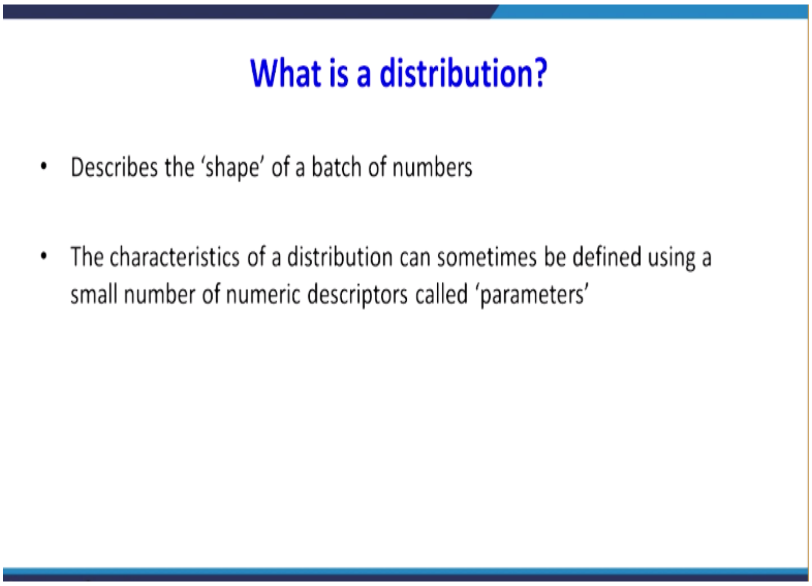
Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 08
Probability Distributions

Good morning students we are entering to the 8th lecture on this course that is a data analytics with the Python. Today the topic is probability distributions. So what we are going to cover today is very interesting topic. We are going to see the some empirical distribution and its properties. Then discrete distribution in the discrete distribution we are going to see Binomial, Poisson, Hyper geometric distributions. The continuous distribution we are going to see the uniform, exponential, normal distribution.

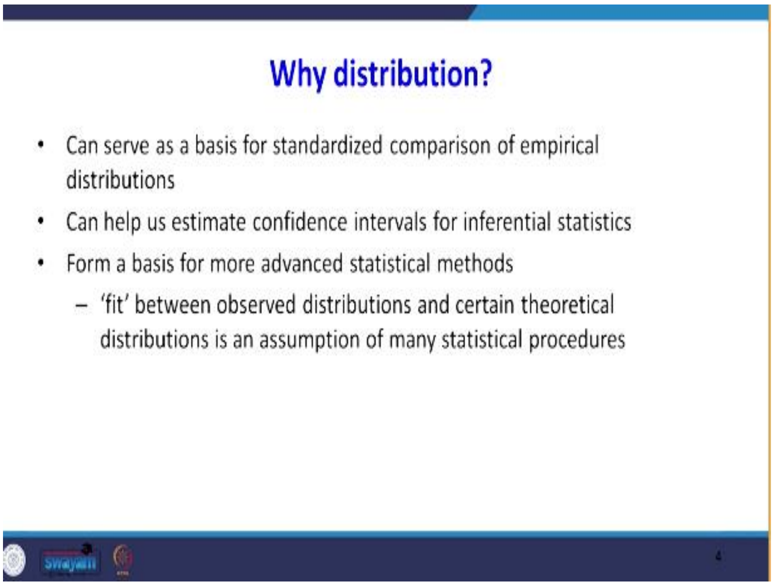
(Refer Slide Time: 01:00)



First up all what is distribution? What is the purpose of studying the distribution? The distributions describe the shape of a batch of numbers that is the meaning of distribution. Suppose the different set of numbers there, you want to show what shape it follows whether it is a bell shaped, we can call it is a normal distribution. If it is forming a rectangular shape, we can call it as a uniform distribution like this that describes the shape of a batch of numbers.

The characteristics of your distribution can sometimes be defined as a small number of numerical descriptors called parameters. So each distributions characteristic is expressed with the help of its parameters. Parameter is nothing but for example normal distribution it has 2 parameter one is mean and variance with the help of that you can draw the distribution that is a parameter.

(Refer Slide Time: 01:53)

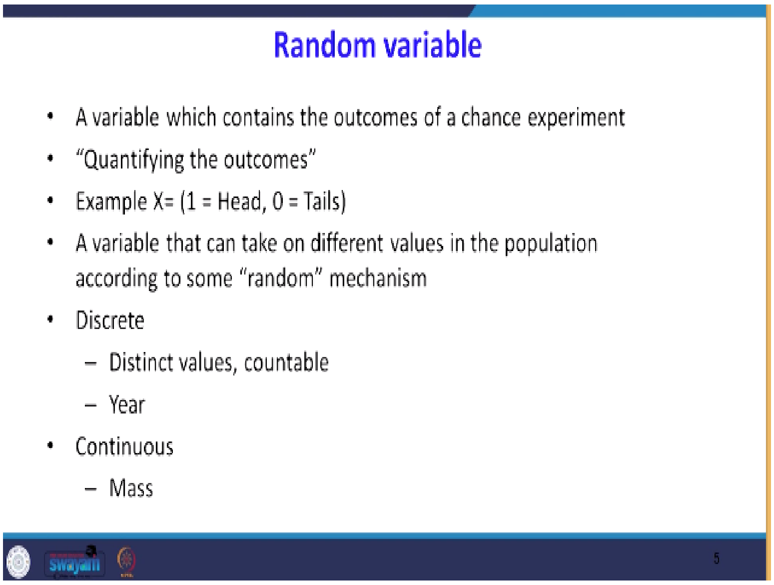


Why distribution? Can serve as a basis for standardized the comparison of empirical distributions because if you want compare with phenomena with the very standard distributions we can come to know that what distribution it follows then it will help you to estimate the confidence intervals for inferential statistics that will see what is the meaning of conference interval incoming classes then form a basis for more advanced statistical methods.

For example, fit between observed distribution and certain theoretical distribution is an assumption of many statistical procedures. Suppose why we have to study the distributions, suppose we are doing your simulation for example the arrival pattern follow Poisson distributions. Suppose certain data you collected if you prove that it is arrival follow Poisson distribution already there is a mean and variance and other population parameters already defined it.

If you are a natural phenomena, you are able to compare with standard distributions that are well defined distribution parameter is there that parameter you can use as it that is a purpose of studying the distribution.

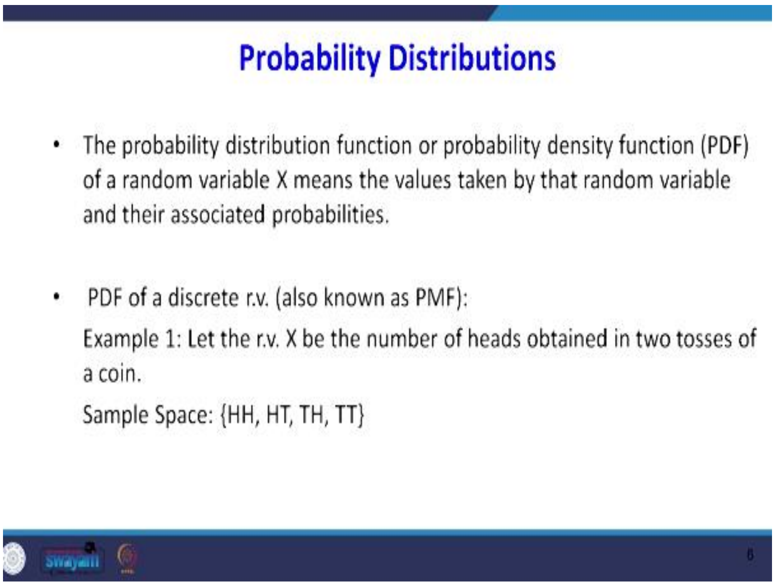
(Refer Slide Time: 03:01)



Then we will go for what is the random variable we want to construct a distribution, it is the relation between X and corresponding probability X, p of x. So here the X is nothing but random variable. A variable which contains the outcome of chance experiment is the random variable is the kind of you are quantifying the outcome suppose we task of the coin for X = 1 is getting head, 0 getting tails. So 1 is nothing but your random variable.

