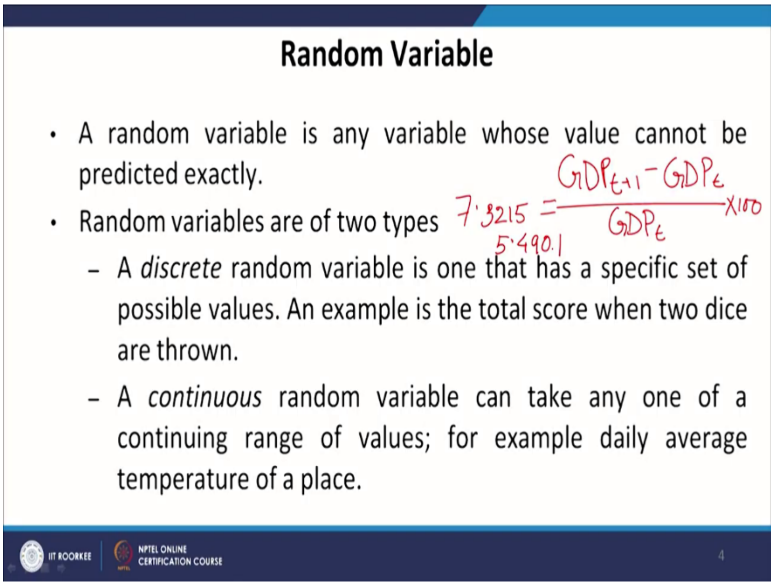
Econometric Modelling
Professor. Sujata Kar
Department of Management Studies
Indian Institute of Technology, Roorkee
Lecture No. 03
Review of Basic Concepts- I

Hello, this is the course on Econometric Modeling. I welcome you to Module 3. The previous two modules have discussed the basics of econometrics, how it is defined, what are its types, the applications and the steps in formulation of economic modeling. Now, the third module basically deals with review of basic concepts. The fourth module also deals with the same thing. So, first we begin with module 3, which takes up review of basic concepts.

So, some statistical concepts primarily, which are very useful in the field of econometrics and will be repeatedly applied or used in the field of econometrics. So, first of all we define random variables.

(Refer Slide Time: 1:21)



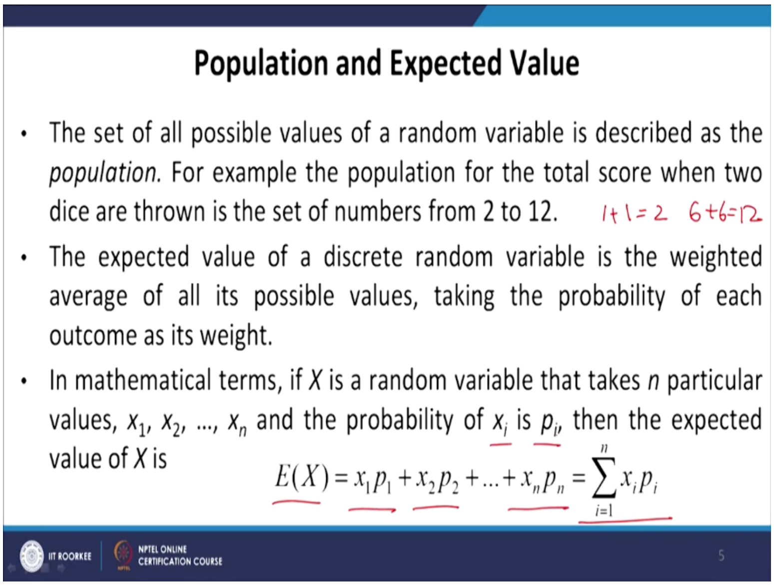
A random variable is any variable whose value cannot be predicted exactly. So, first of all variable implies something, whose values keep on varying. And when we add some randomness to it, then it basically adds certain uncertainties to it. An uncertain event can never be predicted with 100 percent accuracy. Consequently, a random variable can never be predicted exactly.

Random variables are of two types. First of all, discrete random variable (the one that has a specific set of possible values). An example is the total score when two dices are thrown. So, the specific values could be 2 to anything between 12, but the numbers are always whole numbers like 2, 3, 4, 5, 6, 7, and so on. On the other hand, a continuous random variable can take any continuous range of values. For example, the daily average temperature of a place.

