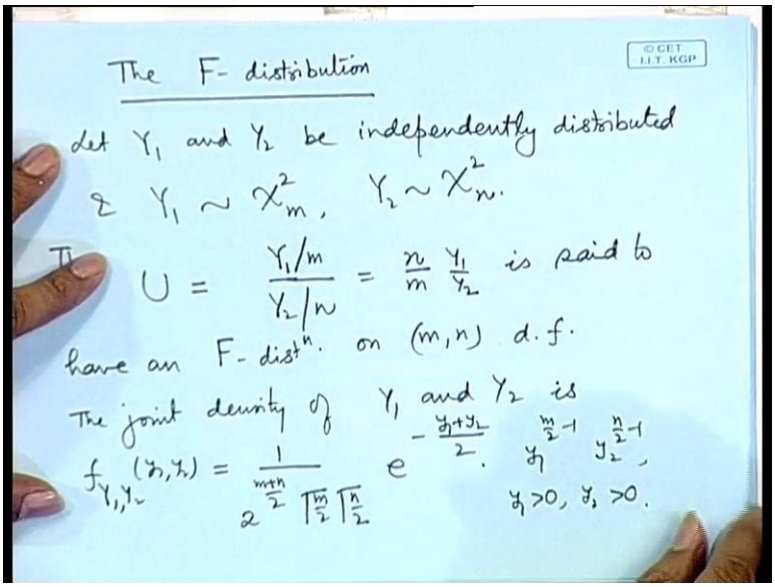
Probability and Statistics
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture - 48
F-Distribution

Sampling distribution which is important and it is used quite frequently is F- distribution.

(Refer Slide Time: 00:27)

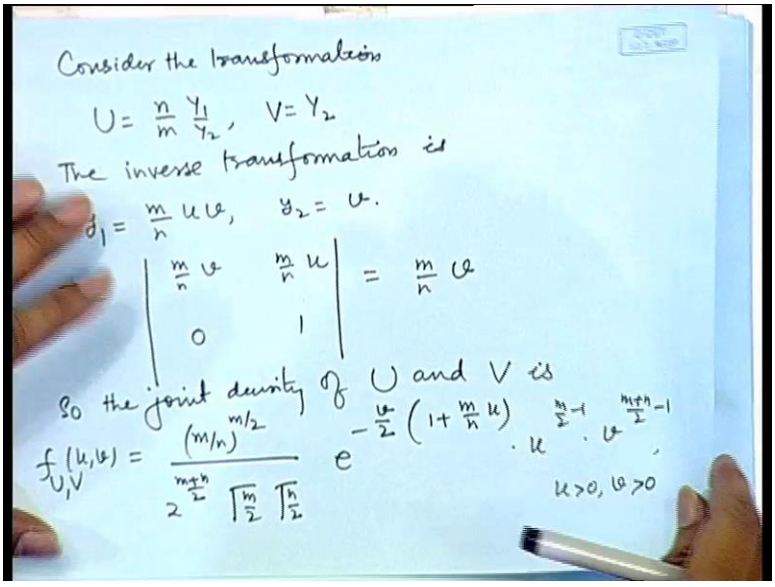


So, now I introduce the F- distribution here. Let Y 1 and Y 2 be independently distributed random variables; and Y 1 follow say chi square distribution on m degrees of freedom and Y 2 follows say a chi square distribution on n degrees of freedom. Then if we define a variables called U as the ratio of this Y 1 and Y 2, but divided by their degrees of freedom that is Y 1 by m divided by Y 2 by n that is basically becoming n by m Y 1 by Y 2, then this is said to have and F- distribution on m n degrees of freedom. Now here one has to notice that two degrees of freedom terms are coming and therefore this order is important. If I am having a numerator chi square variable as m and the denominator chi square as n then we will write the ordered pair m n; that means, if I write n m, it will denote a different F- distribution. Now by our theory of transformation of random variables, U is a function of Y 1, Y 2; therefore, I can use a new dummy variable V and find out the joint density of U and V to derive the probability density function of U.

So, for that purpose we write the joint distribution of Y 1 and Y 2; the joint density of Y 1 and Y 2 is f of Y 1; Y 2. So, basically we multiply the individual densities of Y 1 and Y 2 which are basically chi square densities on m and n degrees of freedom. So, if we combined the coefficients 1 by 2 the power m plus n by 2, gamma m by 2 gamma n by 2; e to the power minus y 1 plus y 2 by 2; y 1 to the power m by 2 minus 1, Y 2 to the power n by 2 minus 1, where both y 1 and y 2 are positive.