So the X, the X is the random variable, that can take the value of 1 and 0, the X value is 1 it is the head. If X value is 0 it is a tail. Variable that can take on different values in the population according to some random mechanism. So the value of 1 and 0, it follows certain mechanism. Random variable can be a discrete it may be distinct values, countable. For example year is a discrete random variable. For example Mass it is a continuous random variable.

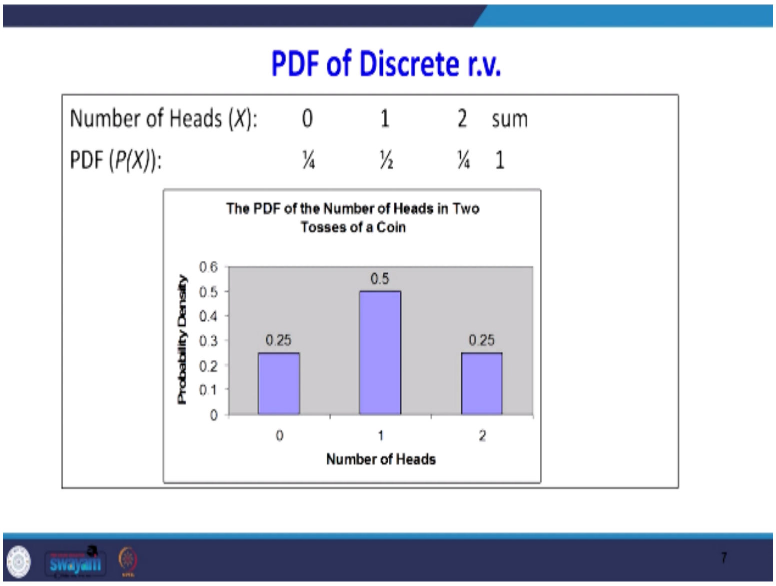
(Refer Slide Time: 04:02)



Then probability distributions, the probability distribution function or probability density function PDF of the random variable X means the values taken by the random variable and their associated probabilities if you make a relation between X and corresponding probabilities p of x or f of x, that if you plot that point that will form your distributions, so PDF of your discrete random variable also known as PMF probability mass function.

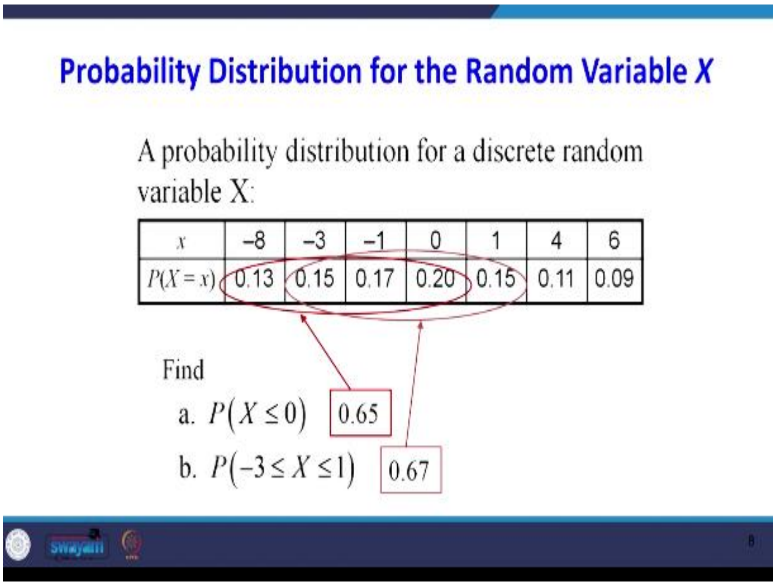
Example let the random variable X be the number of heads obtained in you 2 tosses of your coin. There are 2 possibilities when you toss 2 times 2 tosses, first toss you may get head, second toss you may get head then head tail, tail head, tail tail, so these are the sample space.

(Refer Slide Time: 04:53)



Probability density functions of your discrete random variable. Suppose we are tossing coin 2 times the probability of the 0 head is 1 by 4, the probability of getting one head is 1 by 2, the probability of getting 2 heads 1 by 4 some should be 1. See in the in the X axis, the random variable is taken 0 in Y axis corresponding probabilities marked. So, in X axis random variable and Y axis corresponding probability this is called the distributions.

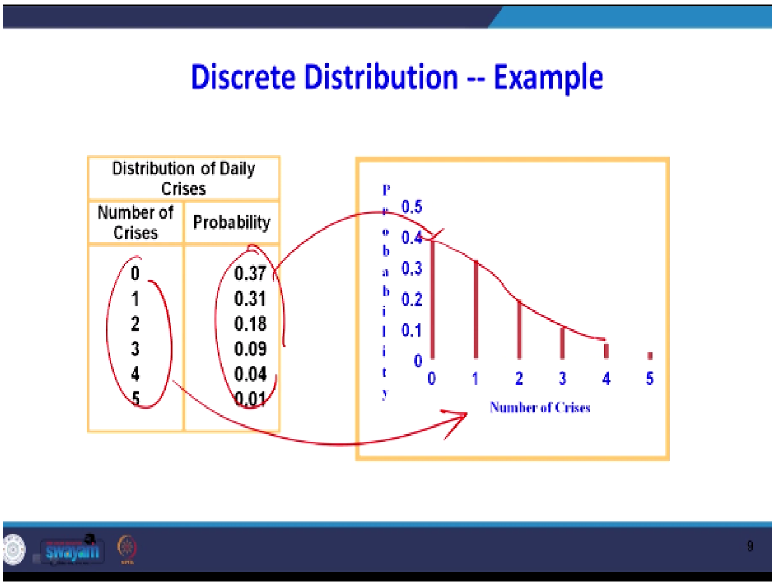
(Refer Slide Time: 05:23)



Now, probability distribution for a random variable X will do a small numerical problem, a probability distribution for discrete random variable X is given. So, X is given corresponding probability distribution is given. So, this is an empirical distribution. Suppose, if you want to know, what is the probability of X < = 0. So, what you have to do? Wherever random variable X is 0 and less than equal to 0 you would add it.

For example 0.20 + 0.17 + 0.15 + 0.13 we will get 0.65. Suppose, if you want to know the probability for the random variable - 3 to 1 - < = X < = 1. So, it would add - 3 to 1. 0.15 + 0.17 + 0.20 + 0.15 and you add it will get 0.67.

