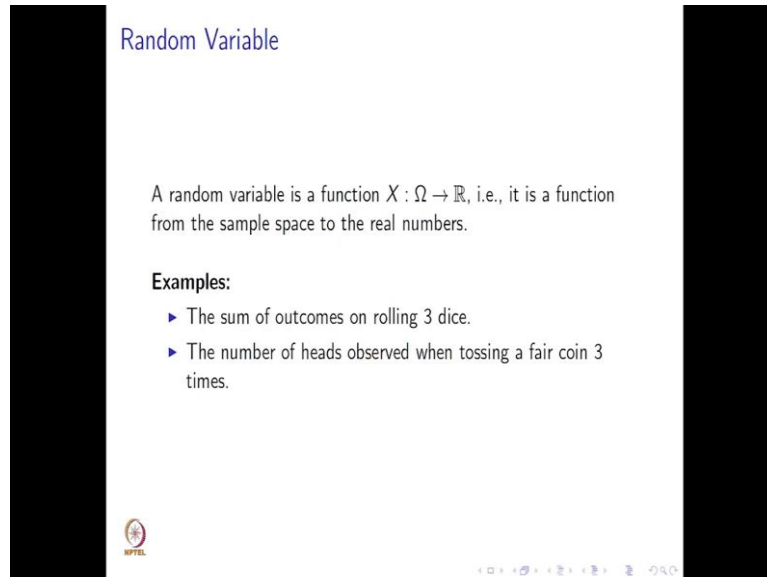


Introduction to Robotics
Professor Balaraman Ravindiran
Department of Computer Science
Indian Institute of Technology, Madras
Lecture - 32
Tutorial - 2: Probability Basics

(Refer Slide Time: 00:13)




Random Variable

A random variable is a function $X : \Omega \rightarrow \mathbb{R}$, i.e., it is a function from the sample space to the real numbers.

Examples:

- ▶ The sum of outcomes on rolling 3 dice.
- ▶ The number of heads observed when tossing a fair coin 3 times.

 ◀ ▶ ⏪ ⏩ 🔍 🔄

One of the important concepts in probability theory is that of the random variable. A random variable is a variable whose value is subject to variations. That is, a random variable can take on a set of possible different values, each with an associated probability. Mathematically, a random variable is a function from the sample space to the real numbers.

Let us consider some examples. Suppose we conduct an experiment in which we roll three dice and are interested in the sum of the outcomes. For example, the sum of 5 can be observed if two of the dice show up 2 each and the other die shows up as 1. Alternatively, the sum of 5 can also be observed if one die shows up as 3 and the other two dice show up 1 each.

Since we are interested in only the sum and not the individual results of the dice rolls, we can define a random variable, which maps the elementary outcomes, that is the outcomes of each die roll to the sum of the three rolls.

Similarly, in the next example, we can define a random variable, which counts the number of heads observed when passing a fair coin three times. Note that in this example, that random variable can take values between 0 and 3, whereas in the previous example, the range of the

random variable is between 3 and 18, corresponding to all dice showing up 1 and all dice showing up 6.

(Refer Slide Time: 01:40)

Induced Probability Function

Consider the previous example experiment of tossing a fair coin 3 times. Let X be the number of heads obtained in the three tosses. Enumerating the elementary outcomes, we observe the value of X as

ω	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
$X(\omega)$	3	2	2	2	1	1	1	0

Instead of using the probability measure defined on the elementary outcomes or events, we would ideally like to measure the probability of the random variable taking on values in its range.

x	0	1	2	3
$P_X(X = x)$	1/8	3/8	3/8	1/8

Consider the previous example experiment of tossing a fair coin three times. Let X be the number of heads obtained in the three tosses. That is, X is a random variable, which maps each elementary outcome to a real number representing the number of heads observed in that outcome. This is shown in the first table.

The first row lists out each elementary outcome and the second row lists out the corresponding real number value to which that elementary outcome is mapped, that is the number of heads observed in that outcome.

Now, instead of using the probability measure defined on the elementary outcomes or events, we would ideally like to measure the probability of that random variable taking on values in its range.

What we are trying to say here is that when we defined probability measure, we were associating each event, that is, subset of the sample space with a probability measure. When we consider random variables, the events correspond to different subsets of the sample space, which map to different values of the random variable. This is illustrated in the second table.

The first row lists out the different values that the random variable X can take and the second row lists out the corresponding probability values, assuming that the coin tossed is a fair coin. This table describes the notion of the induced probability function, which maps each possible value of the random variable to its associated probability value.

For example, in the table, the probability of the random variable taking on the value of 1 is given as 3 by 8. Since there are 3 elementary outcomes in which only one head is observed, and each of these elementary outcomes has a probability of 1 by 8.



(Refer Slide Time: 03:25)

Induced Probability Function

Let $\Omega = \{\omega_1, \omega_2, \dots\}$ be a sample space and \mathcal{P} be a probability measure (function).

