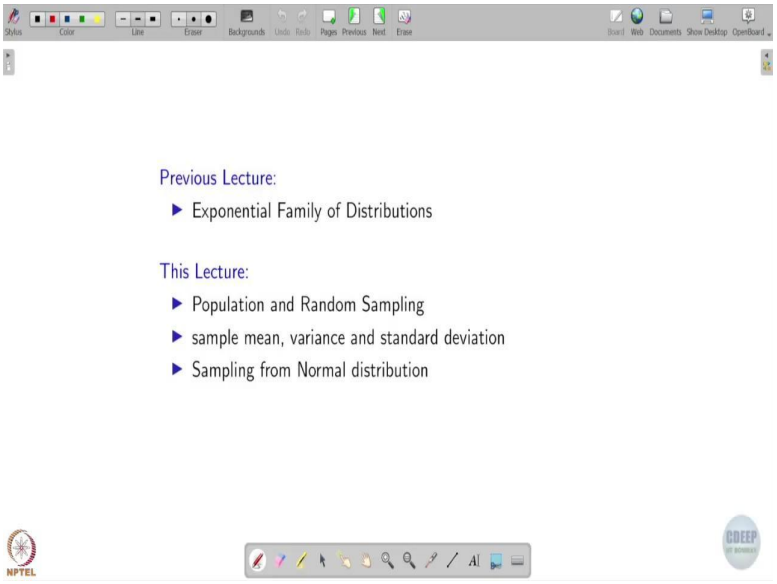
Engineering Statistics
Professor Manjesh Hanawal
Industrial Engineering and Operational Research
Indian Institute of Technology, Bombay
Lecture 21
Random Sampling, Sample means and Sample variance

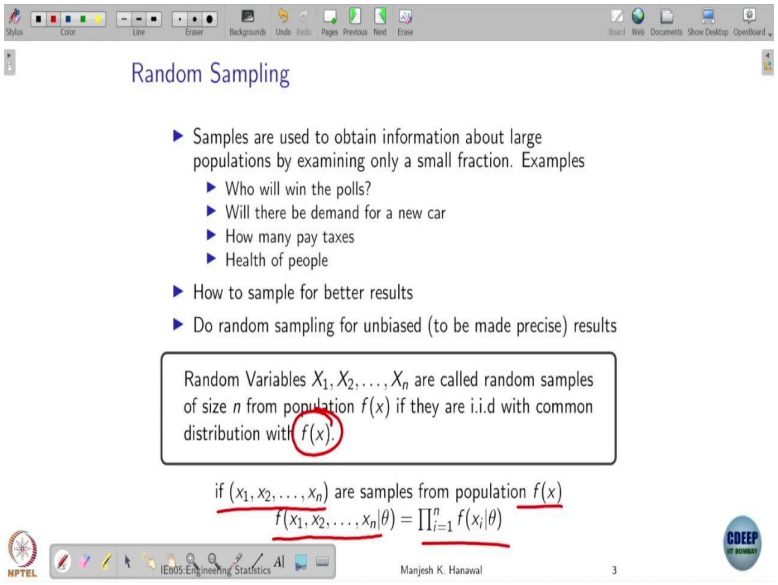
(Refer Slide Time: 0:15)



Now that we have most of the distributions we know, we have written them compactly and put them as an exponential family which has a very compact CPMF or PDF representation. Now, let us start looking into the data itself. How to infer the parameters that we are associated in these distributions? So, for that, if we have to get these underlying parameters, data is fine, but how this, that data should satisfy certain properties or the way the data is generated has to be done in some particular way that is what we are going to know studying sampling.

So, when I say sampling it is about the data points if you let us say there is a black box which is generating data, you do not know what is the underlying distribution according to which it is generating data, you just ask it to give one data point it gave that like a sampling for you. You got one sample like that you can ask more samples and that whole process is called sampling.

(Refer Slide Time: 1:42)



Now little bit motivation on the sampling. Sampling, you can just roughly think as data collection mechanism. Now, for time being forget about this underlying distribution and all. Data is collected in various form for various purposes. And nowadays, especially when there is a large population and we want to understand something about this population, but you cannot go and talk everybody, you may only talk to a small fraction and you could just imagine like you go and talking to them and collecting information that itself can be thought as sampling process.