It can be also some economic numbers. For example, if we calculate growth values of any series. Suppose, if I consider GDP, the growth in GDP is calculated as GDP at time t plus 1 minus GDP at time t divided by GDP at time t, say multiplied by 100 (*Refer to Slide above*). In that case, this calculation can yield any number.

So, we generally say that the economy has grown at a rate of 7 percent, nearly 7 percent or 6 percent, nearly 6 percent, and so on, but the actual values can be like 7.3215 or it can be 5.4901 and so on. So, since these are calculations, they can take any value, and that is how we do not call them discrete random variables. Then we differentiate between population and expected value.

(Refer Slide Time: 3:55)



The set of all possible values of a random variable is described as the *population*. For example, the population for the total score is when two dice are thrown in the set of numbers from 2 to 12. See, we are throwing two dice simultaneously. So, the lowest value that each one can take is 1, so 1 plus 1 makes it 2. The maximum value each one can take is 6, so 6 plus 6, 12, and it can be anything in between that is one dice has come up with 1 and the other dice has come up with 2 so, it adds up to 3 and so on.

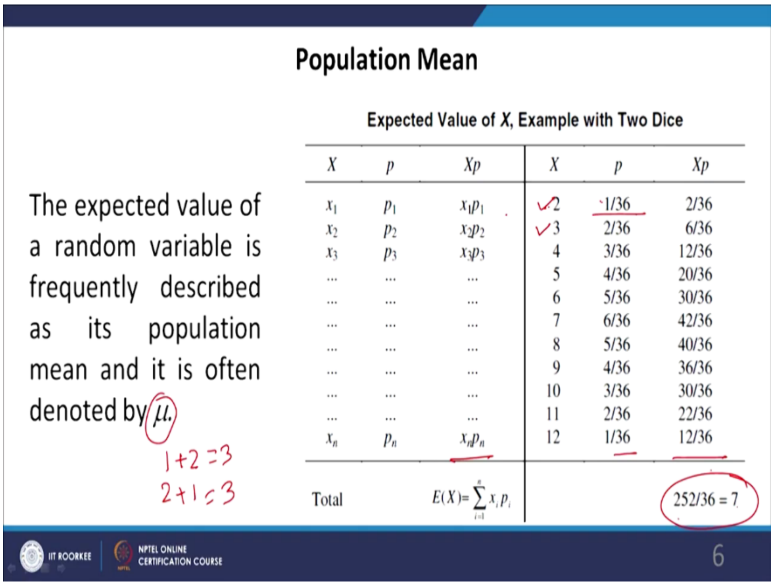
So, the population that is the set of all possible numbers obtained by throwing two dice is between 2 to 12, it contains all possible values, that is population. The expected value of a discrete random variable is the weighted average of all its possible values, taking the probability of each outcome as its weight.

So, I will show you some examples, but before that, we will define it that in mathematical terms, if X is a random variable, then that takes n particular values like x_1, x_2 to x_n and the probability of x_i is p_i, then the expected value of X is calculated following the formula below:

$$E(X) = x_1 p_1 + x_2 p_2 + \ldots + x_n p_n = \sum_{i=1}^{n} x_i p_i$$

So, what is happening here is that the expected value is actually the sum of weighted summation of individual observation. So, when I say and the probability of xi is Pi, then this i is anything from 1 to n. So, xi is occurring with a probability of Pi, which implies that x_1 is occurring with a probability of P_1, x_2 is occurring with a probability of $P_{2,}$ and so on.

(Refer Slide Time: 6:20)



Then, we discuss the concept of the population mean. The expected value of a random variable is frequently described as its population means and it is often denoted by mu. So, the expected value is actually equivalent to the mean of the population observations and it is denoted by μ. So, here is the example.

We continue with the example of the two dice and the possible outcomes *(refer to slide on the population mean above*). So, what are the possible outcomes? As I have explained earlier that when we simultaneously through two dies then the minimum value their sum can take, is 2.