Let X be a random variable with range $\mathcal{X} = \{x_1, x_2, \dots, x_m\}$.

We define the induced probability function \mathcal{P}_X on \mathcal{X} as

$$\mathcal{P}_X(X = x_i) = \mathcal{P}(\{\omega_j \in \Omega : X(\omega_j) = x_i\})$$



From the previous example, we can define the concept of the induced probability function. Let Omega be a sample space and P be a probability measure. Let X be a random variable, which takes values in the range X1 to Xm. The induce probably function PX on X is defined as PX, X equals to small xi equals to the probability of the event, comprising of the elementary outcomes, small omega j such that the random variable X maps small omega j to the value xi.

(Refer Slide Time: 04:06)



Cumulative Distribution Function

The cumulative distribution function or cdf of a random variable X , denoted by $F_X(x)$, is defined by

$$F_X(x) = \mathcal{P}_X(X \leq x), \text{ for all } x$$

Example:

x	$(-\infty, 0]$	$(-\infty, 1]$	$(-\infty, 2]$	$(-\infty, 3]$	$(-\infty, \infty)$
$F_X(x)$	$1/8$	$1/2$	$7/8$	1	1

Induced Probability Function

Consider the previous example experiment of tossing a fair coin 3 times. Let X be the number of heads obtained in the three tosses. Enumerating the elementary outcomes, we observe the value of X as

ω	HHH	HHT	HTH	TTH	THT	HTT	TTT
$X(\omega)$	3	2	2	1	1	1	0

Instead of using the probability measure defined on the elementary outcomes or events, we would ideally like to measure the probability of the random variable taking on values in its range.

x	0	1	2	3
$P_X(X=x)$	$1/8$	$3/8$	$3/8$	$1/8$

The cumulative distribution function or CDF of a random variable X denoted by F_X of small x is defined by F_X of small x equals to the probability of the random variable taking on a value less than or equal to small x for all values of small x .

For example, going back to the previous random variable, which counts the number of heads observed in three tosses of a fair coin, the following table shows the intervals corresponding to the different values of the random variable X along with the corresponding values of the cumulative distribution function.

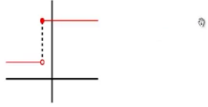
For example, F_X equals, F_X of 1 equals to $1/2$ because the probability that the random variable x has a value of 1, let us just go back to the previous example. Right. The probability that the random variable X has a value of 1 is $3/8$. The probability of x , that the random variable X equals to 0 is $1/8$, and therefore, the probability that the random variable X takes on a value less than or equals to 1 is $1/8$ plus $3/8$ equal to $4/8$ or $1/2$.


(Refer Slide Time: 05:23)

Properties of cdf

A function $F_X(x)$ is a cdf iff the following three conditions hold:

- ▶ (Monotonicity) If $x \leq y$, then $F_X(x) \leq F_X(y)$
- ▶ (Limiting values) $\lim_{x \rightarrow -\infty} F_X(x) = 0$ and $\lim_{x \rightarrow \infty} F_X(x) = 1$
- ▶ (Right-continuity) For every x , we have $\lim_{y \downarrow x} F_X(y) = F_X(x)$





A function is a valid cumulative distribution function only if it satisfies the following properties. The first property simply states that the cumulative distribution function is a non-decreasing function.

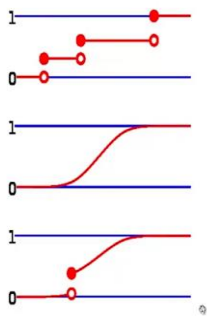
The second property specifies the limiting values. Limit x tends to minus infinity F_X of x equals to 0 and limit extends to infinity F_X of x equals to 1. The third property specifies right-continuity. That is no jump occurs when the limit point is approached from the right. This is also shown in the figure below.


(Refer Slide Time: 06:02)

Continuous & Discrete Random Variables

A random variable X is continuous if $F_X(x)$ is a continuous function of x .

A random variable X is discrete if $F_X(x)$ is a step function of x .





A random variable X is continuous if its corresponding cumulative distribution function is a continuous function of X . This is shown in the second part of the diagram. A random variable X is discrete if its CDF is a step function of x . This is shown in the first part of the diagram.

The third part of the diagram shows the cumulative distribution function for a random variable, which has both continuous and discrete parts.

(Refer Slide Time: 06:30)

Probability Mass Function

The probability mass function or pmf of a discrete random variable X is given by

$$f_X(x) = P(X = x), \text{ for all } x$$

Example: For a geometric random variable X with parameter p ,

$$f_X(x) = \begin{cases} (1-p)^{x-1}p & \text{for } x = 1, 2, \dots \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Properties:

- ▶ $f_X(x) \geq 0$, for all x
- ▶ $\sum_x f_X(x) = 1$

NPTEL

The probability mass function or PMF of a discrete random variable X is given by, f_X of x equal to probability of X equal to small x , for all values of small x . Thus for a discrete random variable, the probability mass function of that random variable gives the probability that the random variable is equals to some value.