So, we consider the transformation in which U is this variable.

(Refer Slide Time: 03:21)

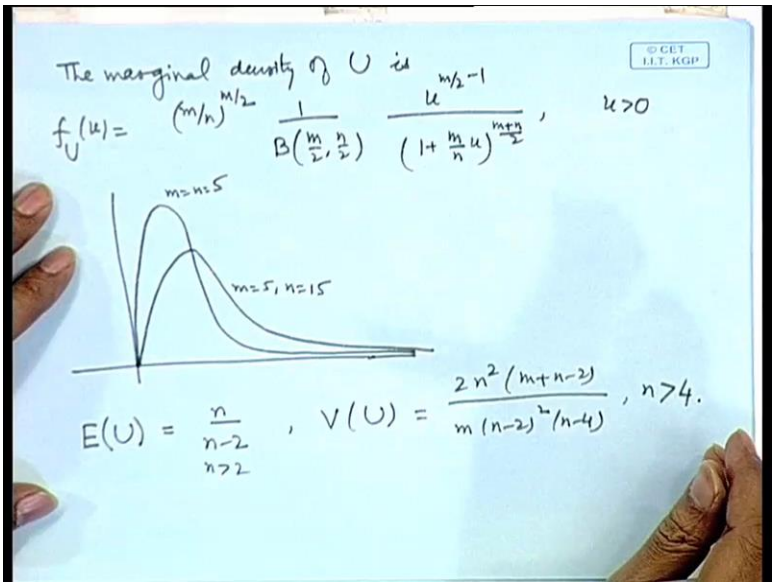


So, consider the transformation U is equal to n y m; Y 1 by Y 2 and V is equal to say y 2. So, the inverse transformation is Y 1 is equal to m by n u v; Y 2 is equal to v. So, the (Refer Time: 03:58) of the transformation can be calculated as m by n v m by n v u; 0 1 which is basically m by n v and since the terms are positive, so absolute value of j will also be the same. So, the joint density of U and V is; now you can observe, we are having here this constant term and we will be replacing y 1 by m by n u v and y 2 by v. So, here v by two terms will come out and this will give additional powers of v, for power of u will be this one only.

So after adjustment of the terms, we can write it as m by n to the power m by 2 divided by 2 to the power m plus n by 2 gamma m by 2, gamma n by 2; e to the power minus v by 2; 1 plus m by n u; u to the power m by 2 minus 1; v to the power m plus n by 2 minus 1, where u and v are positive variables. So, we can integrate with respect to v from

0 to infinity to get the desired density of f random variable. Again if we have observe the integral of this with respect to U is nothing, but a gamma function where the order is m by 2 and the coefficient is 1 plus half into 1 plus m by n u.

(Refer Slide Time: 05:51)

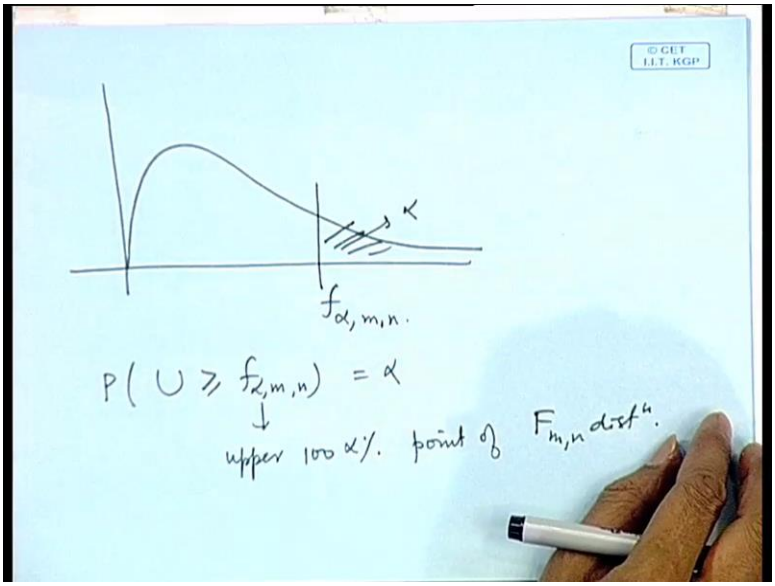


So using a standard argument, the marginal density will turn out to be the marginal density of U is then; f U is equal to m by n to the power m by 2 and after the cancelation of the terms, we will get it as 1 by beta m by 2; n by 2; u to the power m by 2 minus 1 divided by 1 plus m by n u m plus n by 2 for u positive.