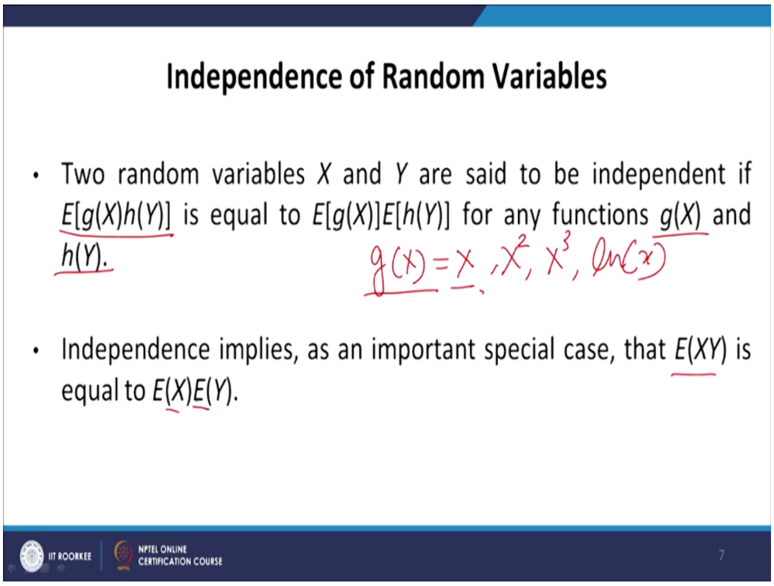
So there can be, the sum can be 3. Now you can see that the possibility of 2 coming up is 1 out of all 36 possibilities.

Similarly, 3 can come up only twice, so that two instances are, first of all, first dice take value 1 and the second dice take value 2. Alternatively, the first dice can take value 2 and the second dice takes the value 1. So, there are two instances when it can be 3. And similarly, we can find out that there are four instances, three instances where the sum can be 4. There are four instances where the sum can be 5 and so on.

The probability of an event occurring is the possibility, the number of times it can occur divided by the total number of times different values of the population takes up. So, 36 is the total number and the possibility that the sum will be 2 is only one, so 1 upon 36 gives us the probability of the sum being 2, and so on.

Finally, in the last column, what we are doing, which is equivalent to x1, p1 we are multiplying x with p. So, 2 multiplied by 1 by 36, and so on. And finally, this gives us our expected value of x, which is the summation xipi *(refer to slide on the population mean above*). So, we sum all this together and arrive at a value of 252 by 36, which is equivalent to 7. So, this tells us that the population mean or the expected value of X is 7 for this particular example.

(Refer Slide Time: 9:09)

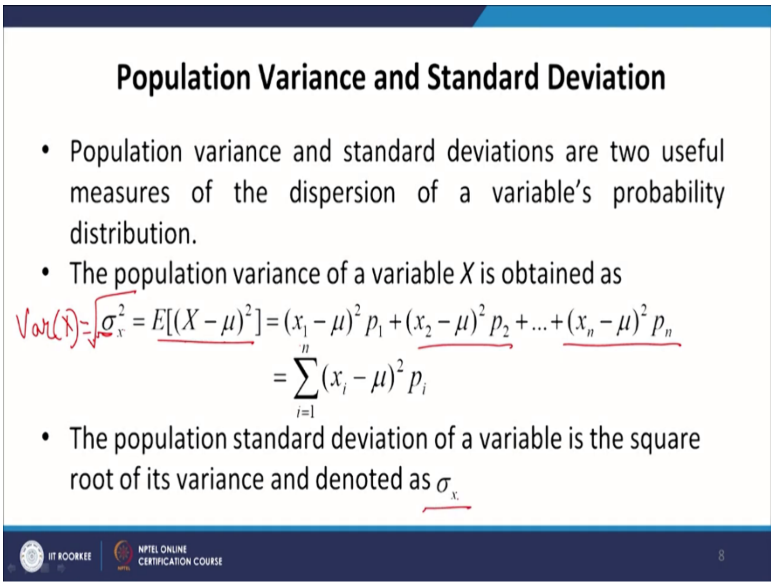


Now we talk about the independence of random variables. Two random variables X and Y are said to be independent. If the expected value of g(X) multiplied by h(Y) is equal to the expected value of g(X) multiplied by the expected value of h(Y), for any functions g(X) and h(Y). So, g(X) and h(Y) here refers to any general functional form of two random variables X and Y. So, g(X) basically refers to some functional form of X, it can be simply the values of X, it can be X's square, it can be X's cube, or it can be the logarithm of X and so on (*see the slide above*). Similarly, it will be for h(Y).

