABILITY THEORY FOR DATA SCIENCE

Prof. Ishapathik Das

Department of Mathematics and Statistics

Indian Institute of Technology Tirupati

Week - 07

Lecture - 37

Marginal Probability Density Function, Independence, and Examples

Let (X, Y) be a continuous bivariate random variable with the joint probability density function $f_{XY}(x, y)$. Suppose this is given. Now, how can we find the marginal probability density function? We want to find the marginal probability density function because X and Y are continuous random variables.

The marginal probability density function of X is given by the derivative of the cumulative distribution function (CDF) of X. It can be expressed as:

$f_X(x) = \frac{d}{dx} F_X(x)$

This is the probability density function.

Now, how do we find this? It is essentially the derivative of $F_X(x)$, which is given by:

$f_X(x) = \frac{d}{dx} \lim_{y \to \infty} F_{XY}(x, y)$

Here, $F_X(x)$ is:

$F_X(x) = \lim_{y \to \infty} F_{XY}(x, y)$

Now, $F_X(x)$ is known to us. To find its derivative, we use the fact that:

$F_X(x) = \lim_{y \to \infty} F_{XY}(x, y)$

This represents the probability that $X \leq x$ and $Y \leq y$. If we want to express this as $y \to \infty$, we use the probability density function. So, we have:

$F_X(x) = \int_{-\infty}^{x} \left[\int_{-\infty}^{y} f_{XY}(u, v)\, dv \right] du$

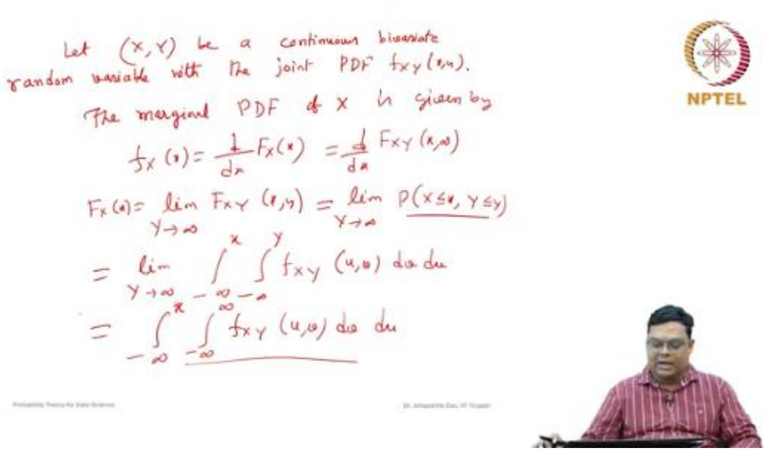
Now, by the definition of integral calculus, the function inside the integral depends on u, and after performing the integration with respect to v, we get:

$F_X(x) = \int_{-\infty}^{x} [\int_{-\infty}^{\infty} f_{XY}(u, v)\, dv]\, du$

Thus, this is $F_X(x)$, and when you take the derivative of $F_X(x)$ with respect to x, you obtain:

$f_X(x) = \frac{d}{dx} F_X(x) = \frac{d}{dx} \int_{-\infty}^{x} [\int_{-\infty}^{\infty} f_{XY}(u, v)\, dv]\, du$

This is the marginal probability density function $f_X(x)$.



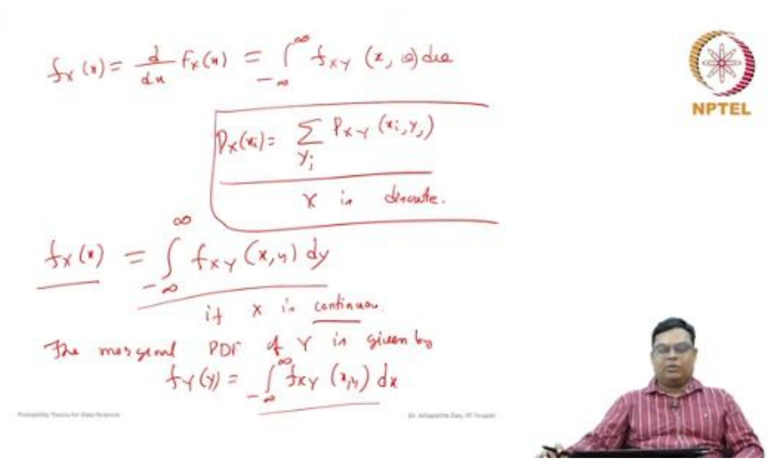
This becomes f(x, y), and now it is becoming because we are taking the limit from $-\infty$ to x, and then y. So, this is actually v, and we took this dv. This is the $f_{xy}(v)\, dv$. This is $\int$ from $-\infty$ to ∞ of f_{xy}. Now, if you apply some integral rules and take the derivative of this function, it becomes evaluated at x, $\int_{-\infty}^{\infty} f_{xy}(x, v)\, dv$.