Obviously, this is a distribution of a positive value would random variable and it is positively skewed. However, the shape will vary depending upon the values of m and n. So, I can just give one example here; if I consider say m and n is equal to 5 then the form of the density is some what like this if I consider say m is equal to 5 and n is say 15, then the form is something like this. So, likewise for different values of m and n you get different shapes of the curves. Here calculation of the moments will make use of different beta functions. However, I will write the mean and variance, the mean of this is n by n minus 2; you may be little bit surprised here that it is dependent only upon second variable because n is not appearing here.

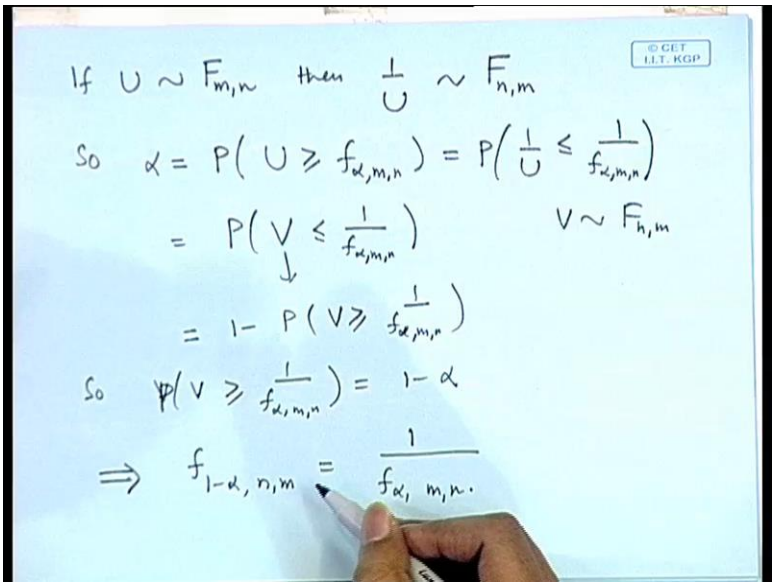
Variance of U is equal to here you need n greater then 2 and variance term is twice n square m plus n minus 2, divided by m into n minus 2 is square into n minus 4; this is valid for n greater then 4, this positively skew distribution.

(Refer Slide Time: 08:00)



So, we concentrate on the points here, so if I have this probability equal to alpha then this point is termed as f alpha m n that is the probability of U greater then or equal to f alpha m n is equal to alpha; that is this is upper 100 alpha percent point of F m n distributions.

(Refer Slide Time: 08:40)

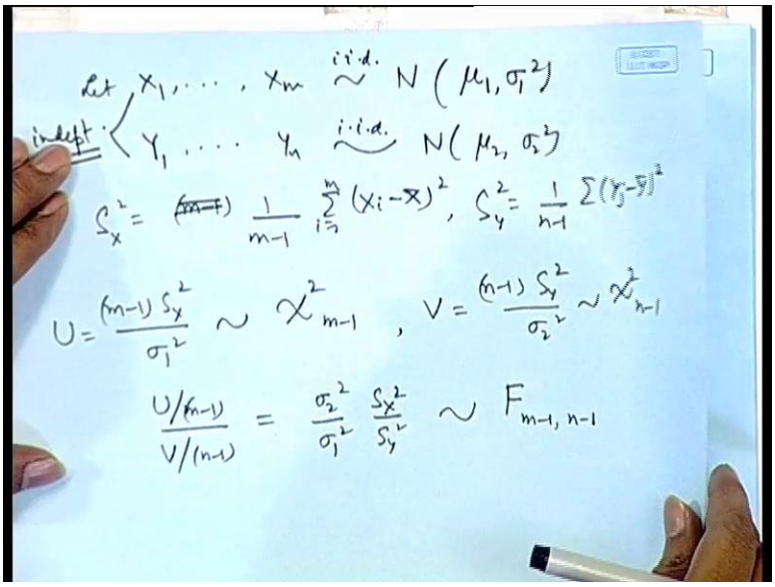


Now by the definition of this f variable, it is clear that if I have U following F m n then 1 by U will follow F n m because 1 by U remains a ratio of chi square variables divided by their degrees of freedom; however, the numerator degrees of freedom have gone to the denominator and the denominator degrees of freedom have gone to the numerator.