For any functional form of Y, we refer to it as h(Y), a general functional form. So, what independence implies is that the expected value of the multiplication of the two functional forms of the two variables is equal to the multiplication of expected values of the two functional forms. Independence implies, as an important spatial case that the expected value of XY is the expected value of X into the expected value of Y.

So, basically, from this general generic functional form, we are coming to a specific functional form, which is only X. So here g(X) is equal to X and that is how the expected value of XY is equal to the expected value of X multiplied by the expected value of Y. Now, this is actually a very important thing to remember. Why this independence of random variables implies it so is actually outside the purview of the course. Here, I am just introducing you to these concepts, which we are going to use in the subsequent modules again, and again.

(Refer Slide Time: 11:15)



Coming to Population Variance and Standard Deviation. Population variance and standard deviations are two useful measures of the dispersion of a variable's probability distribution. So, a probability distribution is often a reflection of what is the distribution of the data generated with a certain level of probability.

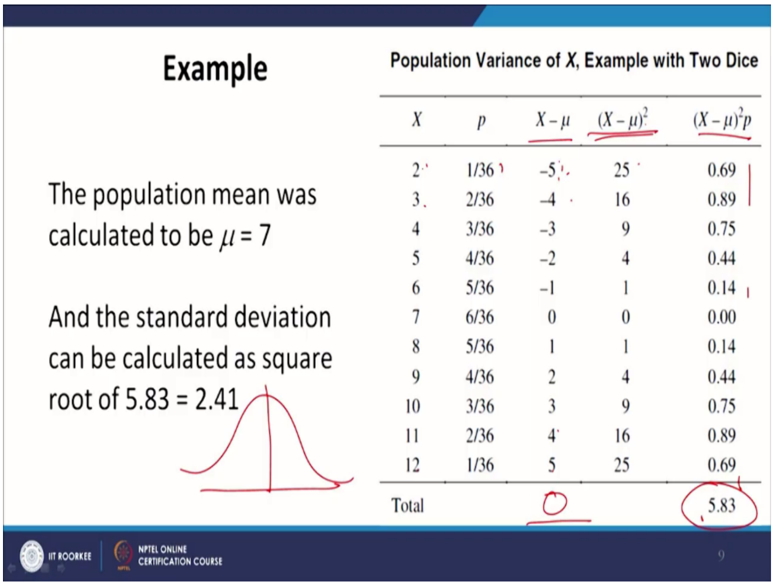
Now, how the data generated or how the distribution is dispersed is actually reflected in terms of population variance and standard deviation. The population variance of a variable X is obtained as:

$$\sigma_x^2 = E[(X-\mu)^2] = (x_1-\mu)^2 p_1 + (x_2-\mu)^2 p_2 + \ldots + (x_n-\mu)^2 p_n$$

$$= \sum_{i=1}^{n} (x_i - \mu)^2 p_i$$

Just a quick recap that μ is the population mean of the variable X here. So, this is how population variances are calculated, and the population standard deviation of a variable is the square root of its variance and is denoted as sigma x. So, sigma square x root over gives us the population standard deviation, which is sigma x.

(Refer Slide Time: 13:39)



Now we take up another example. If you recollect, the population mean was calculated to be 7, so μ is equal to 7. And the standard deviation is calculated from here, which is 2.41, where the variance is actually 5.83 *(refer to slide above, Slide time: 13:39)*. Now you can see that continuing with that example of the sum of two dices thrown together, then what are the outcomes, we have values of X, taking values of X from 2 to 12 and their associated probabilities are also mentioned here.

