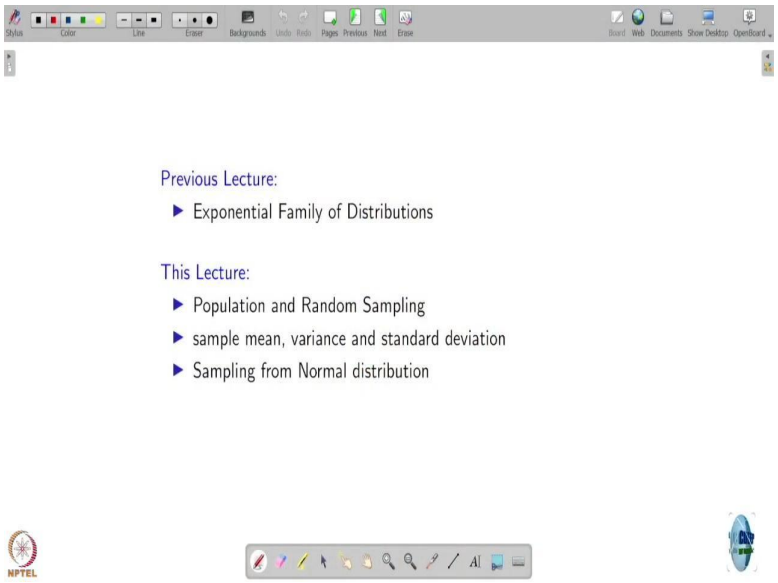
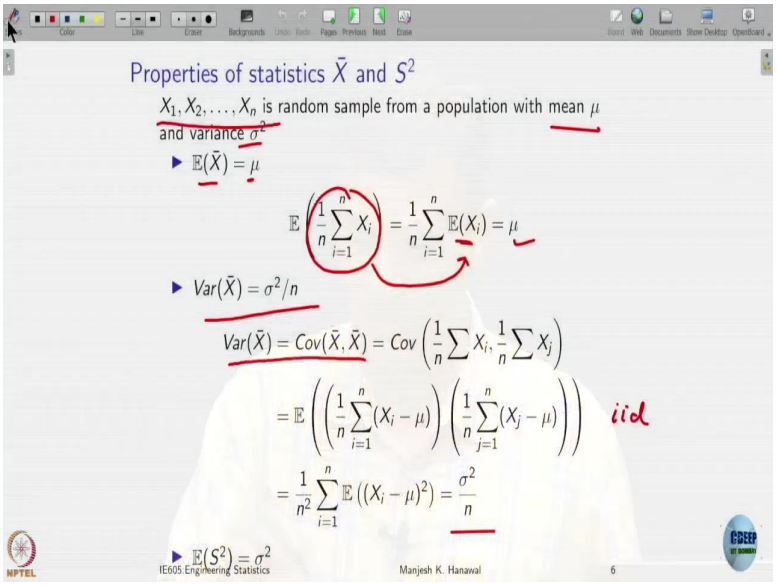
Engineering Statistics
Professor Manjesh Hanawal
Industrial Engineering and Operational Research
Indian Institute of Technology, Bombay
Lecture - 22
Sampling from Gaussian distribution and t-distribution

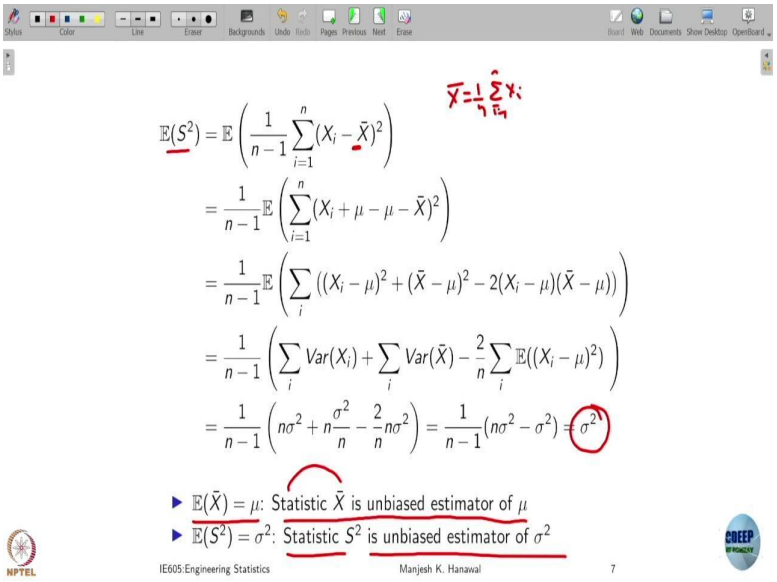
(Refer Slide Time: 0:24)



So, in the last lecture, we started talking about random sampling, we discussed about what we mean by random sampling from a particular population. We said that for us population is nothing but some underlying probability density function and when we samples from this population in an i.i.d fashion, we call that as random sampling, and then we started talking about sample mean and sample variance. So, these were the definitions of sample mean sample variance and sample deviation. Then we looked into some of their properties like expected value of sample mean is nothing but the mean of the underlying population and the variance of the sample mean can be calculated to be sigma square by n.