(Refer Slide Time: 06:12)

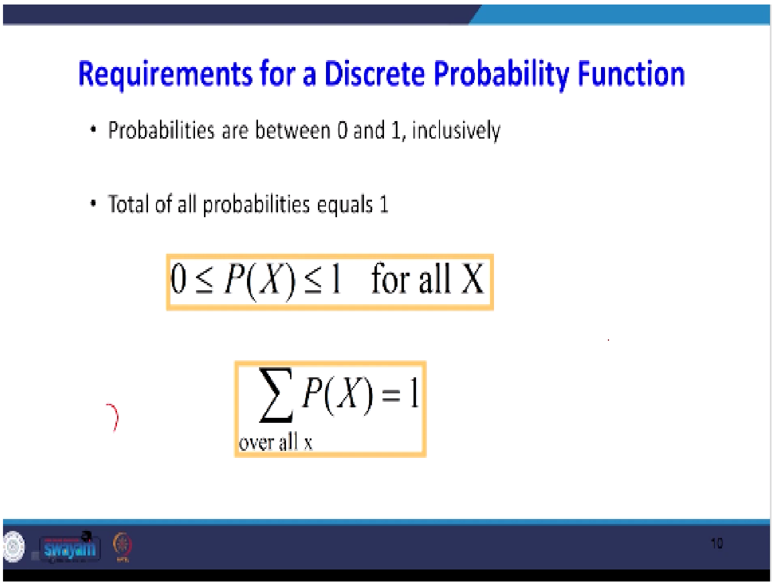


How to plot a discrete distribution? So number of crisis for example is taken the probability of happening that crises is also given for example, the probability of getting 0 crises is 0.37

for one crises 0.31 and so on. So, in X axis you mark the random variable in Y axis you plot the probability. When 0.37 it is a 1 this 1 this is that these are discrete these points, these points cannot be connected in x axis, this random variable has to come into the x axis, this probability has to go to Y axis.

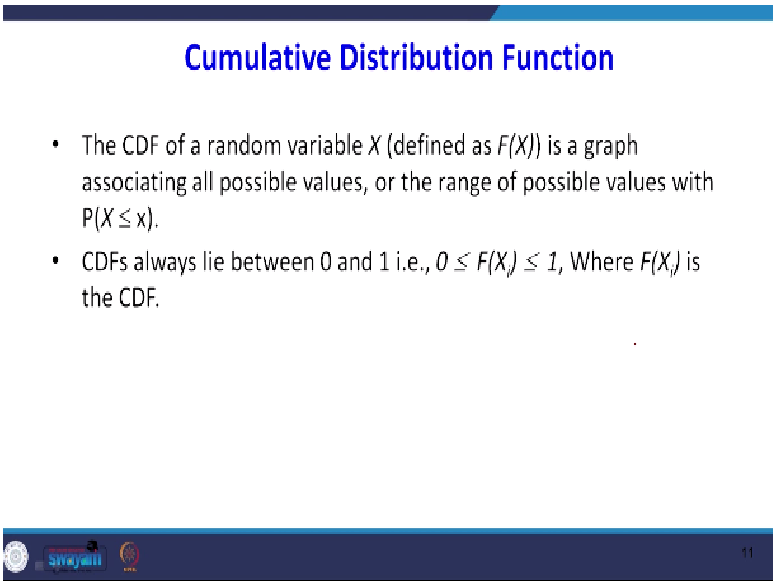
For example, 0.37, this one here what will happen you cannot connect this line because it is a discrete, because you may have an x = 1 when x = 2, when x = 1.5 there is no value, if it is a discrete distribution, you cannot connect these points that is why it is called the discrete distributions.

(Refer Slide Time: 07:17)



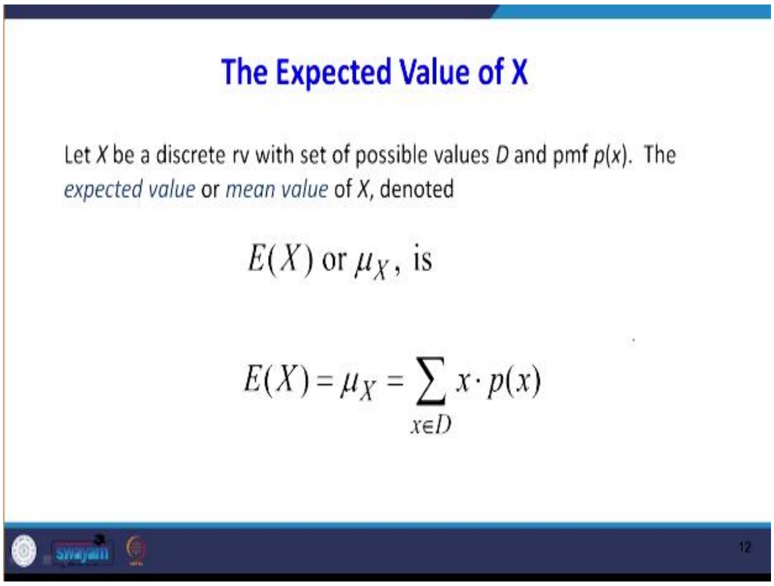
The requirement for the discrete probability density function, so probabilities are between 0 and 1 inclusively. Total of all probabilities equal to 1 and some are probability we have seen that already just 1

(Refer Slide Time: 07:30)



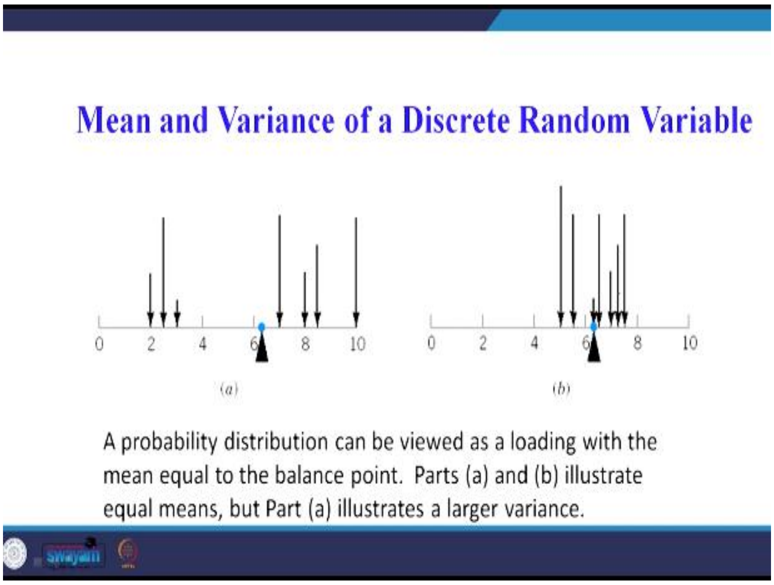
Next term will see cumulative distribution function. The cumulative distribution function of a random variable X defined as F of X is the graph associating all possible values are in the range of possible values with the P of X <= x. Cumulative probability distribution function is just adding the probabilities. The CDF always lies between 0 to 1 that is 0 <= F of x should be <= 1 F CDF cumulative density function.

(Refer Slide Time: 08:04)



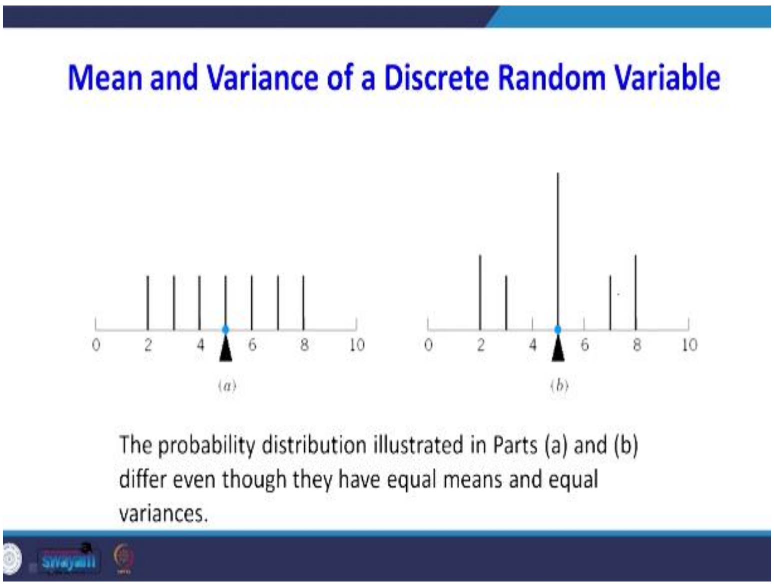
Then there is a very important property is the expected value of X. X be a discrete random variable with the set of possible values of D and pmf is P(x). The expected value or mean value of X is denoted as, generally expect of x or $\mu_x = \Sigma$ x. p(x). So, Σ x. p(x) is your expected value of X.