So, this becomes F- distribution on n m degrees of freedom, so we can derive a formula for the points of F- distribution. So, we defined that f alpha m n is the upper 100 alpha percent point of the F- distribution on m n degrees of freedom. So, probability of U greater then or equal to this is equal to alpha. So, if I write it as probability of 1 by U less then or equal to 1 by f alpha m n; then this 1 by U is a F n m variable that this is probability of sum V less then or equal to f alpha m n; that is this v is F n m variable. So, if I am saying that the V greater then 1 by f alpha m n; it is equality or inequality does not play any role here because of this continuous distributions. So, I am saying probability V greater then or equal to 1 by f alpha m n is equal to 1 minus alpha, but V is F n m distribution, this implies that f 1 by alpha n m is equal to 1 by f alpha m n.

This relationship is used for calculation of the percentage points of the F- distribution and generally the tables are because here it is a two dimensional table, you have m and n both varying and therefore, only for selected values of alpha; the tables are given. Now if they are given for say alpha is equal to 0.05 or alpha is equal to 0.1 then 1 minus alpha becomes 0.95 and 0.9 respectively. So, those tables can be automatically derived from the tables of 0.05 and 0.01 values etcetera.

(Refer Slide Time: 11:20)



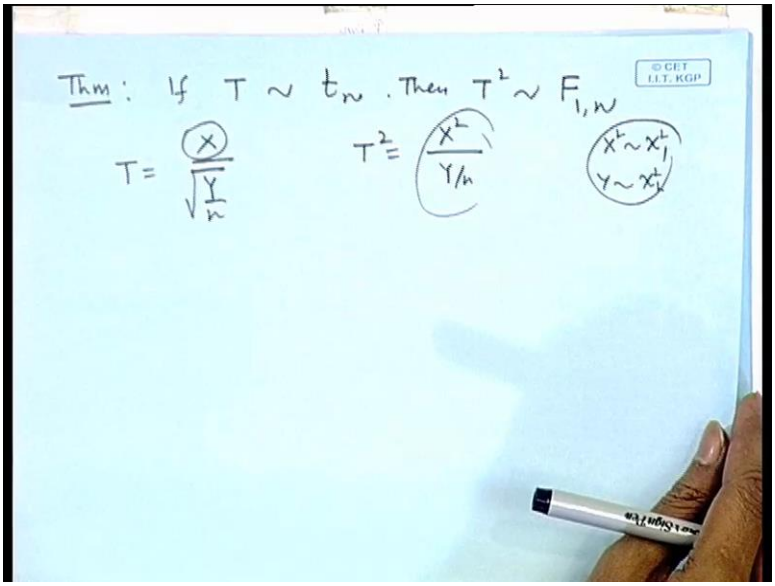
Now, we look at how in the sampling it arises, so if we consider say a random sample say x 1, x 2, x m following normal mu 1, sigma 1 square and say Y 1, Y 2, Y n be a

random sample form normal say mu 2 sigma 2 is square and also I assume that these samples are taken independently.

Let us define the quantity say S x is square as m minus 1; 1 by m minus 1 sigmaXx i minus X bar square i is equal to 1, 2 m and say S y square as 1 by n minus 1 sigma Y j minus Y bar square. Then by the theory of chi square m minus 1, S x square by sigma 1 square follows chi square distribution on m minus 1 degrees of freedom and n minus 1; S y square by sigma 2 square follows, chi square distribution on n minus 1 degrees of freedom; let me call its say U variable and this as the V variable. Then if I take the ratio U divided by m minus 1 divided by V divided by n minus 1. Then this is nothing, but sigma 2 square by sigma 1 square; S x square by S y square that follows F- distribution on m minus 1, n minus 1 degrees of freedom.

So, this relationship or this result is used quite frequently in drawing inferences on ratios of the variances because ratios of the population variances and ratios of the sample variances is occurring here.