Basically, what we finally found is shown here. Similarly, as explained, $F_x(x)$ becomes $\int$ of $f_{xy}(x, \eta)\, d\eta$ when you take the derivative. So, that means without going through this complexity, you can understand how we found it in the discrete case. In the discrete case, we found $P_{xy}(x_i, y_j)$. If you fix x_i and take the sum over the other variable, we found the marginal density function for X in the discrete case, where X is a discrete variable.

Now, for the continuous case, how can we find that? For the continuous case, to determine $f_x(x)$, instead of taking a sum as in the discrete case, we take the integration of the joint density function $f_{xy}(x, y)$ with respect to the other variable. Whenever X is a continuous random variable, the marginal probability density function of X can be determined by integrating the joint probability density function with respect to the other variable. By performing this integration, we obtain the marginal probability density function of X. Similarly, the marginal probability density function of Y, denoted as $f_y(y)$, is obtained by integrating the joint probability density function $f_{xy}(x, y)$ with respect to x over the interval from $-\infty$ to ∞.

In the discrete case, we take the sum, while in the continuous case, we integrate with respect to the other variable, dx. By performing the integration with respect to x, we obtain the marginal probability density function of Y. This is the probability density function of Y. Similarly, we can find the marginal probability density function of X using this formula. Now, we will proceed to the next topic.

I hope this is clear. You have understood how, when the joint probability density function is given, we can find the marginal probability density function of X. Similarly, we have also understood how to find the marginal probability density function of Y. Now, we will work through some examples as well. So, it will be clearer.



Now, we know that whenever X and Y are independent, if X and Y are independent random variables, then the joint cumulative distribution function is equal to $F_x(x) * F_y(y)$ for all $(x, y) \in R$. So, we know that if X and Y are independent random variables, then $F_{xy}(x, y) = F_x(x) * F_y(y)$ for all $(x, y) \in R$.

Now, if you take the derivative d/dx of $F_{xy}(x, y)$, this is nothing but d/dx of $F_x(x)$. Note that it is independent of y, so this becomes $f_x(x) * F_y(y)$.

Similarly, if you take the derivative with respect to y, d/dy of $F_{xy}(x, y)$, this is nothing but d/dy of $F_y(y)$, which is independent of x. So, the derivative will be $f_y(y)$.

Now, the left-hand side is the joint probability density function $f_{xy}(x, y)$, and the right-hand side is the product of the marginal probability density functions $f_x(x)$ and $f_y(y)$. Hence, $f_{xy}(x, y) = f_x(x) * f_y(y)$ for all $(x, y) \in R$. This represents the marginal probability density function of X and the marginal probability density function of Y. Hence, this is true for all $(x, y) \in R$. This is equivalent to what we are getting.

To check whether two random variables are independent or not, there are many approaches. One approach is to find their joint cumulative distribution function $F_{xy}(x, y)$ and check whether it can be expressed as the product of their marginal cumulative distribution functions $F_x(x)$ and $F_y(y)$.

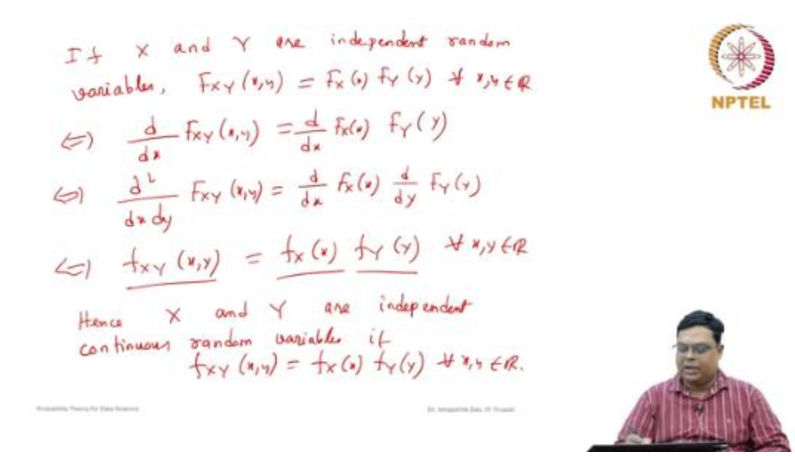
In the case of continuous variables, we can also check whether their joint probability density function $f_{xy}(x, y)$ can be represented as the product of their marginal probability density functions $f_x(x)$ and $f_y(y)$. So, if X and Y satisfy this condition, we can say that X and Y are independent continuous random variables.