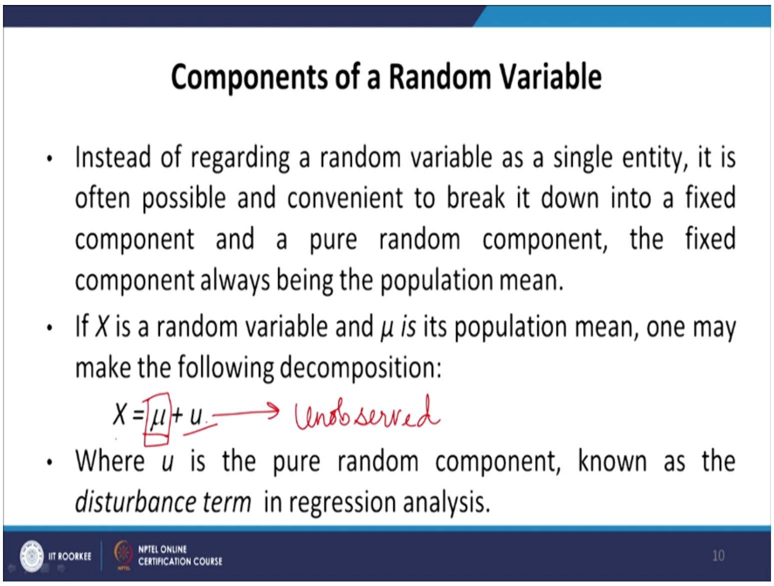
Now *refer to slide above: Slide time: 13:39,* see, X minus μ column, so 2 minus 7 is minus 5, 3 minus 7 is minus 4, and so on. Then we have X minus μ whole square, so I am simply taking the square of this column, so minus 5 squared is 25, minus 4 square is 16 and so on, then, this column is multiplied by the probability values.

So, this is X minus μ square multiplied by p. Summing them up, I arrive at 5.83, which is the population variance, and the square root of 5.83 is 2.41, which is the population standard deviation. So, this is how we calculate population variance and population standard deviation. As you can understand that this ideally tries to tell us that how much individual values have deviated on average from the mean of the series. So, this is the deviation.

Now, if I add the deviation simply, you would see that we would arrive at a 0. So, most often the deviations of individual values from the mean when summed up, would come to 0. Specifically, if we have a nice-looking distribution, that is, everything is centered around the mean, then the deviations would all add up to 0 because the negative and the positive would cancel out.

So, in order to avoid this problem, we first take the square root, then multiply it with the probabilities and sum it up. And when we take the square root, then we actually sum how to take care of this square root thing. So, as a result of which standard deviation could tell us roughly what is on average the deviation of individual values from its mean.

(Refer Slide Time: 16:33)



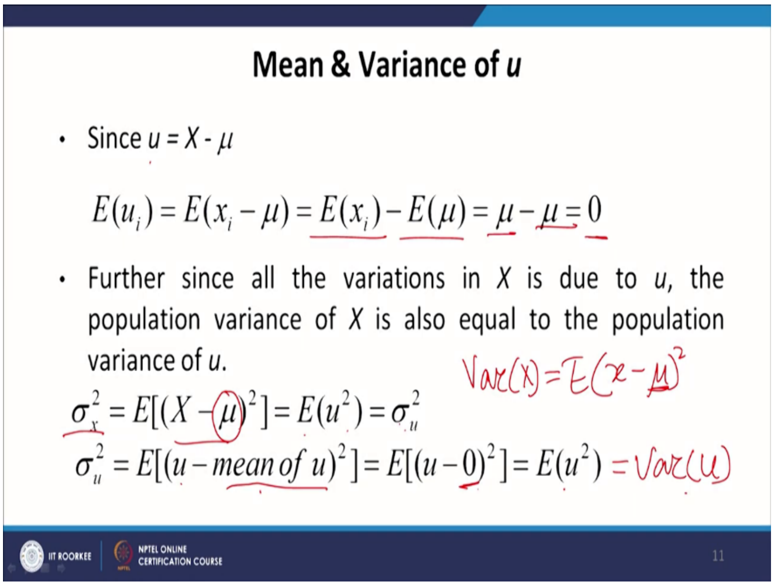
Now, we talk about the components of a random variable. Instead of regarding a random variable as a single entity, it is often possible and convenient to break it down into fixed components and a pure random component, the fixed component always being the population mean. So, basically, we are here trying to bring in some kind of uncertainty because, as I told you in the beginning that random variables have certain randomness in them and randomness is difficult to predict.

So, if it is completely random, which implies there is too much uncertainty, then it is very difficult to predict in order to come up with some level of certainty and predictive ability, we can basically introduce this certainty by dividing a random variable into two components, the first component is a fixed one and the second one is completely random. So, the fixed component most often is actually the mean of the series.