For example, for a geometric random variable X with parameter p , the PMF is given as f_X of x equals to $1 - p$ raised to the power $x - 1$ into p for the values of x equals to $1, 2$, and so on. And for other values of x , the PMF will equals to 0 .

A function is a valid probability mass function if it satisfies the following two properties. First of all, the function must be non-negative. Secondly, the summation over all X , the value of the function summed over all values of X should be equals to 1 .

(Refer Slide Time: 07:32)

Probability Density Function

The probability density function or pdf of a continuous random variable is the function $f_X(x)$ which satisfies

$$F_X(x) = \int_{-\infty}^x f_X(t) dt, \text{ for all } x$$

Properties:

- ▶ $f_X(x) \geq 0$, for all x
- ▶ $\int_{-\infty}^{\infty} f_X(x) dx = 1$

MPTEL

For continuous random variables, we considered the probability density function. The probability density function or PDF of a continuous random variable is the function f_X of x , which satisfies the following. The integral from minus infinity to x , f_X of t dt is equals to the cumulative distribution function at the point X .

Similar to the PMF, the probability density function should also satisfy the following properties. First of all, the probability density function should be non-negative for all values of x . Second, integrating over the entire range, the probability density function should sum to 1.

(Refer Slide Time: 08:17)

Expectation

The expected value or mean of a random variable X , denoted by $E[X]$, is given by

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx \text{ (continuous RV)}$$
$$E[X] = \sum_{x: P(x) > 0} x f_X(x) = \sum_{x: P(x) > 0} x P(X = x) \text{ (discrete RV)}$$

MPTEL

Let us now look at expectations of random variables. The expected value or mean of a random variable X denoted by expectation of X is given by integral minus infinity to infinity x into f_X of x dx . Note that f_X of x here is the probability density function associated with random variable X . This definition holds when X is a continuous random variable.

In case that X is a discrete random variable, we use the following definition. Expectation of X is equal to sum overall x , such that probability of x greater than 0. That is we consider all values of the random variable for which the associated probability is greater than 0, x into f_X of x .

Here, f_X of x is the probability mass function of the random variable X , which essentially gives the associated probability for a particular value of the random variable, thus leading to this definition.

(Refer Slide Time: 09:14)


Example

Q. Let the random variable X take values $-2, -1, 1, 3$ with probabilities $1/4, 1/8, 1/4, 3/8$ respectively. What is the expectation of the random variable $Y = X^2$?

Sol. The random variable Y takes on the values $1, 4, 9$ with probabilities $3/8, 1/4, 3/8$ respectively.
Hence,

$$E(Y) = \sum_x xP(Y=x) = 1 \cdot \frac{3}{8} + 4 \cdot \frac{1}{4} + 9 \cdot \frac{3}{8} = \frac{19}{4}$$

Alternatively,

$$E(Y) = E(X^2) = \sum_x x^2P(X=x) = 4 \cdot \frac{1}{4} + 1 \cdot \frac{1}{8} + 1 \cdot \frac{1}{4} + 9 \cdot \frac{3}{8} = \frac{19}{4}$$


Let us now look at an example in which we calculate expectations. Let the random variable X take values minus 2, minus 1, 1, and 3 with probabilities $1/4, 1/8, 1/4, 3/8$ respectively. What is the expectation of the random variable Y equals to X square?

So in this question, we are given one random variable, the values which this random variable takes, and its associated probabilities. What we are interested is in the expectation of the random variable Y , which is defined as Y equals to X square.

So what we can do is, we can calculate the values that the random variable Y takes along with associated probabilities, since we are aware of the relation between Y and X . Thus, we

have Y taking on the values 1, 4, and 9 with probabilities 3 by 8, 1 by 4, and 3 by 8 respectively.

Given this information, we can simply apply the formula for expectation and calculate the expectation on the random variable Y. This is as follows, giving a result of 19 by 4. Another way to approach this problem is to directly use the relation Y equals to X square in calculating the expectation. Thus, expectation of Y is simply the expectation of the random variable X squared.

So in play, in the formula for expectation, instead of substituting X, we substitute X square. Thus, we have sum of over all x, x square into probability of X equal to x. Calculating the values, we get the same answer of 19 by 4.

(Refer Slide Time: 10:51)

Properties of Expectations

Let X be a random variable and let a, b, c be constants. Then, for functions $g_1(X)$ and $g_2(X)$ whose expectations exist

- ▶ $E(ag_1(X) + bg_2(X) + c) = aEg_1(X) + bEg_2(X) + c$
- ▶ If $g_1(X) \geq 0$ for all x , then $Eg_1(X) \geq 0$
- ▶ If $g_1(X) \geq g_2(X)$ for all x , then $Eg_1(X) \geq Eg_2(X)$
- ▶ If $a \leq g_1(X) \leq b$, for all x , then $a \leq Eg_1(X) \leq b$

SPTEL

Let us now look at the properties of expectations. Let X be a random variable; a, b , and c are constants, and g_1 and g_2 are functions of the random variable X such that their expectations exist, that is, they have finite expectations.