Now, let us see. One way to look into all this is like I am going to let us say X is a random variable, which denotes the score obtained by a particular student in this IE 605, X is that random variable. And all of your realizations of that random variable. I can see what is the score obtained by you, what is the score obtained by you, like that I can think of realization of that random variable. If I got like about 60 students are in this class. If I got the score of the 60 students that I can think of 60 samples of this random variable X.

So, similarly let us think we as a country, we have so many people. Let us look into one particular issue like health status of the Indians, what is the health status of Indians? That is like one value let us say some health has some number which is going to be denoted by X. Now all of us are realization of that. All of us kind of like based on like our health value, like let us say we assuming that like, we have some inherent properties like we as Indians, which is going to go on some medical condition, maybe like all of us are manifestation of that

medical condition, which shows up in that like maybe we can go at each individual ones and ask that the extending this argument suppose that say, this is a classic examples of polls.

We get all the states polls, national level polls, everywhere. And depending on something, maybe let us say all Indians have some tendencies to make, I mean, we have our own way of making opinions about what to do, which party to consider and all. And during these polls, when people do all these opinion polls we have seen like they cannot go and ask everybody that is a massive process that election commission will do. They will only go and talk to certain people that is like a sampling process instead of going and asking everybody they are only going to possibly ask a certain fraction of the people feeling that they are there are like a representative sample of the whole population.

Now, here we can see that, X is a we can think of polling, X is like a random variable, which is going to denote what which party are going to vote. Then I cannot then all of us are like a realizations of that. Now, you can just think of these arguments like suppose there is a new car in the market and all of us have something about like to buy it or not. Now, what, the company who is launching that car want to know that, whether that brand new car will be how much it will be accepted in the market. So, they cannot go and ask everybody. They will only go and ask maybe certain fraction of the population, only collect certain samples.

So, the point here is like, we can just imagine that all of us are like samples when we are looking into the big population and let population is like a realization. What we want to do is we cannot go and ask everything, we will be only sampling certain fraction. Now the question here now comes is how to do better sampling? So, let us take this voting issue itself. When the election commission conducts voting, everybody is going and waiting, like, I assume that mostly all the populations are voting. But now we are doing the opinion poll, you do not have that much of resources going and asking everybody that is a massive work, you are only going to ask certain fraction.

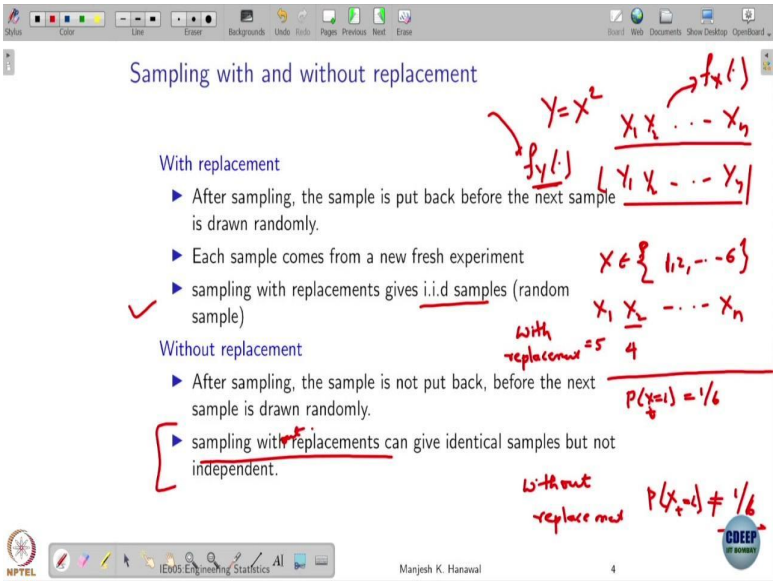
Now how to choose that fraction itself? That is a question. So, that whatever that fraction tells that is actually going to represent actual outcome. Now, this motivation, the question is how to do better sampling. And now, on that front, we are going to refer to one important term called random sampling. So, that it results in an unbiased result. So, what is unbiased? We will make it precise, what is the mathematical meaning of unbiased. I mean, we use the word biased, unbiased very frequently. When something is not favor, we will say, that guy is biased

like, he is not thinking the way I want, but we will make it precise what we mean by unbiased.

Now, we are going to say that the random variables X1, X2 up to Xn, these are called random samples, of a population f of x. Now you see that I am using the word population for f of x, f of x is actually a PDF, but now I am calling it as a population if they are i.i.d with common distribution f of x. So, simply random samples of size n from population f of x, that means there are i.i.d samples drawn from distribution f of x. And here distribution f of x when I say it could be probability mass function or probability density function, depending on you are dealing with a discrete case or continuous case. Is that fine?