So, what we are doing here is that, if X is actually the random variable then μ being the mean of this random variable, is a sort of non-random component, which we also called a deterministic component and u is a purely random component. So, if X is a random variable and μ is its population mean, one may make the following decomposition where X is equal to μ plus u, where u is a pure random component known as the disturbance term in regression analysis.

u actually stands here for the un-observed component. This unobserved component is actually the component, which has certain uncertainties associated with it and is difficult to predict.

(Refer Slide Time: 18:46)



Now, we talk about the mean and variance of few. Basically, the regression analysis put a lot of importance to this concept of u, because, the thing is that, if this u component is not here, which implies that there is no randomness to it, and if there is no randomness to it, then we do not need actually statistics and econometrics. Everything would be deterministic, everything can be predicted with 100 percent certainty and that can be done using mathematics, simple or complex whatever.

So, u actually plays a very important role when you will get into the regression analysis you will see that all the assumptions actually revolved around u, that is, how this unobserved component should behave, what kind of behavior is desirable. And on that basis, the quality of our regression results actually can be determined or on that, the quality of the regression analysis depends to a large extent.

So, u plays a very important role, μ has actually nothing to do because it is a fixed component. So, we talk about different features and characteristics of u, we begin with the mean and variance of u. So,

$$E(u_i) = E(x_i - \mu) = E(x_i) - E(\mu) = \mu - \mu = 0$$

Since u equals X minus μ, the expected value of ui, that is, here we are considering the population mean of the unobserved component. So, what we are doing is that ui is replaced with xi minus μ, and μ is actually a non-random component. So, when we further break up this expected expression we have the expected value of xi minus the expected value of μ.

Now since μ is non-random, it does not have an expected value. It is a fixed number, it is a constant term like 2, 3, 4. Any discrete number or any number like 2.53, 2.64, whatever be it. So, a number does not have any variations associated with it, any probability associated with it. So, expected values or mean values are actually the same. The mean of 2 is 2, the mean of 4 is 2, maybe it does not make any sense altogether, so this actually remains μ. While we know that the expected value of xi is μ, so μ minus μ becomes 0.

So, the expected value of the unobserved component is 0. Now, here in this context, I take you to the previous example that we had taken, where you can see that we also argued that this is x minus μ. And the expected value of this x minus μ, actually appears to be 0. If we calculate the expected values, you would find that it turns out to be 0. So, the thing is that the unobserved component or the error term most often tends to have 0 mean.

Then we talk about it further since all the variation since x is due to u, the population variance of x is also equal to the population variance of u.

$$\sigma_x^2 = E[(X-\mu)^2] = E(u^2) = \sigma_u^2$$

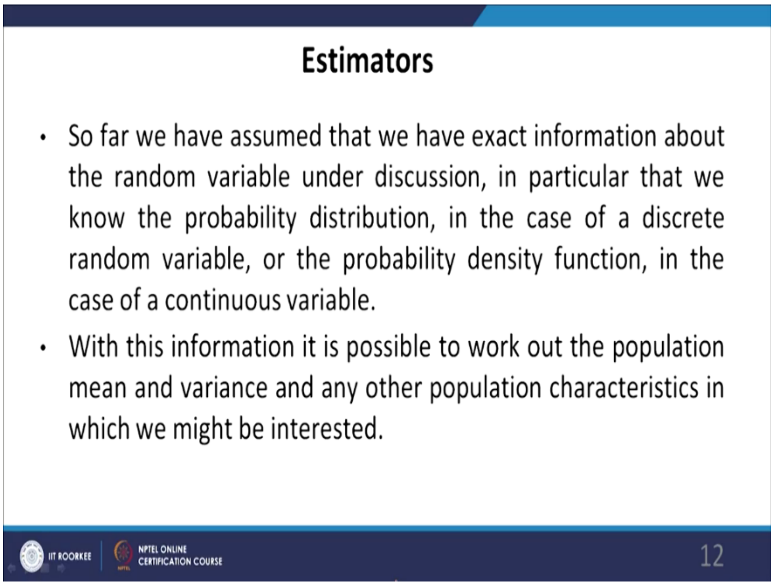
What I am trying to tell here is that the population variance of x is actually the expected value of x minus μ square, μ does not have any variation in it, so the expected value of μ or rather, here the variations in μ is actually 0. So, what I am left with is, this is replaced with u, so the expected value of u square and is equal to sigma u square.