According to the first property, expectation of a into g_1 of X plus b times g_2 of X plus c is equals to a times expectation of g_1 of X plus b times expectation of g_2 of X plus c . This is called the linearity of expectations. There are actually a few things to note here.

First of all, expectation of a constant as equals to the constant itself, expectation of a constant times the random variable is equals to the constant into the expectation of the random variable, and the expectation of the sum of two random variables can also be represented as

the sum of the expectations of the two random variables. Note that here, the two random variables need not be statistically independent.

According to the next property, if a random variable is greater than equals to 0 at all points, then the expectation is also, expectation of that random variable is also greater than equals to 0. Similarly, if one random variable is greater than another random variable at all points, then the expectation of those random variables also follow the same constraint.

Finally, if a random variable has values, which are, which lie between two constants, then the expectation of that random variable will also lie between those two constants.

(Refer Slide Time: 12:31)

Moments

For each integer n , the n^{th} moment of X is

$$\mu'_n = EX^n$$

The n^{th} central moment of X is

$$\mu_n = E(X - \mu)^n$$

The slide includes a small logo at the bottom left and navigation icons at the bottom right.

Let us now define moments. For each integer n , the n^{th} moment of X is μ'_n equals to expectation of X raised to the power n . Also, the n^{th} central moment of X is μ_n equals to expectation of X minus μ raised to the power n . So the difference between moment and central moment is, in central moment, we subtract the random variable by the mean of the random variable or expected value.

The two moments that find most common use are the first moment, which is nothing but μ'_1 equals to expectation of X , that is, the mean of the random variable X and the second central moment, which is μ_2 equals expectation of X minus μ raised to the power 2, which is the variance of the random variable X .

(Refer Slide Time: 13:18)

Variance

The variance of a random variable X is its second central moment.
$$\text{Var}X = E(X - \mu)^2 = E(X - EX)^2 = EX^2 - (EX)^2$$

The positive square root of $\text{Var}X$ is the standard deviation of X .

Note: $\text{Var}(aX + b) = a^2 \text{Var}X$
where a, b are constants

MPTEL

Thus the variance of a random variable X is its second central moment, variance of X equals to expectation of X minus μ whole square. Note that μ is just the first moment, which can be replaced. So it can be replaced by expectation of X . Thus, we have variance of X is equal to expectation of X minus expectation of X whole square.

By expanding this term and applying linear expectations, we will finally get variance of X equals to expectation of X squared minus square of the expectation of X . The positive square root of variance of X is a standard deviation of X . Note that the, when calculating variance, the constants act differently when compared to the linearity of expectation.

This is a very useful relation to remember. Variance of aX plus b is equal to a square into variance of X , where A and B are constants.

(Refer Slide Time: 14:13)

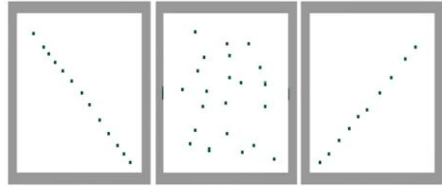
Covariance

The covariance of two random variables, X and Y is

$$\text{cov}(X, Y) = E[(X - EX)(Y - EY)]$$

It is a measure of how much two random variables change together.

COVARIANCE



Large Negative Covariance Near Zero Covariance Large Positive Covariance

NPTEL

The covariance of two random variables X and Y is covariance of X comma Y equals to expectation of X minus expectation of X into Y minus expectation of Y. Remember that the variance of a random variable X is nothing but the second central moment. Thus, the variance of a random variable measures the amount of separation in the values of the random variable when compared to the mean of the random variable.

For covariance, the calculation is done on a pair of random variables and it measures how much two random variables change together. Consider the diagram below. In the first part, assume that the random variable X is on the x-axis and the random variable Y is on the y-axis. We note that as the value of X increases, the value of Y seems to be decreasing. Thus, for this relationship, we will observe a large negative co-variance.

Similarly, in the third part of the diagram, we can see that as the value of variable X increases so does the value of the variable Y. Thus, we see a large positive covariance. However, in the middle diagram, we cannot make any such statement because as X increases, there is no clear relationship as to how Y changes. Thus, this kind of a relationship will give zero covariance.

Now from the diagram, it should immediately be clear that covariance is a very important term in machine learning because we are often interested in predicting the value of one variable by looking at the value of another variable. We will come to that in further classes.

(Refer Slide Time: 16:00)

Correlation

The correlation of two random variables, X and Y is

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$$

Note:

- ▶ For correlation to be defined, individual variances must be non-zero and finite
- ▶ $\rho(X, Y)$ lies between -1 and $+1$

NPTEL

Closely related to the concept of covariance is the concept of correlation. The correlation of two random variables X and Y is nothing but the covariance of the two random variables X and Y divided by the square root of the product of their individual variances.