And now if you have one particular sample, let us say you have obtained n samples, x1, x2, xn and if they are indeed random sample coming from population f of x, then what we know, the joint distribution of them should be written as the product of their marginal PDF. So, this is the definition we have set already. This is just I am using independence here. And identically distribution is coming because they are all following the same distribution. So, the same distribution, the common distribution, sometimes I am also now henceforth calling it as a population.

(Refer Slide Time: 10:48)



Now when I am doing the sampling, I had to be using words with and without replacements. So, what is with and without replacement? In the with replacement, after you do a sampling, the sample outcome, whatever that you are seeing, this will be put back before the next sample is drawn randomly. And each sample comes from a new fresh experiment, like, if you

put it back, all the possibilities are now there, when you want to do the experiment again, all the possible outcomes are there. So, that is why when I put it back, it looks like it is coming from a fresh experiment.

And when I do sampling with replacement, they will give i.i.d samples. On the contrast, if you are going to do without replacement, so, what I will do is after sampling, the sample is not put back. So, you are sampling outcomes get restricted when you do experiment again.

Student: Sir, if y equals to x squares?

Professor: y equals to?

Student: x square than that is also called sample of the population (()) (12:07) sample of the population?

Professor: So, let us say you have this sample X1, X2 up to Xn and from this you got Y1, Y2, Y n using this relation this is your new sample.

Student: Of the same population?

Professor: Yeah because you have this let us say this is coming from some underlying distribution. Now, one you do, y equals to f y this will induce another distribution and y, and all these new samples, they are also coming from this distribution only.

Student: But here the distribution changes from x to x square.

Professor: Right. But yes, now you have to say that these samples are coming from this population, so the population is always, what is the PDF of for PMF you are looking at. Now, when you are going to do sampling with replacement, it can give identical samples but need not be independent, so let us quickly look into some example. Let us say X is some random variable, which is taking 6 value, sorry, from that is this experiment is throw of a die, which is going to give you 6 outcomes. Let us say I am going to repeat this experiment 10 times. When I did first time X1, let us say I got a value 5. What is the probability of getting 5?

Student: 1 by 6.

Professor: 1 by 6 because all the things are there. Now 5 I got. I took that 5. So, that dice, that the same dice I again used to get the second value and let us say I got a value 4, what is this probability?

Student: 1 by 6.

Professor: This is also 1 by 6. So, I am going to use the same dice and again and again. Now let us now say that x equals to 5 I got. Now I removed that face 5. Now my dice is consisting only of 5 faces and that 5 is removed. Now when I got X equals to 4, what is the probability of getting 4?

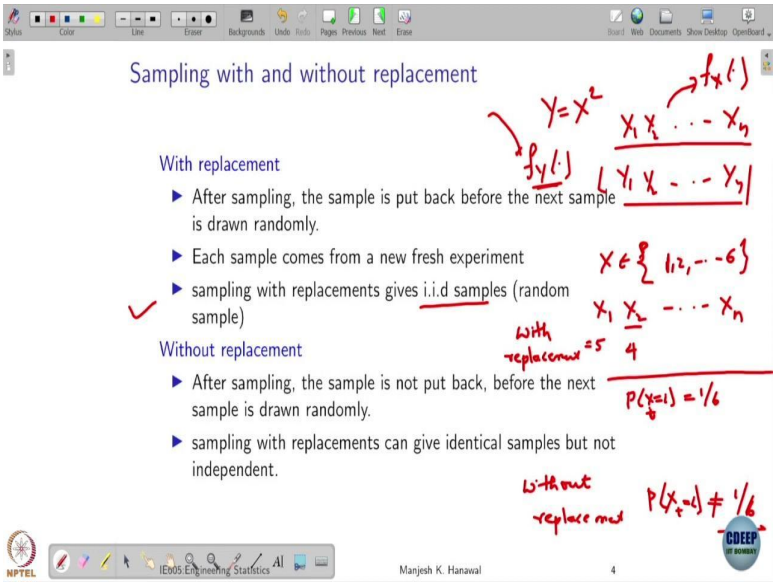
Student: 1 by 5.

