

about sampling distribution we talk about the distribution of a statistic computed from the sample that distribution of that.


Now, here \bar{x} is a sample statistic. What is the distribution of \bar{x} ? Similarly, s^2 is a sample statistic, what is the distribution of s^2 under sampling distribution? We will discuss these two, but please keep in mind in general if we say a θ is equal to. Let it be θ_1, θ_2 like suppose θ_k , there are k parameters for θ . Then definitely, suppose your n is the sample size of statistic. Then corresponding to every population parameter, there will be sample statistic be there. Now, by sampling distribution we say, what is the distribution of this statistic. What is the distribution of this statistic, what is the distribution of this statistic. It may be univariate, it may be multivariate.

So, if you are interested to know collectively, what is the distribution of this statistic vector $k \times 1$ vector, then it will be multivariate distribution, probability distribution. Essentially, by sampling distribution you are talking about probability distribution only. But the difference is you are talking about the distribution of not the variable characterizing the population, rather the statistic computed from the sample distribution of that statistic, clear? It is for example, we have taken the, if we consider the same example again, the profit if you see the 12 months data, 1 to 12. So, your profit is there different values of profit. So, you will calculate mean profit average per month. My question here is, what is the distribution of \bar{x} ? So, with this line today's discussion is sampling distributions.

(Refer Slide Time: 05:41)

Contents

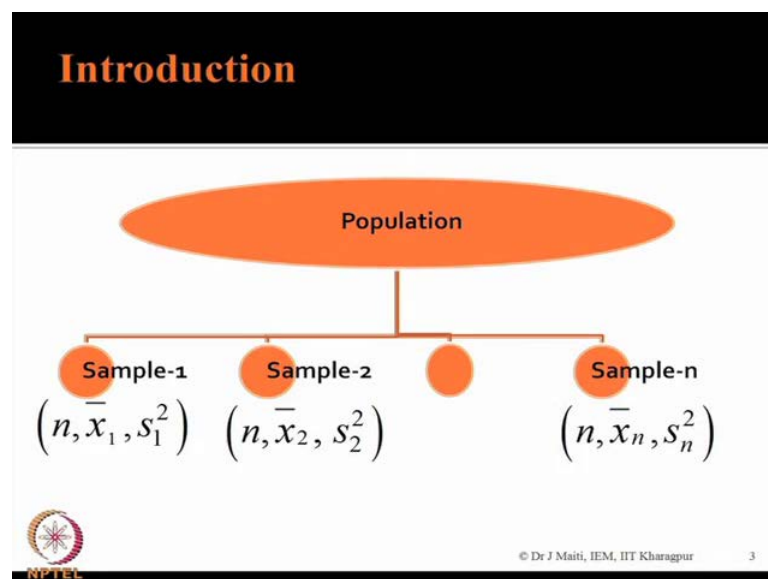
- Introduction
- Sampling distributions
 - z
 - χ^2
 - t
 - F
- Central limit theorem
- Sampling strategy
- References



© Dr J Maiti, IEM, IIT Kharagpur 2

And you see that there are unit normal distribution? z stands for unit normal distribution, the chi-square distribution, t distribution, F distribution. These are the most popular, mostly used, widely used distribution for sample statistics. Then we will discuss central limit theorem. Finally, sampling strategy, because how do you collect the data? It all depends on what strategy you will be adopting, because the data collection process if wrong or faulty analysis will not give you good result. So, you may be wondering that where from there these.

(Refer Slide Time: 06:36)



Suppose you have collected one sample. For example, you collected the height of people for a particular community, suppose n data points.

(Refer Slide Time: 06:41)

C	X
1	x_1
2	x_2
...	...
n	x_n

Statistic is a r.v.

Sample	\bar{x}	s^2
1	\bar{x}_1	s_1^2
2	\bar{x}_2	s_2^2
...
n	\bar{x}_n	s_n^2

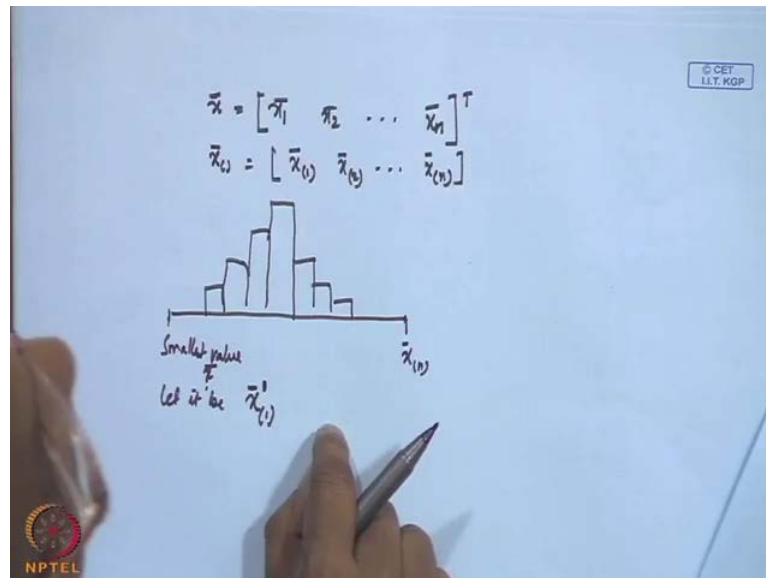
You have collected 1, 2, dot, dot, dot, n. Your value x value is x_1, x_2, \dots, x_n . You are thinking that I have computed \bar{x} here and s^2 here. And these are all because once you collect data, this the these are the fixed data. So, your \bar{x} is also a fixed value, s^2 is also a fixed value for a particular sample. Then where from this distribution is coming? The distribution concept is coming, because if you do the same sampling next time, there is no guarantee that you will get the same value of \bar{x} , the same value of s^2 . That is why we in statistics, all statistics are or otherwise each statistic is a random variable statistic.