(Refer Slide Time: 08:32)



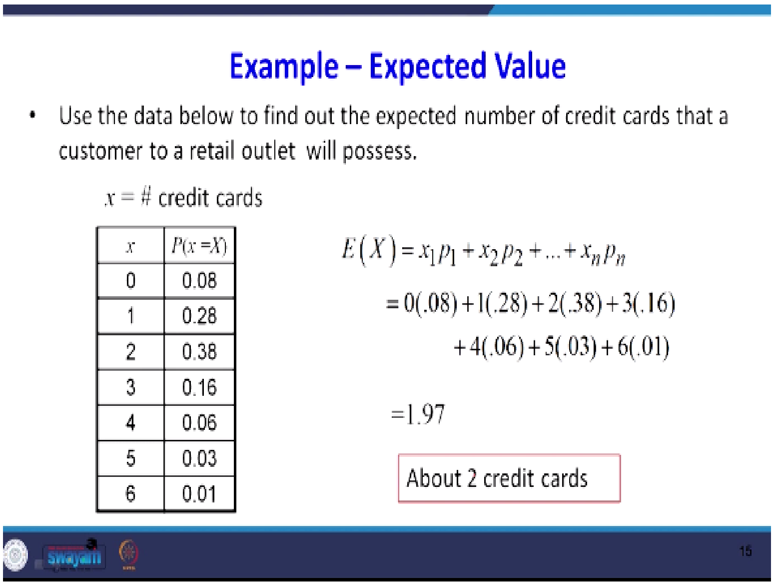
What is the meaning of this mean and variance of your discrete random variable is look at this there are picture (a) and picture (b) the left side, the mean is same for both the distribution, but look at the variance. The left side figure it shows that lot of variances that items figure it is less variances, the probability distribution can be viewed as, are viewed as you are loading with the mean equal to the balance point. So mean is nothing but it is like kind of a balance point for which the distribution lies. Part (a) and part (b) illustrate equal means, but part (a) illustrates a larger variance.

(Refer Slide Time: 09:12)



See the second case mean and variance of discrete random variable, the probability distribution illustrated in parts a and part b differs even though they have equal means and equal variances the shape of the distribution is different.

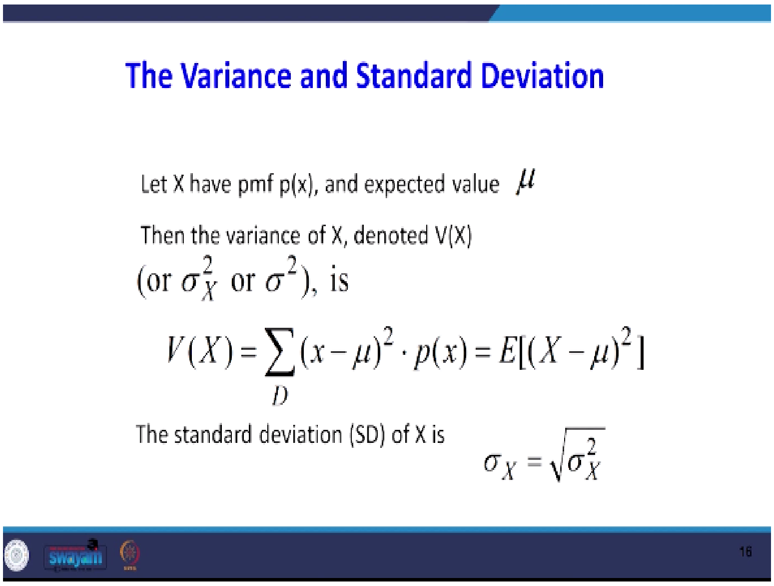
(Refer Slide Time: 09:25)



Now, we will see how to find out an expected value use the data below to find out the expected number of credit cards that a customer to retail outlet will process. So X is a random variable. There is a how many number of credit cards customers having the P (x) equal to X corresponding probability. So Zero P (x) is .08. That means probability of a person having 0 credit card is 8 % probabilities a person to have for example, 6 credit cards is 1 %.

So how to find out the expected value multiplied by x and corresponding probability to submit. So, 0 (0.08) + 1 (0.28) + 2 (2.38), and so on, + 6 (.01) = 1.97. You can make them 2. That means the customers, they can have an average of 2 credit cards that was any customer if you take randomly average that customer can have 2 credit cards. Here an example of meaning of this what is mean.

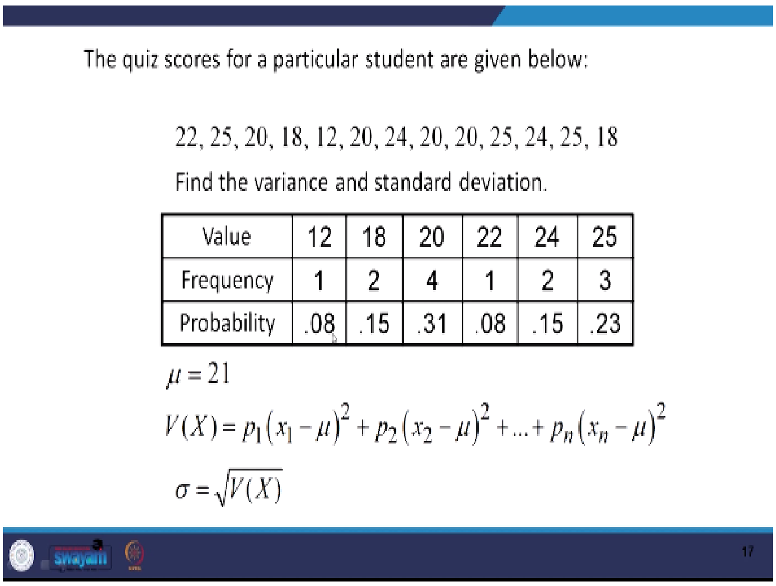
(Refer Slide Time: 10:31)



Now we will see how to find out the variance and standard deviation of an empirical distribution. Previously we have seen $\mu_x = \Sigma$ x. p(x), I will see how to find out the variance of an empirical distribution. Let X how the pmf of p of x and the expected values mu. We know already the mean of an empirical distribution. Now we have to find out the variants of the empirical distribution, then the variance of X denoted as V(X) or $\sigma_x{}^2$ or σ^2.

The variance of X = Σ (x - μ)^2. p(x), variances can be denotes E(X - μ)^2 . The standard deviation is the square root of this.

(Refer Slide Time: 11:14)



We will see you an example, a quiz scores for a particular student are given below 20, 25 and so on find the variance and standard deviation. So, before knowing the standard deviation, first you have to find out the mean because the mean is required. So, the mean if you add and divided by corresponding elements, number of elements you will get 21. For example first we will construct a frequency distribution, you see 12 is repeated by 1 time, 18 is repeated by 2 times, 20 is repeated by 4 times, 25. For example 25 is repeated by 3 times.