Professor: 1 by 5. Now, this is a difference. Now, when I get this is let us say this is like a with replacement and now when I am doing with replacement P of X is always 1 by 6 that is uniform, but now I did sorry this is with replacement. And now when I do without replacement for this need not this is let us say t this is this and this is equals to i. This need not be 1 by 6, if you have removed something which have that faces whatever you observed you have not considered again then this need not be all the time 1 by 6. So, that is what we said sampling with replacements this results in i.i.d identical distributions all the time.

But, if you do sampling without replacement that is no more going to be identically distributed. One thing you need to notice is when we are going to do the sampling with replacement also it is not necessarily that it is always going to be i.i.d. In some cases, you may happen that it may be not independent because if you are not going to perform the experiment in an independent fashion that could be dependency there. So, you have to be that is what will careful. Whenever you are doing the sampling you are doing sampling with or without replacement and what going forward will always assume that we are going to do sampling with replacement all the time, so, that we always get i.i.d samples. Is the distinction between with or without sampling clear? Maybe we should say without.

So, when we do, let us read this sampling without replacement can give identical samples. So, this statement is a little ambiguous. Let me see what I wanted to say here.

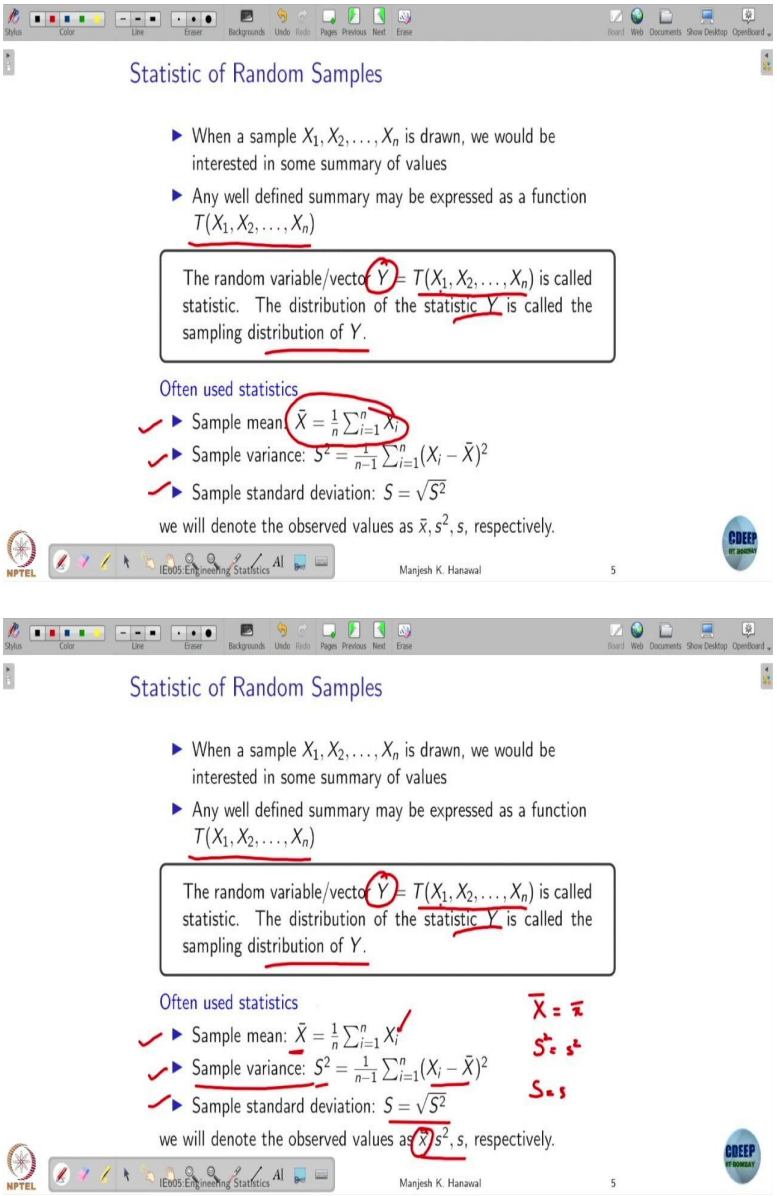
(Refer Slide Time: 18:05)



See, sampling with replacement is definitely going to be given identical samples sorry identically distribution, because the distribution remains the same. But now whether it is going to be independent or not, depends on the way you construct your experiment. If you are ensuring independence that is true then it is not only identically distributed this is also going to be independent samples.