(Refer Slide Time: 13:23)

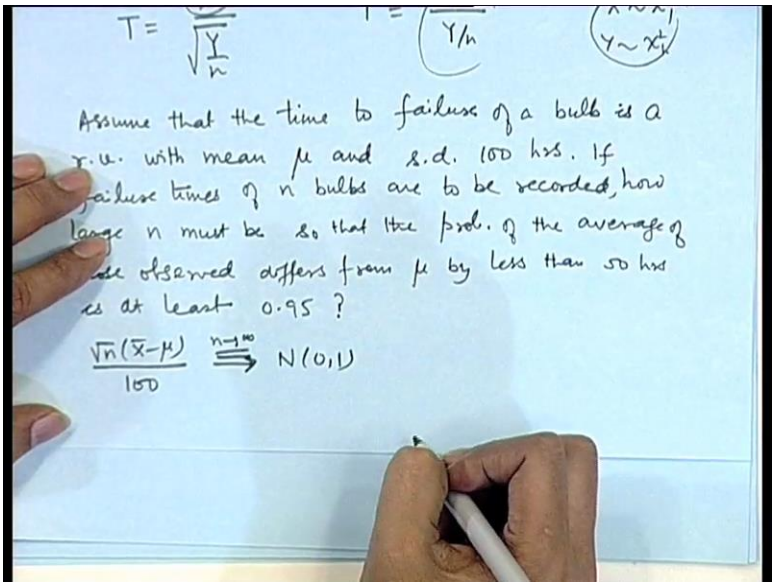


Another relationship which is coming here is that if say T follows t distribution on n degrees of freedom then T square follows F- distribution on 1 and m degrees of freedom. One can prove it by direct transformation by writing down the density of t and making the transformation U is equal to t square there and writes and compare with the forms of the densities. However, one can look at an easy representation see we can write T as X

divided by a root Y by n, where X is a standard normal and Y is a chi square variable, so if I look at t square that is X square by Y by n.

Now, this X square will be chi square on one degrees of freedom and Y will be chi square on n degrees of freedom. So, this is nothing, but the definition of an f variable on 1 and n degrees of freedom. Sampling distributions are extremely useful in a statistical inference and they are used all the time, I will give a couple of applications here.

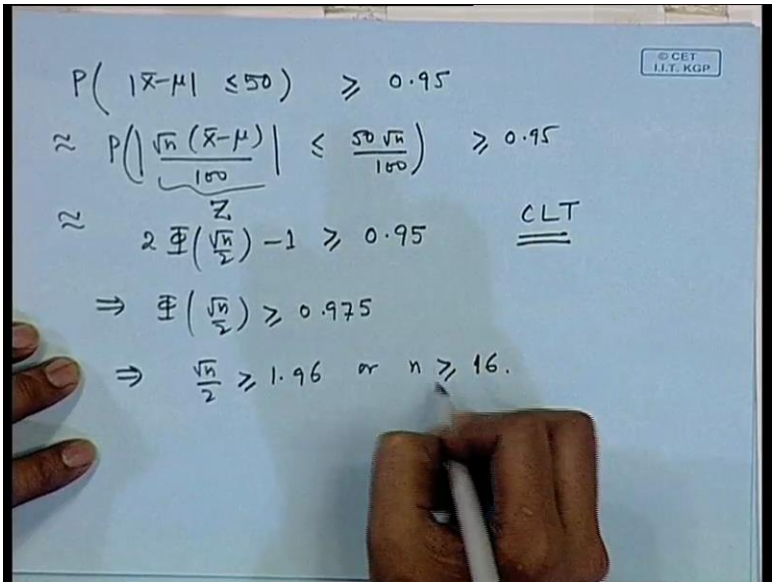
(Refer Slide Time: 14:47)



Assume that the time to failure of a bulb is a random variable with say mean mu and standard deviation say 100 hours. If failure times of n bulbs are to be recorded, how large n must be so that the probability of the average of those observed differs from mu by less than 50 hours is at least 0.95.

That means how much should be my sample size such that the sample average and the population mean should differ Y less then 50 percent and this probability should be at least 0.95. So, we can make use of the central limit theorem here because the only information about the distribution that we are having is that; it is a particular distribution with certain mean and certain variance. So, the conditions for application of the central limit theorem are valid here. So, we will have a root n x bar minus mu by sigma that is 100; this will be approximately normal 0, 1 as N becomes large.