(Refer Slide Time: 1:30)

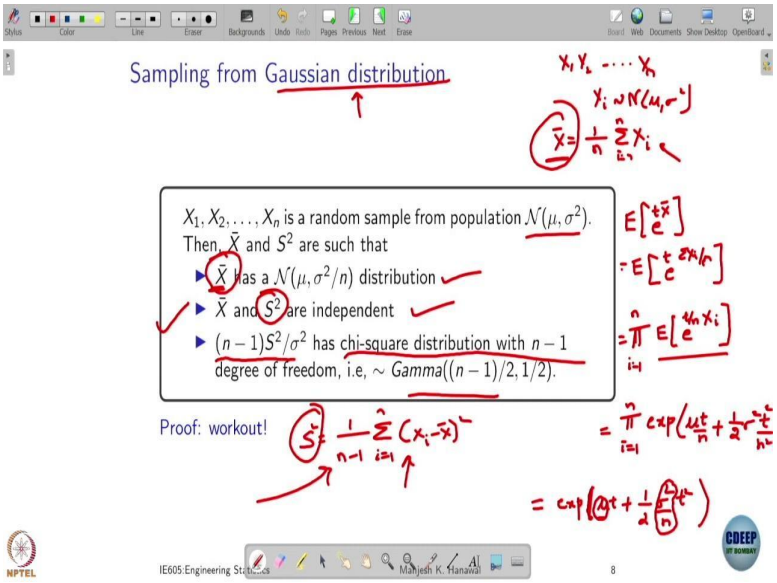




And we also said that the expected value of standard sample variance is nothing but the variance of the underlying population. And we actually showed this I just skipped a computation and I asked you to verify this. I hope you verified. So, you notice that like if I say some computation to verify, please verify this may come in your quiz or any place and at that time you should not repent.

And then we said that when the expected value of sample mean is equals to the mean of the underlying population, then the sample mean which is basically one of the statistics we consider, we call that statistics as unbiased. And similarly, the another statistics we have a statistic we have is sample variance and we said that sample variance is also unbiased.

(Refer Slide Time: 2:42)



Now, let us move on. When you do sampling and all, we have to do all the analysis with only some finite number of samples. Maybe when you do samples, when you collect samples from people, maybe you will collect maybe 100 people, 200 people, 300, 400, maybe 1000 or maybe at max few lakhs. And based on those samples, only you have to do analysis or try to understand what is the underlying mean and variance of the population.

And always collecting samples is a expensive task. You should not assume that, like I will just to go and get as many samples I want. No. Like if you have noticed, like when all these trucks are being tested, different vaccines are such tested, they were trying to do it to different people and see what is the effect of that drug. And based on that they were trying to come up how how efficient their drug is.

So, giving it to someone, and then seeing that how drug affects him. That is like collecting one sample. Naturally, asking more people to come for this drug testing is a very expensive thing. Like if by mistake something happens to anybody, that is a lot of liability is there so, getting more samples is not always easy and you should not be assuming that sample set to like I will get as many samples.

And another thing, if you want to do a go and do market survey, like if you are launched a new product and wanted to see how many people will be interested. You have to set up a whole process of getting the sample. You have to put somebody to and talk to people like that maybe you have to employ somebody. You have to pay him for that. And he has to go and talk to people. So, his travel needs to be arranged, all these things. These are all they cost

money. And I mean all these people who do poor prediction, like you see that they put a mammoth system for this, they put a lot of people. They will send to different regions. They will talk to people. That is again involves a lot of money.

And now things are not simple, most of the samplings, maybe people put it on just opinion polls like just to make some web page and keep on sharing it and maybe like a through Google form or something, ask people to give their inputs that is like asking somebody is like, you are basically getting one sample from that person. But that itself is also going to take maybe cost you something and it is not that everybody you can approach anybody like if you start sending everybody will just flag it as spam and they will not get any response. So, a lot of issues are there and that is why we should always be ready to see what kind of information we have from a given number of samples.