Now I am just trying to like this is I was just trying to highlight the fact here that even if you do with replacement, it may do's to identical underlying distribution can remain identical, but it may not remain independent if you are going to bring in some dependency there. So, if I do without that statement breaks, but let us say let us read it in this fashion only. But it is just like the experiments are not done in identical fashion. Even if you do replacement, but if you do not do identically sorry, if you do not do experiments in an independent fashion i.i.d things may not remain. So, that is what now we are going to follow henceforth is whenever I am saying I am doing a random sampling, I mean is it is with replacement and also it is done independently so that I get i.i.d samples.

(Refer Slide Time: 19:40)



Now, you have samples. You have random samples. Next question is what you are going to do with that? Now with this sample, you are actually trying to extract some information. Now you want to get some summary of this. And whenever we are trying to do summary basically we are looking into some function of this T, some function of the samples, which we are going to denote it as T.

Now, let us say I am output some function on the samples and I got a value Y. Usually this we got by processing this data sample that is called statistic. Notice that here you are X1 to X n, now you are summarizing them by getting one random variable Y which we are through some function T that is called statistic. And whatever the distribution of resultant Y that is called sampling distribution of Y. Now, what are the possible operations I can do, you can any

operations you want, if you have n samples, the T function can be anything. But we will focus on some often used statistics, one is simple take the average of all the samples you have. And this is called sample mean.

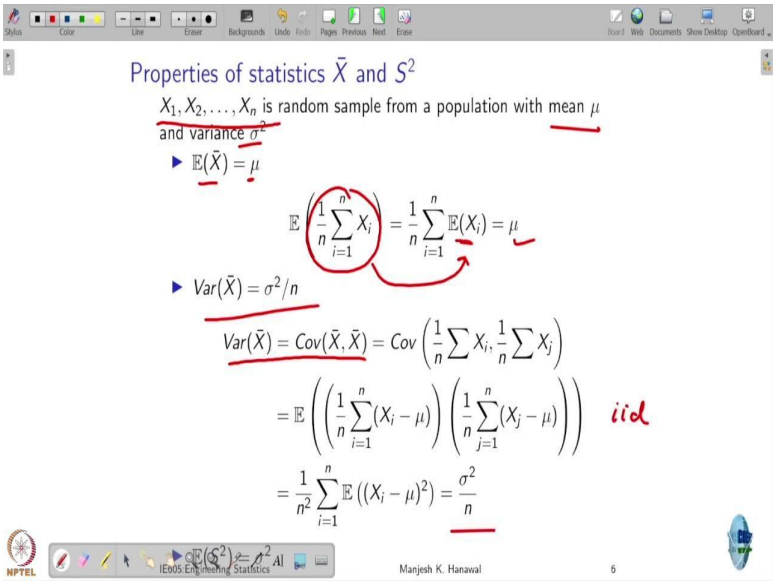
Another thing you do something more you first compute sample means subtracted from the samples and then square them. And divided by anything you want like maybe let us say n minus 1. This we are going to call it a sample variance and denoted by S square and the square root of this sample variance we are going to call it as standard deviation. So, I am highlighting this because these are the most common statistics we are going to use, sample mean sample variance and sample standard deviation.

The realized values of the sample means, sample variance x, we are going to denote it by small letters x bar, s square and s. So, notice that here all I am just representing the samples as random variables, not the particular realizations, but when you plug in the particular realization whatever the value you got, you are going to, let us see, is this sample mean here x bar is a random quantity?

Student: Yes.

Professor: Yes, because that is gotten by averaging random variables. So, this x bar can take some particular value, x bar small x bar, and this s square is also random variable, whatever the realization, it can take, we are going to denote and similarly, s square can also take some particular small value realization or we are going to denote it like s.

(Refer Slide Time: 22:59)



Now, let us look into some other properties of this sample means statistics and sample variance we have. Let us say these X1, X2, Xn they are random samples coming from a population with mean, mu and variance sigma square. The expected value of this sample mean is always going to be the underlying mean value of that population. So, now, let us say why is that so, X bar is nothing but this quantity. You should take the expectation of this. This is nothing but average of this expectation. Is this step correct from this to this?