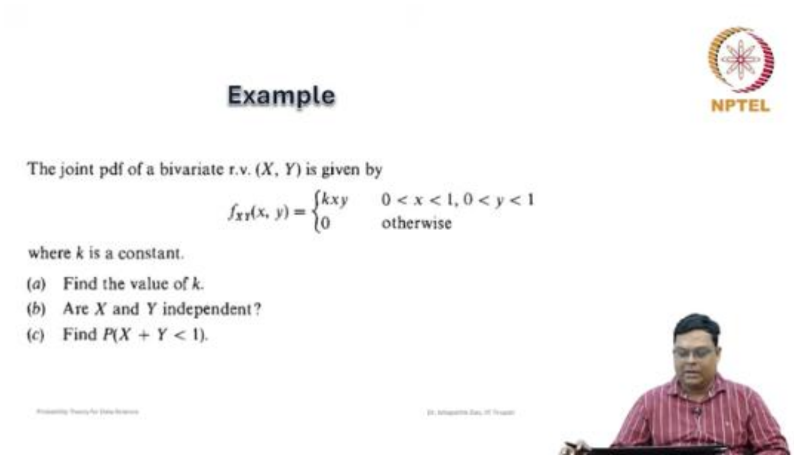
If the joint probability density function $f_{xy}(x, y)$ can be represented as the product of their marginal probability density functions $f_x(x)$ and $f_y(y)$ for all $(x, y) \in R$, then X and Y are independent random variables.

If their joint probability density function $f_{xy}(x, y)$ can be represented as the product of their marginal probability density functions, then we say that X and Y are independent random variables.

These are the concepts related to continuous random variables: their cumulative distribution function $F_{xy}(x, y)$, joint probability density function $f_{xy}(x, y)$, how to find the marginal probability density function $f_x(x)$ or $f_y(y)$ from the joint probability density function $f_{xy}(x, y)$, and how to check if two random variables are independent by verifying if the joint probability density function $f_{xy}(x, y)$ is equal to the product of their marginal probability density functions $f_x(x)$ and $f_y(y)$ for all (x, y).

So, now we can discuss some examples. Let's discuss this example. Here is an example: Suppose the joint probability density function of a bivariate random variable X and Y is given by this.





The joint probability density function $f(x, y) = k * x * y$, where $0 \leq x \leq 1$, $0 \leq y \leq 1$, and $f(x, y) = 0$ otherwise. Here, k is a constant. Find the value of k. Are X and Y independent? Find the probability that $X + Y < 1$. Let us discuss this.

The joint probability density function of a bivariate continuous random variable is given by $f(x, y) = k * x * y$, where $0 \leq x \leq 1$ and $0 \leq y \leq 1$; it is equal to 0 otherwise.

The joint probability density function of a bivariate continuous random variable is given by $f(x, y) = k * x * y$, where $0 < x < 1$ and $0 < y < 1$. Here, k is a constant, meaning it is a real number. Now, we need to find the value of k.

To find k, we check the properties of the joint probability density function. Since it is a probability density function, it should satisfy the properties of a joint probability density function.

1. $f(x, y) \geq 0$: Since k is a constant and x, y are between 0 and 1, the value of f(x, y) is positive. Therefore, k must be a positive value. If not, f(x, y) = 0 otherwise. This means $k \geq 0$.

2. The integral over the region should be equal to 1: ∫ from -∞ to ∞ ∫ from -∞ to ∞ of f(x, y) dx dy = 1. Outside the region, f(x, y) = 0. Within the region $0 \leq x \leq 1$ and $0 \leq y \leq 1$, we have:

∫ from 0 to 1 ∫ from 0 to 1 of f(x, y) dx dy = 1.

Substituting f(x, y) = k * x * y:

∫ from 0 to 1 ∫ from 0 to 1 of k * x * y dx dy = 1.

Now, perform the integration. First, integrate with respect to y, then with respect to x:

∫ from 0 to 1 [k * x * (y^2 / 2) | from 0 to 1] dx = 1.