Now, the question is how to go about analyzing this? So, to simplify things, what we will do is we will assume underlying population is Gaussian. This is our first simplification and try to understand how the analysis goes about, and then we will consider what happens if it is non Gaussian . If suddenly if when the things are non question, things becomes complicated, somehow Gaussian distribution is something which is very amenable for analysis, mathematical analysis, that is why most of the times we consider from the theoretical point of view we assume Gaussian distribution, that is why this is one of the popular distribution because it is easy to analyze.

Now, when we have samples, we are going to do some statistics on them. It could be mainly sample mean or sample variance. And now try to understand some of the properties of this when the underlying population is Gaussian distribution. Now, if you have samples X1, X2 are all random samples and the underlying population is Gaussian with mean, mu and variance sigma square the sample mean itself is going to be Gaussian distributed with mean, mu and variance sigma squared by n.

Let us now, is this obvious? So, what you are saying, let us say we have this X1, X2 up to Xn and each of these Xi is mu sigma square and now, we are saying X bar is 1 by n summation Xi. How do you check that this X bar is Gaussian distributed, what is the measurement method you have?

Student: Moment-generating function

Professor: One possibility is moment generating function. See that what is the moment generating function of X bar? How we are going to find that? It is going to be 1 upon e to the power t x you do it for some X and that this is nothing but e to the power t summation Xi by n.

Now, I can write it as product of 1 to n expectation of e to the power t by n Xi, can I do this? Why this product is correct?

Student: Independence

Professor: Because they are independent. Now, what is its value? You know what is the moment generating function of a Gaussian distribution. Compute it at what point?

Student: t/n

Professor: t by n here. What is that value? Can somebody tell me. Exponential plus half, is this correct? I do not I am just taking you and now let us simplify this. Now with this it is going to be exponential. I am adding now all of them. So, mu t becomes mu t plus half sigma square t square by n. Now, what is this corresponds to, what this distribution corresponds to?

Now, if you compare it with a template of the normal distribution this is going to have a mean of mu and variance of sigma square by n and that is why this is going to be Gaussian with parameter mu and variance sigma square by n. Now, x squared we have.

Now, what is our S square? S square is 1 by n summation, now, what we are saying is this X bar and S square they are independent. So, only first look does it make sense? Because they all these X bar now, we are talking about this X bar and this S square. Both X bar and S square they all depend on the same set of samples Xi. See, Xi is let us take on particular Xi that Xi is there, and that Xi is also there in some so, both of them do we expect it to be independent? Or at least what your intuition says like? It is hard to believe that, this X bar and S square which both of them depends on the same set of samples, they are independent. But for us, we will go with the definition of independence, what is the definition of independence?

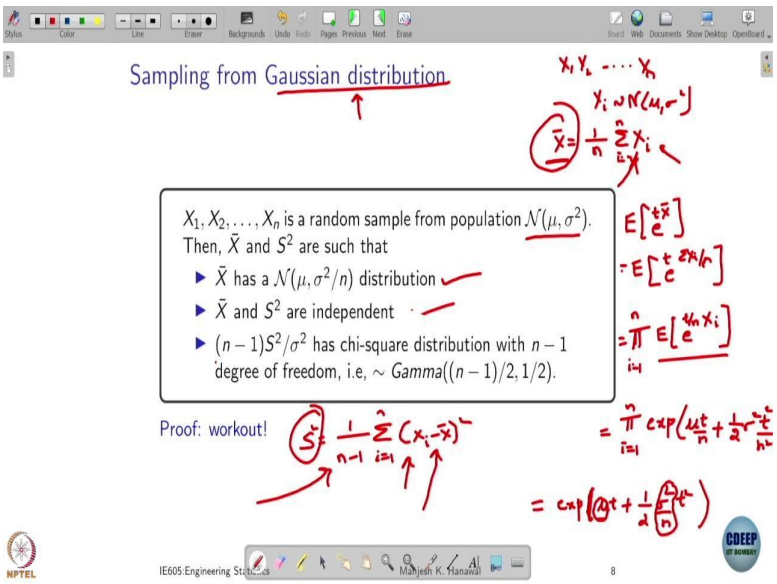
If you compute the distribution of X bar and compute the distribution of S square, and if you look into the joint distribution, they should split into the marginals and it actually happens for this whenever it is Gaussian. And that is why we are going to call I mean, that is why it is a result that they are independent. And in fact, you can show this and this required some good

amount of calculations. That is why I am skipping that. And it is there in the book, which I am going to share and I will refer you which chapter and which section you have to look into for this calculations. What we are again going to do is you are going to find now you have already have X bar distribution you already have.