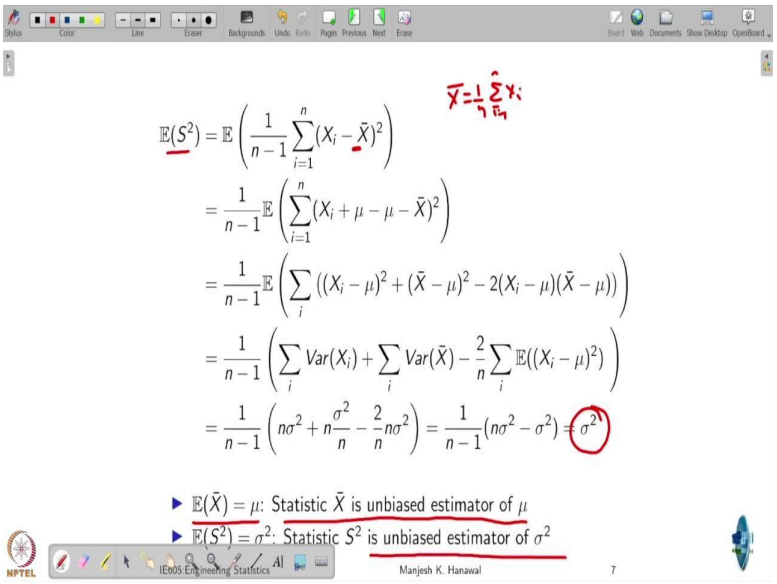
Student: Yes.

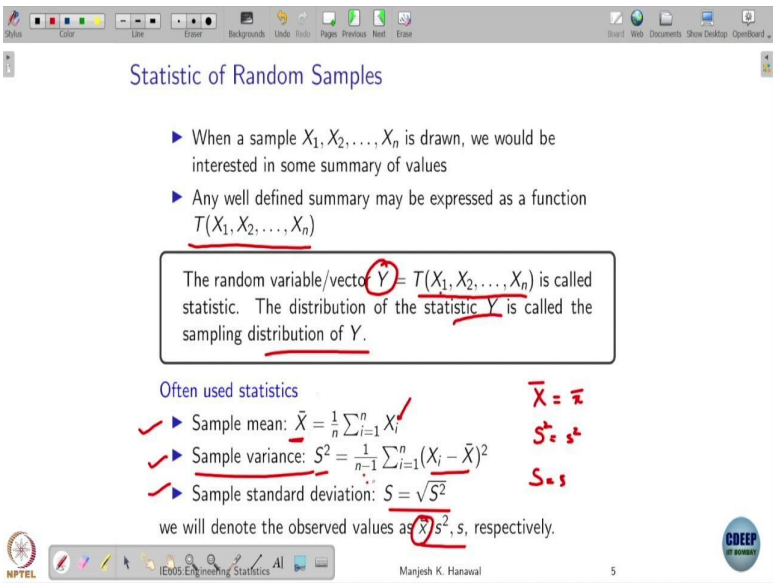
Professor: Why?

Student: Linearity operation.

Professor: Linearity operation and now, expectation of Xi is all mu. And if I add mu n times and divide by n this is going to be simply mu. Now, similarly, you can also compute the variance of the sample mean. Actually, it is always easier to compute variance thinking this as a covariance, like we already discussed that variance of X is nothing but covariance of that random variable with itself. If you just use this and expand and after doing some simplification, and using the fact that these are all i.i.d samples, you will see that this is sigma squared by n.

(Refer Slide Time: 24:38)





Now, let us look into the next one sample variance. If you look into the sample variance, and again compute its expected value, this needs to be done certain manipulation because X bar here is nothing but average of so, X bar here is nothing but Xi by n. So, if you plug in that, appropriately, you will see that and whatever the definition we will get, we are going to see that this value is exactly equals to sigma square.

So, notice that if you people noticed, when I took a sample mean, I divided by n, but when I took sample variance, I did not divide it by n even though I did n terms here, I divided it by n minus 1. Only when I did that n minus 1, you will see that expected value is going to be sigma square. Had I not divided by n, this would have not been sigma square. This will be some extra factor would have come here.

So, now the definition is, whenever expected value of this sample mean is mu that is the underlying mean value, this statistics X bar is called unbiased estimator of mu. Notice that now I am calling X bar as an estimator. And whenever this expectation of X bar is mu, we are calling it as unbiased estimator. And similarly, expectation of sample variance is also sigma square. And in this case, the statistics X squared is called unbiased estimator of your variance term. So, sample mean and sample variance, they are the estimators for mean and variance further they are unbiased. So, let us stop here.