Now how I am coming at it, so the expected value of sigma u square is actually the expected value of u minus the mean of u. Now we have already seen the mean of u is actually 0. So, the mean of u is 0, so here I am left with the expected value of u square. An expected value of u square is actually variance of u that is what I had explained earlier only.

$$\sigma_u^2 = E[(u - mean\ of\ u)^2] = E[(u-0)^2] = E(u^2)$$

So, this is a common practice that if I write variance of x, then this is the expected value of x minus mean of x whole square. So, if the mean of x is 0, then this becomes the expected value of x square. So, the same thing is happening here. So, this is the variance of u and that is how this is the variance of u.

(Refer Slide Time: 23:49)



Now, we talk about estimators. What is the concept of an estimator? How are they different from estimates?

So far, we have assumed that we know the probability distribution in the case of a discrete random variable or the probability density function in the case of a continuous random variable.

With this information, it is possible to work out the population mean and variance and any other population characteristics in which we might be interested. So, we are so far dealing with population, but the problem is that most often population is not available to us or population is not observable to us.

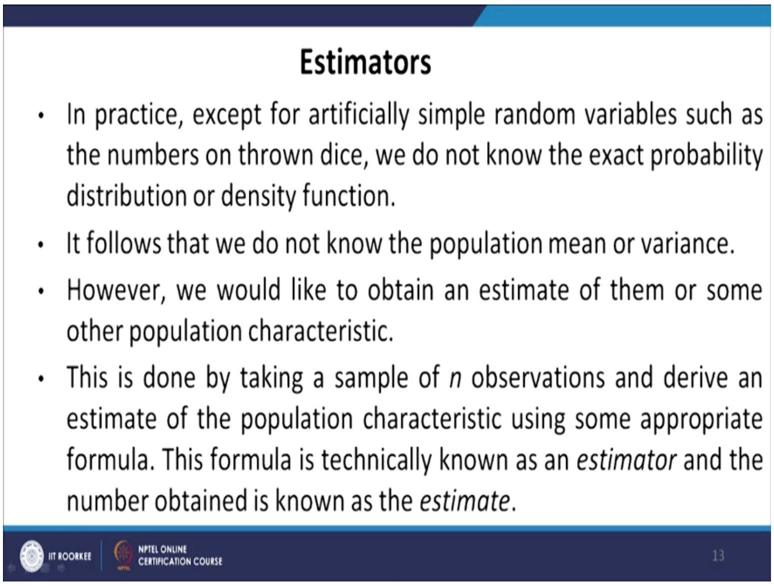
For example, if I want to find out the impact of parental education on a child's education. For that, I may need to collect data from some number of individuals. So, if I want to consider the population. Suppose my study focuses on India, this means, I must consider all the parents residing in India or those who are residents of India and their children. Suppose, I also specify the children's education.

So, I begin with children's education. Suppose I consider an individual and I tried to find out whether that individual has completed 10 standard or not. So, the education criteria or the education levels could be measured as whether someone has completed 10 standard or not, then whether someone has completed 12 standard or not, then further graduation, post-graduation, professional education and so on.

So, I can have several criteria for a child's education, and for parenting education, I will then note down that the child has so education and his or her parents have this education. I am just simply talking about the problem with the population. It is difficult for any individual to collect data from all the parents and all the children who are residents of India or who are residing in India.

So, even if my population is confined to India, this actually becomes a Herculean task. So, what we try to do is that we take up a sample. So, this is an example, where we are actually not able to work with the population, population is not available to us. The example of dice was pretty simple, where we could work with the population, but most often it is not possible to work with the population. So, what do we do in that case?