Basically, correlation is a normalized version of covariance. So the correlation will always be between minus 1 and 1. Also, since we used the variance of the individual random variables in the denominator for correlation to be defined, individual variances must be non-zero and finite.

(Refer Slide Time: 16:39)

Probability Distributions

Consider two variables X and Y, and suppose we know the corresponding probability mass functions f_X and f_Y

Can we answer the following question:

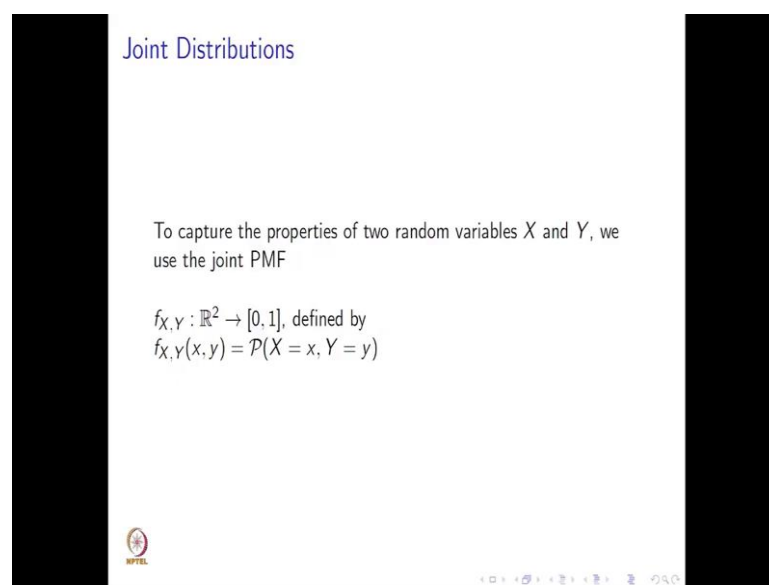
$$P(X = x \text{ and } Y = y) = ?$$

NPTEL

In the final part of this tutorial on probability theory, we will talk about probability distributions and list out some of the more common distribution that you are going to encounter in the course. Before we proceed, let us considered this question.

Consider two variables, X and Y , and suppose we know the corresponding probability mass function f_X and f_Y corresponding to the variables X and Y . Can we answer the following question? What is the probability that X takes a certain value small x and Y takes a certain value small y . Think about this question. If you answered no, then you are correct. Let us see why.

(Refer Slide Time: 17:21)



Joint Distributions

To capture the properties of two random variables X and Y , we use the joint PMF

$$f_{X,Y} : \mathbb{R}^2 \rightarrow [0, 1], \text{ defined by}$$
$$f_{X,Y}(x,y) = \mathcal{P}(X = x, Y = y)$$

The slide includes a small logo at the bottom left and navigation icons at the bottom right.

Essentially, what we were looking for in the previous question was the joint distribution, which captures the properties of both random variables. The individual PMFs or PDFs in case that random variables are continuous, capture the properties of the individual random variables only but miss out on how the two variables are related.

Thus, we define the joint PMF or PDF, $f_{X, Y}$ as the probability that X takes on a specific value small x and Y takes on a specific values small y for all values of X and Y .

(Refer Slide Time: 17:57)

Marginal Distributions

Suppose we are given the joint PMF

$$f_{X,Y}(x,y) = \mathcal{P}(X = x, Y = y)$$

From this joint PMF, we can obtain the PMF's of the two random variables

$$f_X = \sum_y f_{X,Y}(x,y) \quad (\text{marginal PMF of R.V. } X)$$
$$f_Y = \sum_x f_{X,Y}(x,y) \quad (\text{marginal PMF of R.V. } Y)$$

MPTEL

Suppose we are given the joint probability mass function of the random variables X and Y. What if we are interested in only the individual mass functions of either of the random variables? This can be obtained from the joint probability mass function by a process called marginalization.

The individual probability mass function thus obtained is also referred to as the marginal probability mass function. Thus, if we are interested in the marginal probability mass function of random variable X, we can obtain this by summing the joint probability mass function over all values of Y.

Similarly, the probability mass function of, the marginal probability mass functional of random variable Y can be obtained by summing the joint probability mass function for all values of X. Note that in case the random variables considered here are continuous, we substitute summation by integration and PMFs by PDFs.

(Refer Slide Time: 18:59)

Conditional Distributions



Like joint distributions, we can also consider conditional distributions

$$f_{X|Y}(x|y) = \mathcal{P}(X = x | Y = y)$$

Using conditional probability definition, we have

$$f_{X|Y}(x|y) = f_{X,Y}(x, y) / f_Y(y)$$

Note that the above conditional probability is undefined if $f_Y(y) = 0$.