This becomes ∫ from 0 to 1 k * x * (1^2 / 2 - 0) dx = ∫ from 0 to 1 (k * x / 2) dx = 1.

Now integrate with respect to x:

k / 2 * ∫ from 0 to 1 x^2 / 2 dx = 1.

k / 2 * (x^3 / 3 | from 0 to 1) = 1.

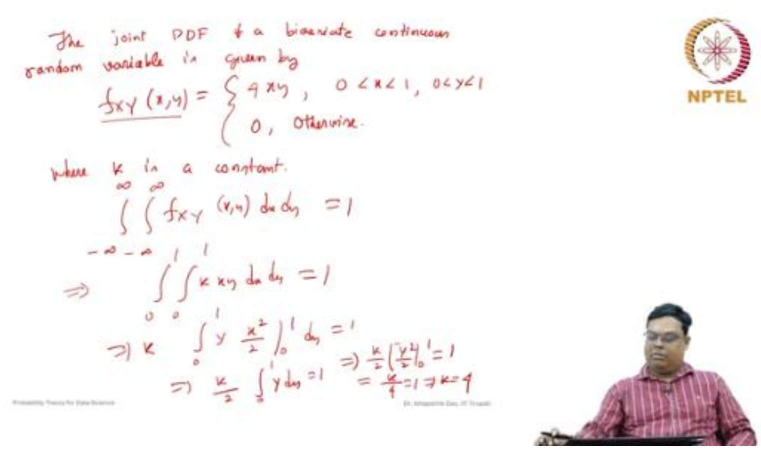
k / 2 * (1^3 / 3 - 0) = 1.

k / 2 * (1 / 3) = 1.

k / 6 = 1, so k = 6.

Therefore, the value of k is 6.

Next, we check whether X and Y are independent. To check this, we first need to find the marginal distributions.



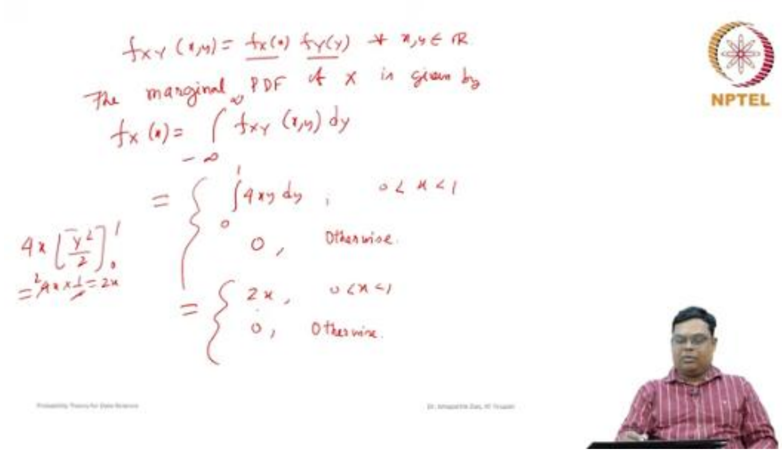
The definition of independence is that we need to show that $f_{xY}(x, y) = f_x(x) * f_Y(y)$ for all $x, y \in \mathbb{R}$. So, to make this relationship true, we first need to find $f_x(x)$ and $f_Y(y)$. To check this relationship, we need to know the marginal distributions. The marginal probability density function of X is given by the marginal probability mass function $f_x(x)$, derived from the joint probability density function. This is how we find the marginal probability density function of X. We integrate from $-\infty$ to $+\infty$ over y, considering the whole region.

The function is kxy when $0 \leq x \leq 1$. So, this is the integral from $-\infty$ to $+\infty$ of k, which we found to be 4, times 4xy dy. Otherwise, the result is 0. So, then from $-\infty$ to $+\infty$, you have to integrate 4xy dy. Since 4x is a constant, the integration will be $4x * \int y \, dy$, which is $4x * (y^2 / 2)$.

Okay, one minute. So, it is not $-\infty$ to $+\infty$. The value of 4xy is valid only when $0 \leq y \leq 1$. Similarly, x must also satisfy $0 \leq x \leq 1$. Otherwise, the integral will be 0.

So, that is why this is $4x * (y^2 / 2)$, and the limits are from 0 to 1. So, this value is 2, and the 2s cancel out. y^2 becomes 1 when you subtract 0 from 1. This gives 1 / 2. So, the 2s cancel again, and you get $4x * (1 / 2)$, which simplifies to 2x. So, finally, what we get is 2x, for $0 \leq x \leq 1$, and 0 otherwise. This is the marginal probability density function.