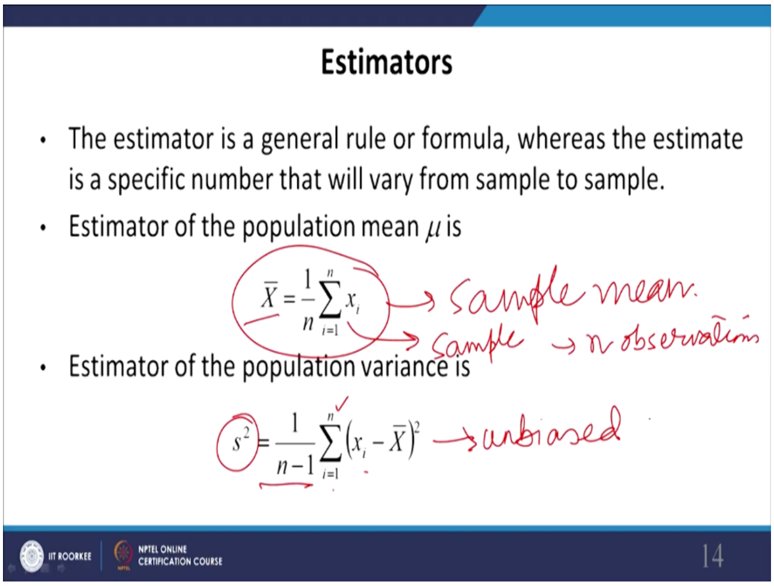
(Refer Slide Time: 26:55)



In practice, except for artificially simple random variables such as the number of thrown dice, we do not know the exact probability distribution or density function, because we do not have the population observed most often. So, it follows that we do not know the population mean or variance. However, we would like to obtain an estimate of them or some other population characteristic, this is done by taking a sample n, sample of n observations, and derive an estimate of the population characteristic using some appropriate formula.

This formula is technically known as an estimator and the number obtained is known as the estimate. So, estimators are the formula, which is giving us sample estimates of the population parameters. The population is not available, I have taken a sample. I am estimating the population parameters using the sample. So, *the formula used to estimate the population parameters is known as estimators, and what we obtain from the sample as the estimates of the population parameters are known as estimate.*

(Refer Slide Time: 28:14)



The estimator is a general rule or formula, whereas, an estimate is a specific number that will vary from sample to sample. The estimator of the population mean is:

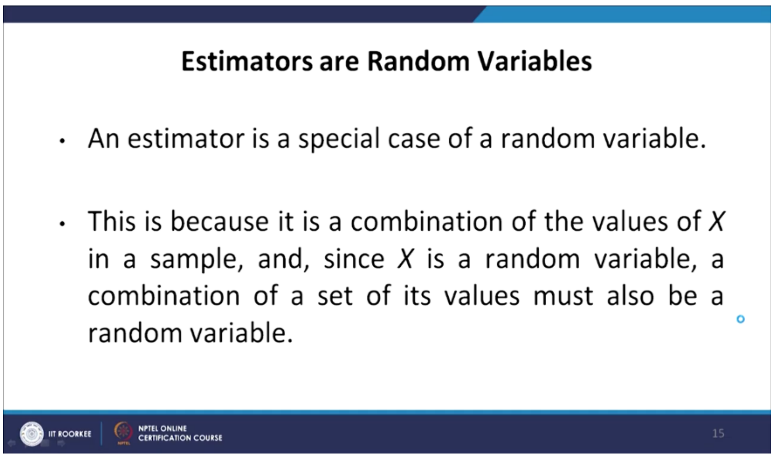
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The estimator of the population variance in a similar fashion is denoted by:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2$$

For the time being, we just note down this formula, which is an unbiased estimator of the population variance.

(Refer Slide Time: 29:51)



Estimators are random variables. It is a special case of random variables because it is a combination of the values of X in a sample. And since X is a random variable, a combination of a set of its values must also be a random variable. What it tries to tell us is that I have taken a sample of say n individual, so say 100 individuals, I continue with my previous proposition or problem that I am interested in finding out the impact of parental education on child education.

So, first of all, I have picked up 100 children and their parents, so 100 mothers 100 fathers. So I have roughly 200 parents and 100 children. Now, if I choose a different sample or someone picks up a different sample, that sample could be another 100 children and another set of parents, so this makes basically estimators a random variable because I can have different estimates of the population mean and various using different samples.

So, that is why it is said that an estimator is actually a combination of the values of X in a sample. And since X is a random variable, a combination of a set of its values must also be our random variable.

(Refer Slide Time: 31:24)

So, this is what is the first part of the introduction or a review of the basic concepts. These are the books that have followed in order to come up with or present these ideas. I will next continue with the review of basic concepts in the next module as well. Thank you.