Then we have to find the probability, the probability is nothing but the relative frequency as I told you one definition of probability is relative frequency. So what is the cumulative frequency here first you have to find a total frequency 1 + 2 =3, 3 + 4 =7, 8, 10, 13 there is a cumulative frequency. So, the probability here we are obtaining by using the concept of relative frequency. So the relative frequencies we are adding all the frequency that is a total. So 1 divided by corresponding sum of all frequencies 2 divided by some of frequencies. now the mu we can find out mu in another way also, we know that already we are done with this relative frequencies $\Sigma F \times M / \Sigma F$, how to find out the mean, sigma of expected value X. p (x), 12 x 0.08 + 18 x 0.15, + 20 x .31 + 22 x .08,

$\Sigma x. p(x) = 21$. One way you can add all the values you can divided by number of elements.

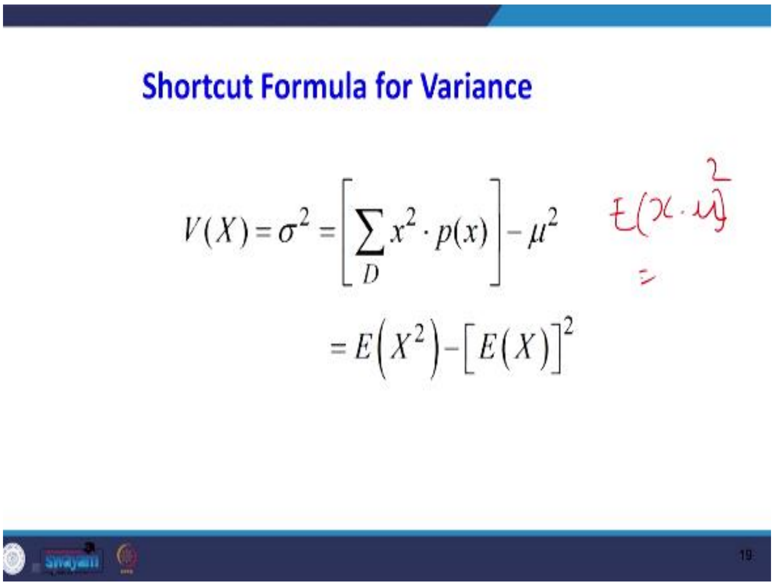
Otherwise from this empirical distribution, what x is given x is 12, 18, 12, probabilities given. So, if you want to know the mean x into p of x, now we are going to find out the variance.

(Refer Slide Tim: 13:40)

$$V(X) = .08(12-21)^2 + .15(18-21)^2 + .31(20-21)^2$$
$$+ .08(22-21)^2 + .15(24-21)^2 + .23(25-21)^2$$
$$\boxed{V(X) = 13.25}$$
$$\sigma = \sqrt{V(X)} \quad \boxed{= \sqrt{13.25} \approx 3.64}$$

So p 1 variance = .08 .X_1, Where $X_1 = (12 - \mu)^2$ +, p 2 , that is , .15 . X_2 That is, $(18 - \mu)^2$ and so on when you add it you will get the variance and the mu take square root of will get this. You see that and going back $.08\ (12 - 21)^2 + .15(18 - 21)^2 + .31\ (20 - 21)^2$, when you simplify the variances 13.25 standard deviation is 3.64. So, what do I have done seen this problem, the data is given first you were constructed here empirical distributions, then we will use the formula of needed variants to find out the mean and variance.

(Refer Slide Time: 14:13)

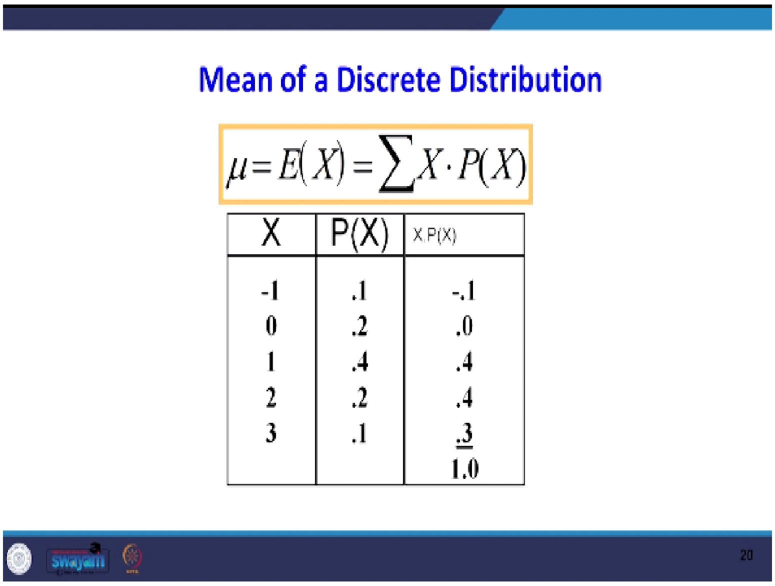


Another shortcut formula to find the variances is nothing but we the E $(x - \mu)^2$ for example already we seen E $(x - \mu)^2$, when you square it and simplify it, you will get this formula. So $V(X) = [\Sigma x^2 p(x)] - \mu^2$

$= E(X^2) - [E(X)]^2$

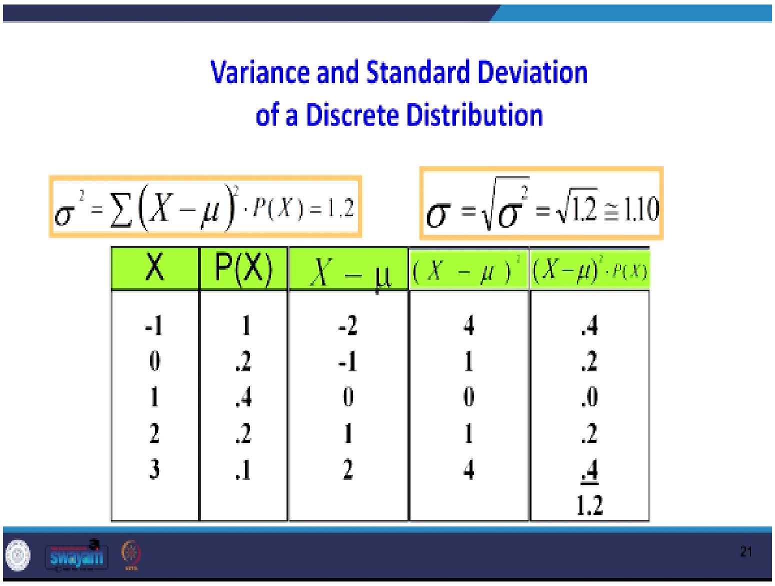
just you expand it will get this answer.

(Refer Slide Time: 14:48)



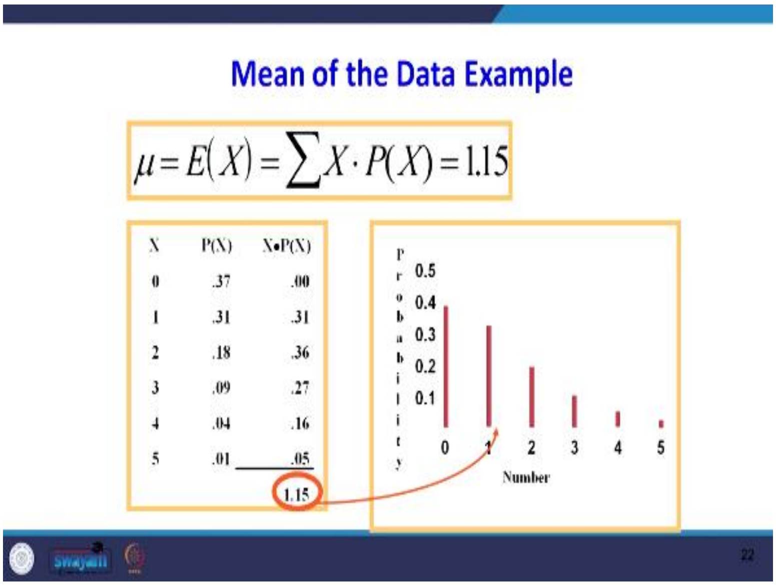
So let us find out the meaning of your discrete distribution, the formula for finding the mean mu equal to expected value of X that is X .p (x). So X is given p(x) is given, multiply X and p(X) after doing that, when you sum the sum is 1. So the mean of this empirical distribution is 1.