Now you need to figure out S square distribution. But finding the distribution of S square is going to be hard. Instead, what you are going to do is we will look into their joint moment generating function, and we will show that they are joint moment generating function actually splits through that you are going to show they are independent and this is one of the important properties that we are going to use later. And thanks to the special structure of Gaussian distribution, this property holds. Otherwise, if you replace this by any other distribution, this need not hold another property.

Now, this sample variance, if you multiply by n minus 1 and divided by sigma square, this has a chi square distribution with n square degrees of freedom, n minus 1 degrees of freedom. So, notice that X bar has Gaussian distribution and when you S square when you divide and multiply accordingly, it will have chi square distribution with n minus 1 degrees of freedom. And by our definition, we know that chi square distribution is related to the gamma distribution. So, chi square distribution with n minus degrees of freedom is nothing but gamma distribution with parameters n minus 1 by 2 and half.

(Refer Slide Time: 15:31)



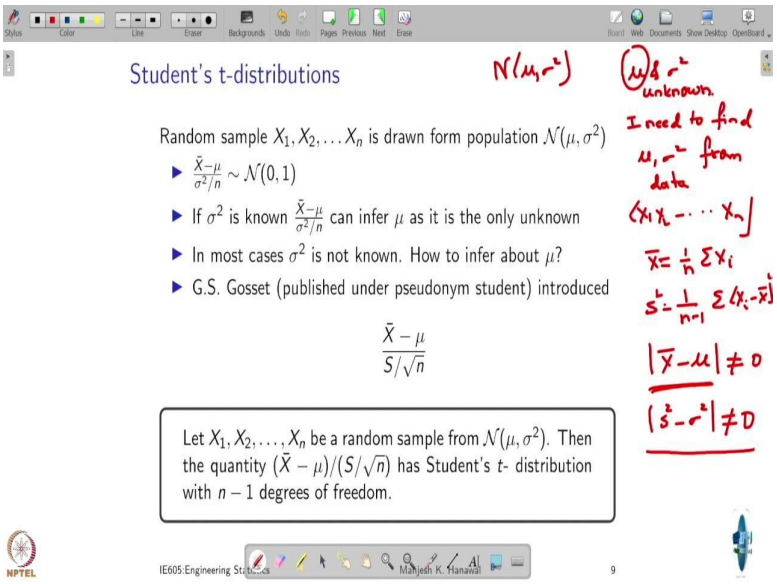
So, now at this point you may be wondering X bar consist like when I defined X bar it has n random variables in it and each one of them is like a Gaussian random variable in this case.

And S square also consists of this n random variables then why it is not n degrees of freedom like I can think of these are like n components and which all of them are like independent. They can vary in an arbitrary fashion, why is that they are not n degrees of freedom, why it is that n minus 1?

Student: X bar acts as a constraint for bringing the values (()) (16:17).

Professor: So, even though we have n components, but there is an X bar here and that X bar is affecting my S square the average is affecting and because of that, you can just work out that like even with that, if you take the n minus 1 components, the nth components get fixed because of this x bar part here. So, because of that one degrees of freedom get reduced and you will end up with n minus 1 degrees of freedom.

(Refer Slide Time: 17:03)



Now, let us continue to analyze this random sampling of this Gaussian distribution itself. And when we try to analyze this. Say, I will tell you at this point, what is our ultimate goal? We want to get the parameters from samples. Let us say we already know it is a Gaussian. Already we have put the structure that is mu and sigma square but I may not know mu and sigma square, mu and sigma square unknown. And what I have is I have access to the data. I have samples X1, X2 up to Xn and I need to mu and sigma squared from data that is X1, X2 this random sample. From this random sample at best I can find the sample mean and sample variance but this X bar is it same as mu that I want mu but this X bar is not same as mu.

So, that mu there will be some difference. This need not be 0 and similarly, this S squared minus sigma squared this need not be same. Now, what I want to understand, how much is this actual difference is, if I had to compute this only n sample, I know that this is not going to be 0. It is going to be some positive value, but how much is that? And how can I quantify that? Similarly, this also I know this is not going to be 0. But how much is that error? And I need to quantify that.

So, for that now we will start thinking about how to do that. And for that we will use all the Gaussian properties because in general, we cannot compute it for any distribution easily, but we focus on Gaussian distribution and try to get it. So in that process, we are going to now in doing that we will end up studying something called Student t-distribution.