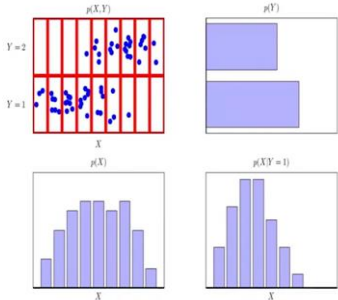


Like joint distributions, we can also consider conditional distributions. For example, here we have the conditional distribution f_X given Y , which is the probability that the random variable X will take on a some value small x , given that the random variable Y has been observed to take on a specific value small y .

The relation between conditional distributions, joint distribution, and marginal distributions is shown here. This relation should be familiar from the definition of conditional probability that was seen earlier. Note that the marginal distribution f_Y of y is in the denominator and hence, it must not be equals to 0.

(Refer Slide Time: 19:41)

Example



The overall idea of joint, marginal, and conditional distributions is summarized in this figure. The top left-figure shows the joint distribution and describes how the random variable X, which takes on nine different values is it related to the random variable, Y which takes on two different values.

The bottom-left figure shows the marginal distribution of random variable X. As can be observed in this figure, we simply ignore the information related to the random variable Y. Similarly, the top-right figure shows the marginal distribution of a random variable Y.

Finally, the bottom-right figure showed the conditional distribution of X, given that the random variable Y takes on a value of 1. Looking at this figure and comparing it with the joint distribution, we observe that in the bottom-right figure, we simply ignore all the values of X for which Y equals to 2. That is the top half of the joint distribution.

(Refer Slide Time: 20:43)

Bernoulli Distribution

Consider a random variable X taking one of two possible values (either 0 or 1). Let the PMF of X be given by

$$f_X(0) = \mathcal{P}(X = 0) = 1 - p \quad (0 \leq p \leq 1)$$
$$f_X(1) = \mathcal{P}(X = 1) = p$$

This describes a Bernoulli distribution

$$E[X] = p$$
$$\text{var}(X) = p(1 - p)$$

MPTEL

In the next few slides, we will present some specific distributions that you will be encountering in the machine learning course. We will present a definition and list out some important properties for each distribution. It would be a good exercise for you to work out the expressions for the PMFs or PDFs and the expectation and variances of these distributions on your own.

We start with our Bernoulli distribution. Consider a random variable X taking one of two possible values, either 0 or 1. Let the PMF of X be given by f_X of 0 is equal to the probability that the random variable X takes on a value of 0 equals to 1 minus p, where p lies between 0. And f_X of 1 equals to probability that the random variable X takes the value 1 equals to p.

Here P is the parameter associated with the Bernoulli distribution. It generally refers to the probability of success. So in our definition, we are assuming that X is equal to 1 indicates a successful trial and X equals to 0 indicates a failure.

The expectation of a random variable following the Bernoulli distribution is p and the variance is p into 1 minus p . The Bernoulli distribution is very useful to characterize experiments which have a binary outcome, such as in tossing a coin, we observe either heads or tails or say, in writing an exam, where you have pass or fail. Such experiments can be modeled using the Bernoulli distribution.

(Refer Slide Time: 22:14)

Binomial Distribution

Consider the situation where we perform n independent Bernoulli trials where

- ▶ probability of success (for each trial) = p
- ▶ probability of failure = $1 - p$

Let X be the number of successes in the n trials, then we have

$$\mathcal{P}(X = x|n, p) = \binom{n}{x} p^x (1 - p)^{n-x}$$

where $\binom{n}{x} = \frac{n!}{(n-x)!x!}$
and $0 \leq x \leq n$

$$E[X] = np$$

$$\text{var}(X) = np(1 - p)$$

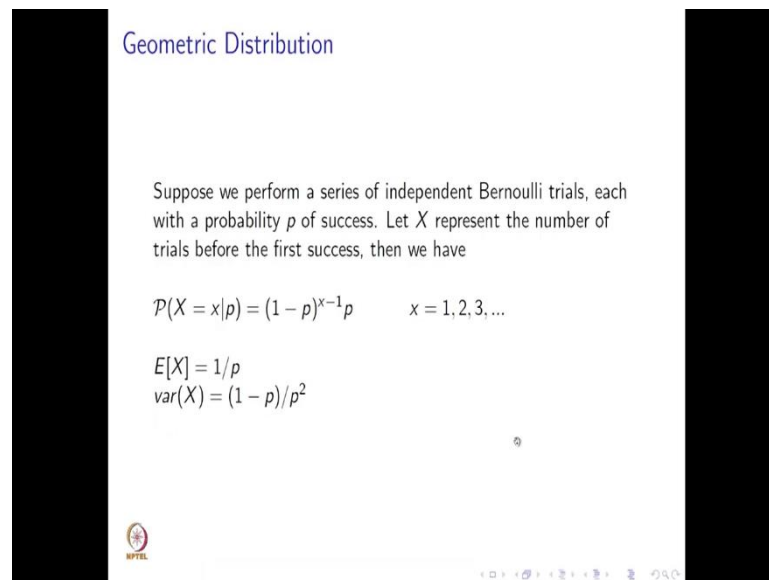
Next, we look at the binomial distribution. Consider the situation where we perform n independent Bernoulli trials, where the probability of success for each trial equals to p and the probability of failure for each trial equals to 1 minus p .