(Refer Slide Time: 15:07)



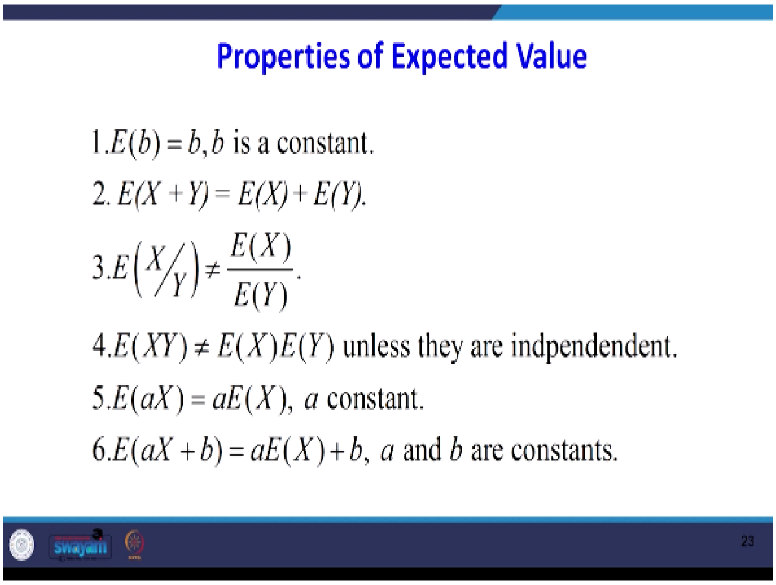
That is find out the variance and standard deviation of this empirical distribution that is a discrete distribution. So σ^2, we know $(X – u)^2$. p (x). So X is given p(x) is given first to find out (X – u) then $(X – u)^2$, then multiply $(X – u)^2$ by P(x) and sum it we are getting 1.2. So the variance is 1.2 you take square root the standard deviation is 1.10.

(Refer Slide Time: 15.37)



Suppose then another distribution say X is given p of x is given X into p of X. So when you plot it the mean you see that the mean or mean not be exactly 1 or 2 or 3 mean value may in between 1 and 2. So mean value need not be discrete only the random variable is discrete here.

(Refer Slide Time: 16:00)



Some of the very important properties of expected values suppose the expected value of a constant is constant only. When you want to multiply 2 random variable E (X + Y). we can write E(X) + E (Y). E (X \ Y) is not division it is a conditional, it is kind of a conditional probability. So E (X \ Y) need not be will not be equal to E (X) divided by E (Y) and same thing E (XY) is not equal to E (X) multiply E (Y) unless they are independent.

If they are independent, you can write E (XY) equals to E (X) and E (Y) otherwise we cannot written. So, if a random variable come along with the constant that constant can be removed out of this expected value. For example E (a X) the a can be brought left side. So, a E (X), here a is the constant. So, easy that it is in format E (a x + b) that can be brought a left side. So, a E(x), then when your expect b value constant this constant itself. It will become a E (x + b) where a and b or constant.

(Refer Slide Time: 17:11)

Properties of Variance

1. Var(constant) = 0
2. If X and Y are two independent random variables, then
   Var(X + Y) = Var(X) + Var (Y) and
   Var(X - Y) = Var(X) + Var (Y)
3. If b is a constant then Var(b+X) = Var(X)
4. If a is a constant then Var(aX) = a^2Var(X)
5. If a and b are constants then Var(aX+b) = a^2Var(X)
6. If X and Y are two independent random variables and a and b are constants then Var(aX+bY) = a^2Var(X) + b^2Var(Y)

Then properties of variances; variances of a constant is 0. If X and Y are 2 independent random variable, then Var(X + Y) = var(X) + var(Y). Var(X – Y) = var(X) + var(Y), it should be very carefully here, support there are 2 groups there are group 1 and group 2. If you want to know the difference in the variance, you too add their variances b a constant, then variances of b + X because variances of b will become 0 it will become only variances of X.

If a is constant than variances of a X is because variances Ax the square term and you bring left side of the bracket would write a square and variances of X. There are proof is therefore this if a and b are constant than variances of a X + b equal to a square variances of X, the variance of B will become 0. The answer is a square variances of X. If X and Y are 2 random variable and a and b are constant, then var(a X + b Y) = a^2 var(X) + b^2 var(Y).

(Refer Slide Time: 18:24)

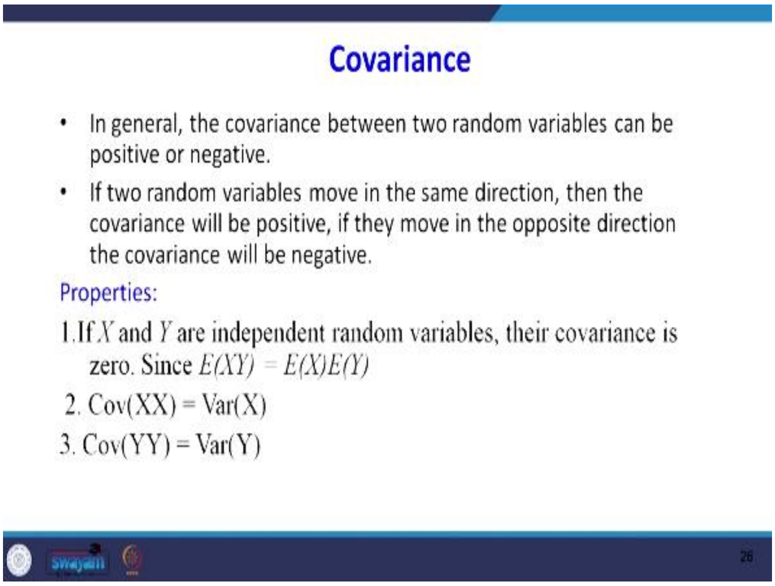
Covariance

Covariance: For two discrete random variables X and Y with $E(X) = \mu_x$ and $E(Y) = \mu_y$, the covariance between X and Y is defined as
Cov(XY) $= \sigma_{xy} = E(X - \mu_x)\, E(Y - \mu_y) = E(XY) - \mu_x \mu_y$

25

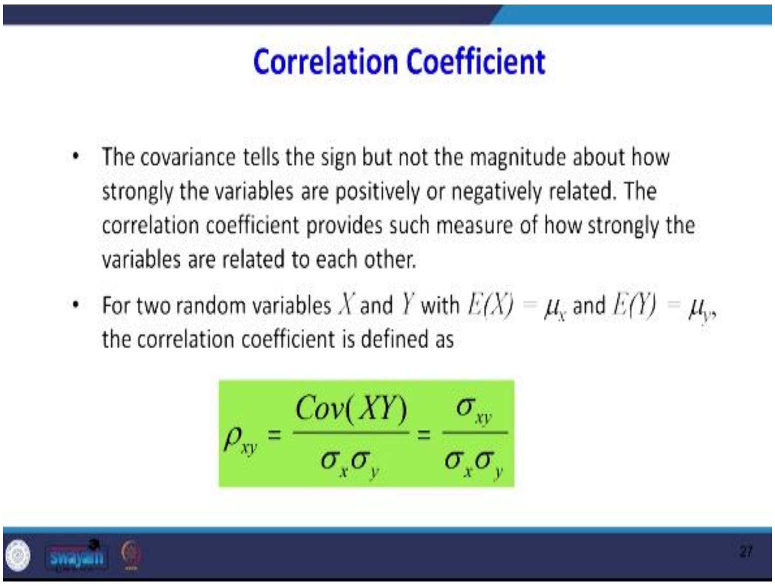
Then covariance for 2 discrete random variable X and Y, E (X) = μ_x and E (Y) = μ_y, then covariance between X and Y is the defined as covariance of X Y equal to can be written as σ x y = E (X - μ_x) and E (Y - μ_y) and simplify it will get E (XY) - μ_x . μ_y , that is a covariance.