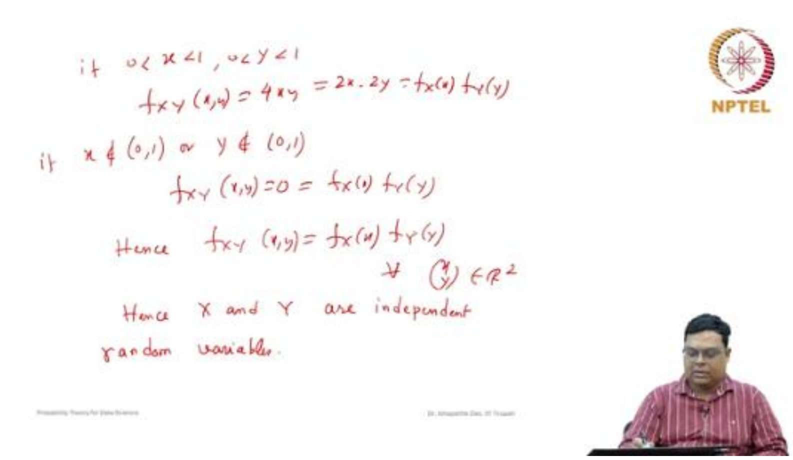
So, if the computation is correct, you can check it because it has to be a probability density function. The integration from 0 to 1 should equal 1. So, you can see that $x^2 / 2$, with 2 and 2 canceling, and in the limit from 0 to 1, this equals 1. That is why this is a probability density function. Now, similarly, the marginal probability density function of Y is given by $f_Y(y)$, which is nothing but the integral from $-\infty$ to $+\infty$ of $f_{xY}(x, y)$.



Now, we will do the integration with respect to x. So, similarly, we will do that, and you can understand now. This will be nothing but the integral from 0 to 1 of 4xy with respect to x, whenever $0 \leq y \leq 1$; otherwise, it is 0. After performing this integration, it is actually the same by symmetry. So, it is 2y whenever $0 < y < 1$; otherwise, it is 0.

Now we can see that, if $0 \leq x \leq 1$ and $0 \leq y \leq 1$, $f_{xY}(x, y)$ is nothing but 4xy, which is the same as 2x * 2y, or $f_x(x)$ * $f_Y(y)$. For any other region, if x is outside this interval or y is outside this interval, then $f_{xY}(x, y) = 0$, which is the same as $f_x(x)$ * $f_Y(y)$. So, basically, f_{xY} will be 0 if either x or y is outside this interval, or if both are outside. On the right-hand side, if x does not belong to the interval, $f_x(x)$ will be 0, making the product 0. Similarly, if y does not belong to the interval, $f_Y(y)$ will be 0, and the product will be 0.

If both are outside the region, then both will be 0. So, that is why if either of these conditions hold, we can conclude that $f_{xY}(x, y) = f_x(x) * f_Y(y)$ for all $x, y \in \mathbb{R}$, or we can say that this vector belongs to $\mathbb{R}^2$. Hence, from the definition and properties of independence, we conclude that X and Y are independent random variables. Finally, the next question is to find the probability that $X + Y < 1$. How can we compute this kind of question, specifically the probability of $X + Y < 1$?



Let us first represent this graphically, as it will help us understand the problem better. This is the region from 0 to 1, and this is where the density function exists, from 0 to 1. These are the points (0, 1), (1, 0), and (1, 1). So, this density function is non-zero inside this region. Now, the question asks about $X + Y < 1$.

The line $X + Y = 1$ is represented by this line. This line corresponds to $X + Y = 1$, and $X + Y < 1$ means we are interested in the probability of the joint random variable taking values inside this region. So, to find this probability, we need to integrate over the region where $X + Y < 1$, using the joint probability density function $f_{XY}(x, y)$ and integrating with respect to x and y. However, this is not simply 4xy for the entire region; it is only 4xy within the intersection of the square where X and Y are between 0 and 1. We need to focus on that specific region to find the correct value.