Let X be the random variable, which represents the number of successes in the end trials. Then we have probability that the random variable X will take on a specific value of small x , given the parameters n and p equals to n choose X , that is, the number of combinations of observing X successes and in n trials into p raised to the power x into 1 minus p raised to the power n minus x .

Note that here x is going to be a number between 0 and n . The expectation of a random variable following the binomial distribution equals to np , and the variance equals to n into p into 1 minus p . The binomial distribution is useful in any scenario where are conducting multiple Bernoulli trials, that is experiments in which the outcome is binary.

For example, suppose we have a coin, suppose we toss a coin 10 times, and want to know the probability of observing three heads. Given the probability of observing a head in an individual trial, we can apply the binomial distribution to find out the required probability.

(Refer Slide Time: 23:36)



Geometric Distribution

Suppose we perform a series of independent Bernoulli trials, each with a probability p of success. Let X represent the number of trials before the first success, then we have

$$P(X = x|p) = (1 - p)^{x-1}p \quad x = 1, 2, 3, \dots$$
$$E[X] = 1/p$$
$$\text{var}(X) = (1 - p)/p^2$$

The slide includes a small NPTEL logo at the bottom left and navigation icons at the bottom right.

Suppose we perform a series of independent Bernoulli trials each with a probability p of success. Let X represent the number of trials before the first success, then we have probability that the random variable X will take a value small x given the parameter p is equals to 1 minus p raised to the power X minus 1 into p . This definition is quite intuitive.

Essentially, we are trying to calculate the probability that it takes us small x number of trials before observing a first success. This can happen if the first x minus 1 trials failed, that is, with probability 1 minus p and the trial succeeded, that is, with probability p .

A random variable which has the, this probability mass function follows the geometric distribution. For the geometric distribution, the expectation of the random variable equals to 1 by p and the variance equals to 1 minus p by p square.

(Refer Slide Time: 24:37)

Uniform Distribution

A continuous random variable X is said to be uniformly distributed on an interval $[a, b]$ if its PDF is given by

$$f_X(x|a, b) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

$E[X] = (a + b)/2$
 $\text{var}(X) = (b - a)^2/12$

MPTCL

In many situations, we initially do not know the probability distribution of the random variable under consideration but can perform experiments which will gradually reveal the nature of the distribution. In such a scenario, we can use the uniform distribution to assign uniform probabilities to all values of the random variable, which are then later updated.

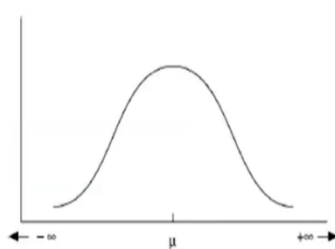
In the discrete case, say the random variable can take n different values. Then we simply assign a probability of $1/n$ to each of those n values. In the continuous case, if the random variable X takes values in the closed interval, a comma b , then its PDF is given by f_X of x given parameters a comma b equals to $1/(b - a)$, if x lies in the closed interval, a comma b , and 0 , otherwise.

For a random variable following the uniform distribution, the expectation of the random variable X equals to $(a + b)/2$, and the variance equals to $(b - a)^2/12$.

(Refer Slide Time: 25:46)

Normal Distribution

A continuous random variable X is said to be normally distributed with parameters μ and σ^2 if the density of X is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad -\infty < x < \infty$$


The diagram shows a bell-shaped curve representing the normal distribution. The x-axis is labeled with $-\infty$, μ , and $+\infty$. The y-axis represents the probability density function. The curve is symmetric and centered at μ .

A continuous random variable X is said to be normally distributed with parameters μ and σ^2 if the PDF of the random variable X is given by the following expression. The normal distribution is also known as the Gaussian distribution and is one of the most important distributions that we will be using. The diagram represents the famous bell-shaped curve associated with the normal distribution.

(Refer Slide Time: 26:12)

Importance of Normal Distribution

Roughly, the central limit theorem states that the distribution of the sum (or average) of a large number of independent, identically distributed variables will be approximately normal, regardless of the underlying distribution.