(Refer Slide Time: 18:54)



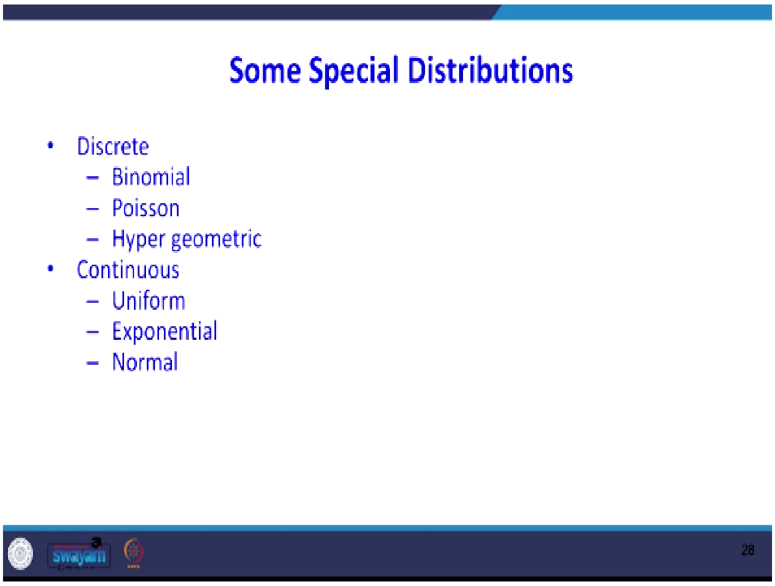
In general, the covariance between 2 random variable can be positive or negative. If random variables move in the same direction, then the covariance will be positive. If they move in the opposite direction, the covariance will be negative. Properties of covariance, if X and Y are independent random variables their covariance is 0. Since E (X Y) = E(X), E(Y) is independent covariance there would not be any covariance. Covariance (X X) is variance of X. Similarly, covariance of YY is simply variance of Y.

(Refer Slide Time: 19:31)



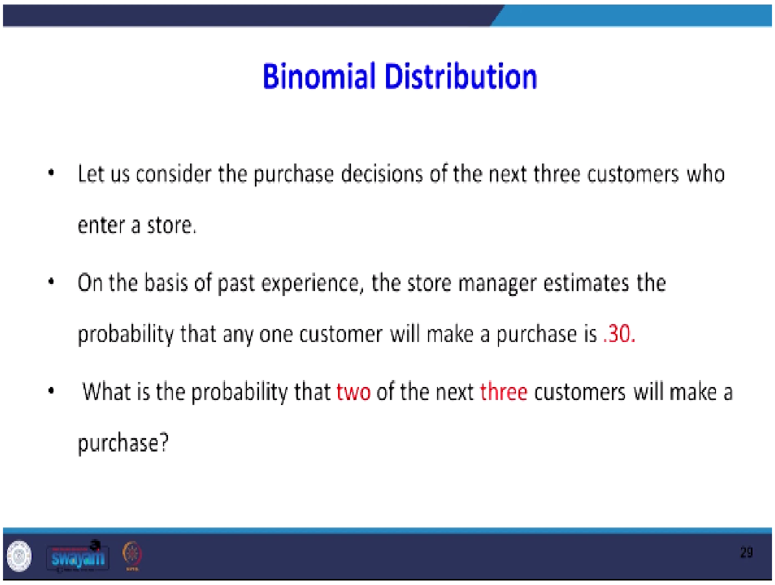
Then correlation coefficient, the covariance tells the sign but not the magnitude about how strongly the variables are positively or negatively related. The correlation coefficient provides such measures of how strongly the variables are related to each other. The variance is only giving the direction not the magnitude, but the correlation it is giving the magnitude for 2 random variables X and Y, E (X) = μ_x and E (Y)= μ_y. The correlation coefficient is defined as covariance of X Y divided by σx and σ y.

(Refer Slide Time: 20:06)



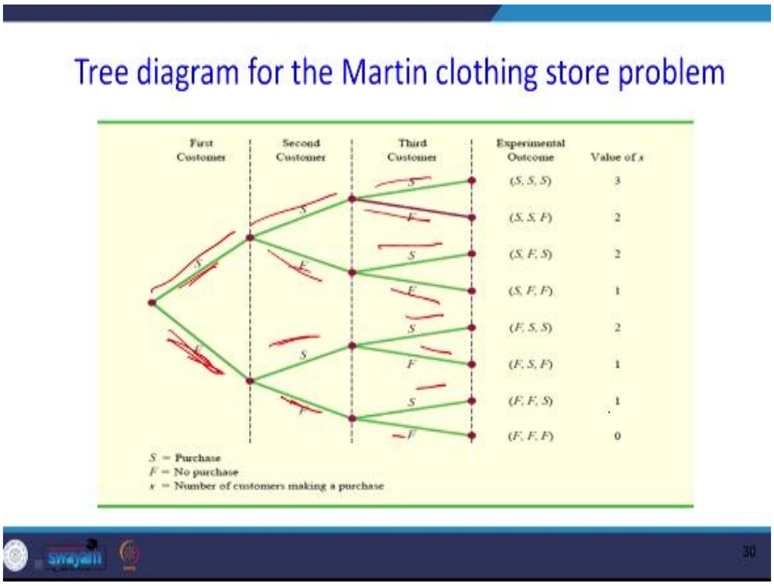
Dear students now are going to some special distributions will study some special distribution in a discrete category and continuous category. The discrete will study about the binomial distribution and Poisson distribution and Hyper geometric distribution. Continuous category which will study we are going to study uniform exponential and normal. In this class I will explain the theory and corresponding its parameters outer end of the class will use Python to find out various parameters various mean and variance of your distributions and corresponding probabilities in the practical class.

(Refer Slide Time: 20:46)



First one is the binomial distribution. Let us consider an example to explain the concept of binomial distribution. Let us consider the purchase decision of the next 3 customers who enter in store there are 3 customers going enter the store and the basis of past experience, the store manager estimates that the probability that any one customer will make your purchase is 0.30. What is the probability that 2 of the next 3 customers will make a purchase?