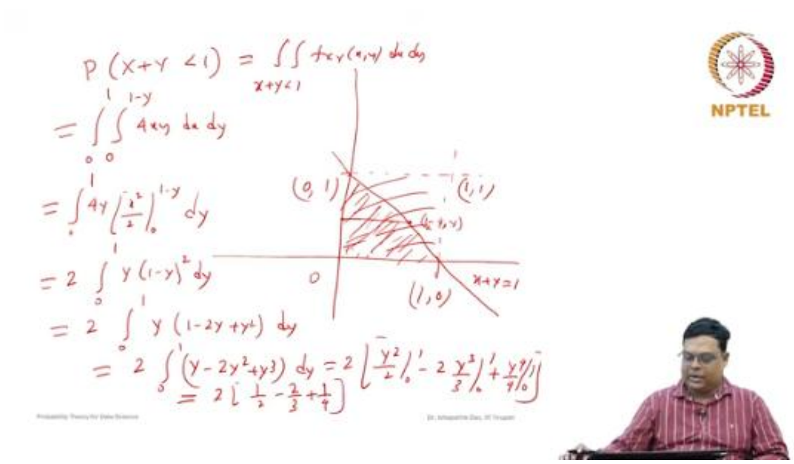
So, how will we do that? The region will be such that whenever $0 \leq x \leq 1$ or $0 \leq y \leq 1$, for any fixed value of y, x will not go beyond 1. It is limited to this value. What is this value? This value is given by the equation $X + Y = 1$.

For a fixed value of y, this value is 1 - y, and for any value of y, the x-coordinate will be 1 - y. This means the integration will go from 0 to 1 - y, with the function 4xy, then dx. The limit for y will be from 0 to 1. So, this becomes the double integral from 0 to 1, and from 0 to 1 - y, with the function 4xy, dx and then dy. So, we have to do the integration now.

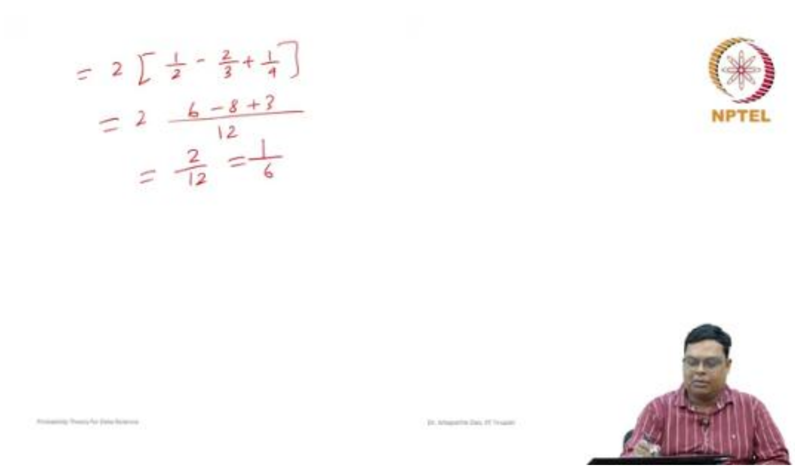
The limits are from 0 to 1, and for the inner integral, the limits are from 0 to 1 - y. The first part is 4y, which is constant here, and the integration of x^2 is divided by 2. So, we have the limits from 0 to 1 - y, and then we integrate with respect to x. After this, we get 2y. Then, we integrate from 0 to 1, and the result is y * (1 - y).

So, x^2 multiplied by $(1 - y)^2$, divided by 2, gives us $(1 - y)^2$. Then, we integrate y^2 with respect to y. We need to integrate this expression from 0 to 1 for $y * (1 - y)^2$. Expanding this, we get $1 - 2y + y^2$. Therefore, the integral becomes 2 * the integral from 0 to 1 of the expression $y - 2y^2 + y^3$ with respect to y.

Then, we need to perform the final integration. This is nothing but $y^2 / 2$, integrated from 0 to 1, minus $2y^3 / 3$, integrated from 0 to 1, plus $y^4 / 4$, integrated from 0 to 1. So then what we are getting finally is equal to 2 * (1 / 2 - 2 / 3 + 1 / 4). So let us do the final calculation, which is equal to 2 * (1 / 2 - 2 / 3 + 1 / 4). So this is equal to 2 * (1 / 2) * (1 - 2 / 3 + 1 / 4). This simplifies to 6 - 8 + 3, giving us 1 / 12.



This is nothing but 1 / 6. So, if these computations are correct, then this probability is nothing but 1 / 6. So, this is the answer. This should be the answer. Please check if these computations are correct; otherwise, you can check again to see if the answer matches.



Next, we will discuss another numerical example. I hope you have understood this numerical example, where we discussed how to find the constant k for the joint probability density function and how to check if X and Y are independent. First, we need to find the marginal probability density functions of X and Y, and then check if they are independent. After that, we will compute the probability that $X + Y < 1$.