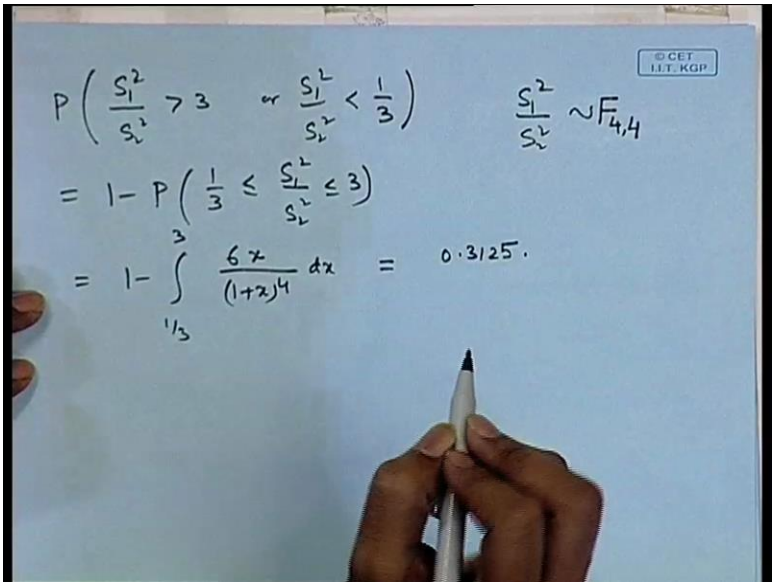
(Refer Slide Time: 16:58)



So, by this statement we have the condition here that probability of modulus x bar minus mu less then or equal to 50. We want to put this to be greater then or equal to 0.95, so we approximate this probability by converting to; by making use of the central limit theorem. So, this is less then or equal to 50 root n by 100, so this is approximately a standard normal random variable then this probabilities I can replace this variable by Z, where Zis a standard normal variable. So, this can be written in terms of the CDF that is twice phi root n by 2 minus 1 greater then or equal to 0.95. So, I have made use of the central limit theorem here and this gives us phi of root n by 2 greater then or equal to 0.975. So, from the tables of the normal distribution root n by 2 must be greater then or equl to 1.96 or n must be greater then or equla to 16.

So, we need minimum sample size 16, so that the sample average and the population average do not differ by more than 50 and the probability of that should be at least 0.95. Let me give one more problem here.

(Refer Slide Time: 18:31)



So, we consider independent random samples of size 5 from two normal populations and they have the same variance. So, what is the probability that the ratio of the larger to the smaller variance exceeds 3, so this problem is important in the following sense, see we have taken 2 samples from the same population. Now we want to check whether the there is too much variability in the sampling process. So, we look at the variances of the 2 samples and one of them will be naturally larger and one will be smaller. So, we are saying that the ratio of the larger to the smaller exceeds 3 what is the probability of this event.

So, if we write in terms of S 1 is square and S 2 is square basically we are requiring here what is the probability that S 1 square by S 2 square is either greater then 3 or S 1 is square by S 2 is square is less then 1 by 3 what is the probability of this. Now if we have taken the samples of size 5 each, then by the formula that n minus 1 by m minus 1; S 1 square by this 1, we get that S 1 is square by S 2 is square follows F- distribution on 4 and 4 degrees of freedom; that means, for calculation of this; we have to look at the either the tables of F 4 4 and we write down the density of F 4 4 here.

Fortunately the density of F 4 4 becomes a quite simple form. So, we can write this as 1 minus probability 1 by 3 less then or equal to S 1 sqare by S 2 is square less then or equal to 3 and this turns out to be 1 minus 1 by 3 to 3 and the density function of F 4 4 is 6 x

divided by 1 plus x to the power 4. So, this integration can be done easily and the value turns out to be 0.3125.

So, various problems which relate to the sample means or the sample variances or the comparison of the means or comparison of the variances can be solved using sampling distributions. In the portion of point estimation confidence interval estimation and testing of hypothesis, we will have frequent uses of these sampling distributions. So, in particular we have considered normal distribution itsef has a sampling distribution because for the large sample, any sample mean will be approximately normally distributed under certain condition of course.

And then when we are sampling from normal distributions, then certain functions which are related to the means and the variances; they are having chi square, t, and F-distribution. So these are in particular 4 important sampling distributions. There are many more sampling distributions, but they are not as frequenctly used in practice.