(Refer Slide Time: 21:18)

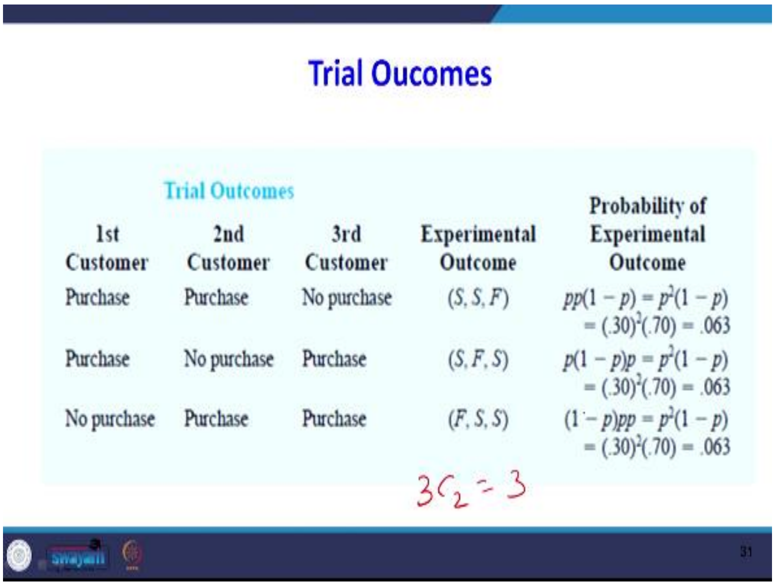


Now look at this the tree diagram the first customer there is a 2 possibility, S is the purchase F is no purchase, X is the number of customers making purchase. So we will see that is the end here. Now, what is happening the first customer he can purchase or not purchased second customer different possibilities, third customer different possibilities. Now we look at the experimental outcome, this possibility, look at this possibility, success success success.

Look at this possibility success success failures look at this personal success failure Success, then success failure failure, failure success success, failure success failure, failure failure success, failure failure failure. So, we have written all possibilities. Now, the question is out of 3 customers what is the probability that 2 customers will make a purchase? What is the meaning SSS all 3 customer have purchased.

So value of x equal to 3 random variable second case to customer how purchased third customer did not buy. So, here x is 2 because here the X is taken the number of customers making purchase the first possibility x = 3, the second possibility is 2, the third possibility 2, the fourth possibility is 1,.. 2, 1, 1, 0. Now the question is, what is the probability that 2 out of 3 customers will make a purchase, See that? There is a possibility.

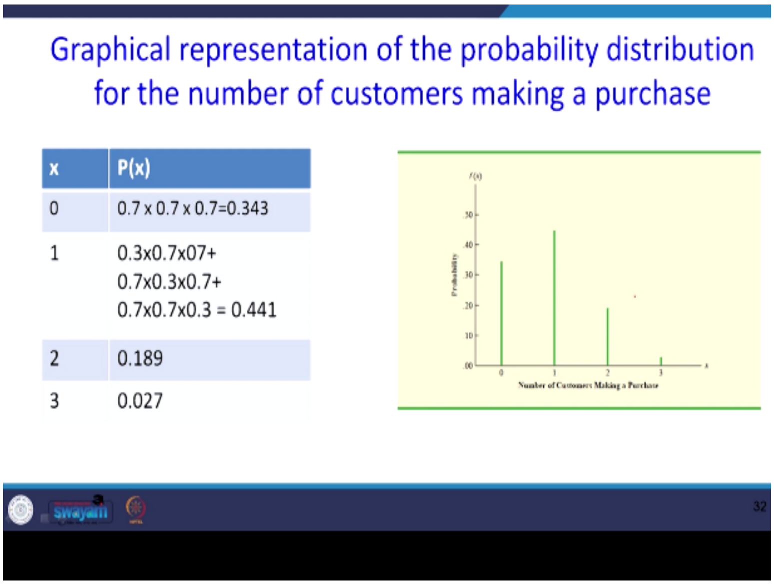
(Refer Slide Time: 23:00)



The first customer: it is the possibility, the SSF, SFS FSS what is the probability of success is p,p and 1 - p we know p is .3. So 0.3^2x0.7 = .063. For second category also we are getting p, first success p failure 1 - p again success is p. So, p^2 multiplied by (1 – p). so.3 square .7 equal to .063 then third possibility failure, 1 - p success success p p ,So, p square 1 – p = . 063, .063, .063.

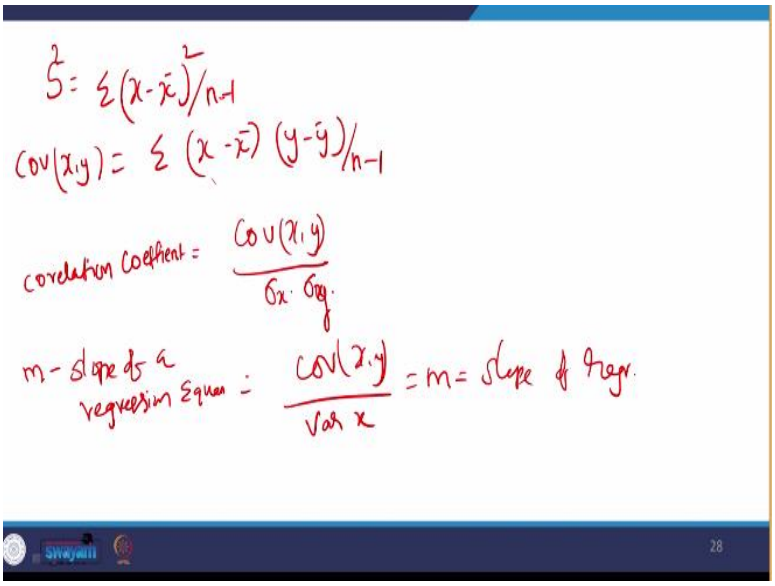
Now, here we actually do here the possibilities $^3C_2$ because the question is asked out of 3 customers, what is the probability that 2 customer will buy? So it is a $^3C_2$, the value of $^3C_2$ is 3. That is why 1, 2, there are 3 possibilities when you go back when you go back how many 3 is there 1 2 3 possibilities there is the meaning of value $^3C_2$.

(Refer Slide Time: 24:20)



Now we will find out the probability if x = 0 that mean nobody is buying.

(Refer Slide Time: 24:33)



Students we have studied so far variance, covariance, correlation coefficient, just how to make it in relation. For example, we know the variance formula, variance equal to Σ(x - x bar) ^{2}. So variances for 1 variable, suppose if you want to know 2 variable if you want to know for 2 variable, this variance will be called this covariance.

So covariance is Σ(x - x bar) .(Y - Y bar) variances divided by n - 1. So, Σ(x - x bar). (Y - y bar), here also n - 1. Variance, covariance, next one correlation coefficient is covariance x, y divided by standard deviation x, standard deviation y. Now, you see that this is a variances, this is a covariance, correlation coefficient. So, for correlation coefficient when you divide covariance to a corresponding standard deviation you will get correlation coefficient.

Next we will say ‘m’ that is called slope of the regression, slope of a regression equation. So, there is nothing but you were covariance(x, y) divided by variances of x, the first one is variance, covariance, correlation coefficient and slope of regression equation, you see that all are having some relationship for the variance is only for one variable, what is the meaning of variance?

How each value is away from its mean that deviation square of the deviation, then some of the deviation, then the mean value of that some of the deviation it will give you the variance for covariance there are 2 variables there how each variable is moving away from its own mean. So, Σ(x - x bar) (y - y bar) divided by n - 1. If you want to know correlation coefficient, that covariance is divided by its corresponding standard deviation look at correlation coefficient.

If you want to know slope and regression equation, if you divide covariance of x, y divided by corresponding variances of x you will get m that is nothing but slope of the regression equation. Dear students so for we have seen what is the need for studying the distribution?, then we have seen how to construct a discrete probability distribution after constructing how to find out the mean and variance of a discrete distribution then we have seen the properties of expected value.

Next we have seen the properties of the variance. Then we have seen how this mean, variances, covariance are interrelated. The next class we will continue some discrete distributions and some continuous distribution in detail. Thank you