Multivariate Normal Distribution

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

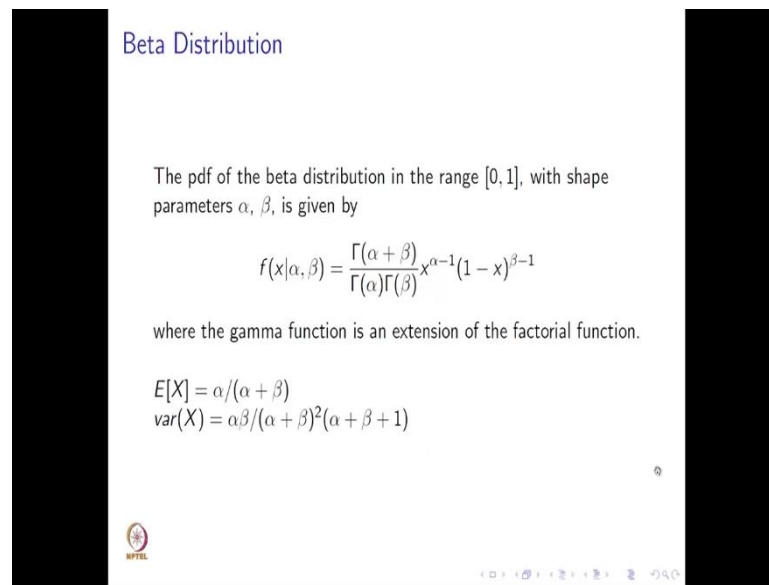
where

- ▶ μ is the D -dimensional mean vector,
- ▶ Σ is the $D \times D$ covariance matrix, and
- ▶ $|\Sigma|$ is the determinant of the covariance matrix

The importance of the normal distribution is due to the central limit theorem. Without going into the details, the central limit theorem roughly states that the distribution of the sum of a large number of independent, identically distributed variables will be approximately normal, regardless of the underlying distribution.

Due to this theorem, many physical quantities that are the sum of many independent processes, often have distributions that can be modeled using the normal distribution. Also, in the machine learning course, we will be often using the normal distribution in its multivariate form. Here, we are presented the expression of the multivariate normal distribution, where μ is the D-dimensional mean vector and Σ is D cross D covariance matrix.

(Refer Slide Time: 27:02)



Beta Distribution

The pdf of the beta distribution in the range $[0, 1]$, with shape parameters α, β , is given by

$$f(x|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$$

where the gamma function is an extension of the factorial function.

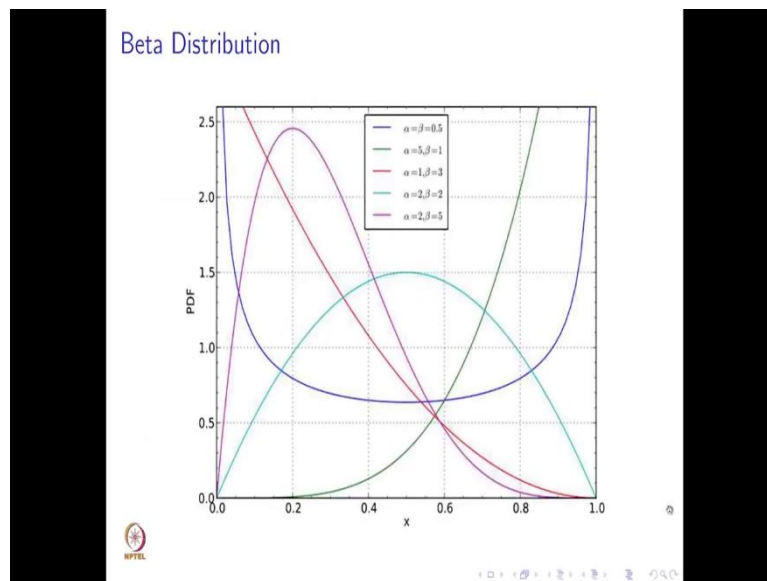
$$E[X] = \alpha / (\alpha + \beta)$$
$$\text{var}(X) = \alpha\beta / (\alpha + \beta)^2 (\alpha + \beta + 1)$$

MPTEL

The PDF of the beta distribution in the range 0 to 1 with shape parameters, alpha and beta is given by the following expression, where the gamma function is an extension of the factorial function.

The expectation of a random variable following the beta distribution is given by alpha by alpha plus beta. And the variance is given by alpha beta by alpha plus beta whole square into alpha plus beta plus 1.

(Refer Slide Time: 27:29)



This diagram illustrates the beta distribution. Similar to the normal distribution in which the shape and position of the bell-curve is controlled by the parameters μ and σ^2 , in the beta distribution, the shape of the distribution is controlled by the parameters α and β .

In the diagram, we can see a few instances of the beta distribution for different values of that shape parameters. Note that unlike the normal distribution, a random variable following the beta distribution takes values only in a fixed interval. Thus, in this example, the probability that that variable takes a value less than 0 or greater than 1 is equals to 0.

This ends the first tutorial on the basics of probability theory. If you have any doubts or seek clarifications regarding the material covered in this tutorial, please make use of the forum to ask questions.

As mentioned in the beginning, if you are not comfortable with any of the concepts presented here, do go back and read up on it. There will be some questions from probability theory in the first assignment, so hopefully, going through this tutorial will help you in answering those questions. And note that the, we will be having another tutorial next week on linear algebra.