In this figure, you see this figure that in sample, n samples, sample 1 to sample n collected from a particular population and the sample size is n for all the cases. Sample size means the number of observation collected per sample. And if you compute the mean and variance from each sample, you will be getting for first sample. It is \bar{x}_1 and s_1^2 , \bar{x}_2 and s_2^2 for second sample, and it is obvious that all those values may not be same, actually it will be different. Suppose, if I say that sample, sample 1, sample 2 like, sample n .

Then if you calculate your mean and calculate variance, \bar{x}_1 is the first sample mean, \bar{x}_2 is the second sample mean, like this \bar{x}_n is the n th sample mean. s_1^2 is

the first sample, standard variance and s^2 square is the second sample variance. Like this the variance s^2 . Now, as these are not same value, what will happen if you can draw a histogram?

(Refer Slide Time: 09:33)




How you can find out the smallest value of \bar{x} ? Let it be, if I say \bar{x} within bracket 1 bar, originally what happened? Originally you have \bar{x}_1 bar, \bar{x}_2 bar, like this \bar{x}_n bar. If I say this is a vector \bar{x} which is this one. This T stands for transpose, then you are ordering them from smallest to largest, then the same thing. What will happen if I say that some order we are giving? Then this will be \bar{x}_1 bar, \bar{x}_2 bar, \bar{x}_n bar. So, this is the smallest one and the largest one is \bar{x}_n bar.

So, if you plot you will get, you will develop histogram. You may find out a distribution, it may be like this. It may be like this. So, that means what we are trying to say, that if you collect sample 1 after another from the same population of same size for the same variable, you will get different values for that sample statistic. And that sample and that is why the sample statistic is a random variable. And you have a probability density function for a random variable and that density function is the sampling distribution function.

(Refer Slide Time: 11:43)

Sampling distribution

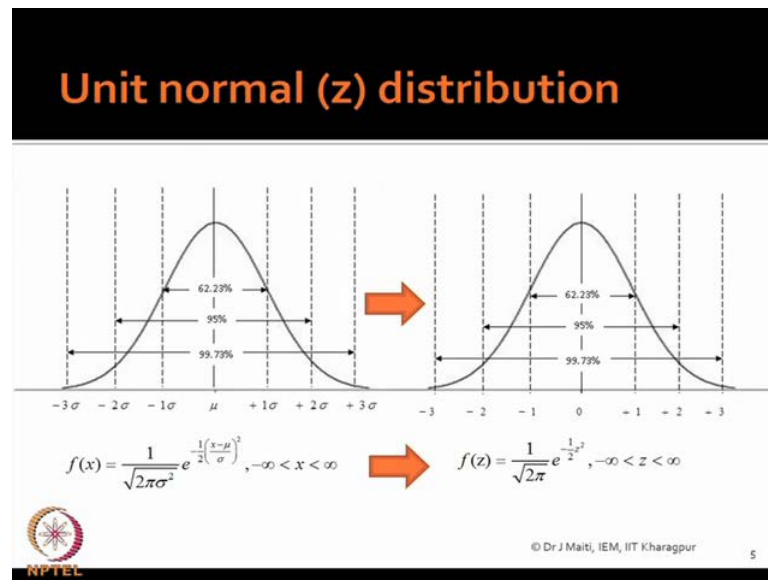
- Distribution of a statistic
- z distribution
- χ^2 distribution
- t distribution
- F distribution

 Dr. J. Maiti, IEM, IIT Kharagpur 4

Now, you have to know that what are the distribution that comes under sampling distribution. I told you earlier that there are four most popular sampling distribution, like z distribution, chi square distribution, t distribution and f distribution. And these concept is very, very vital concept, I am telling you and very, very important. Later stage is when we talk about any multivariate model. For example, for multiple regression, there are several beta coefficient. You will be finding out that regression coefficient, these regression coefficient through data, Sample data you will be estimating. They are parameter, they are basically the statistic.

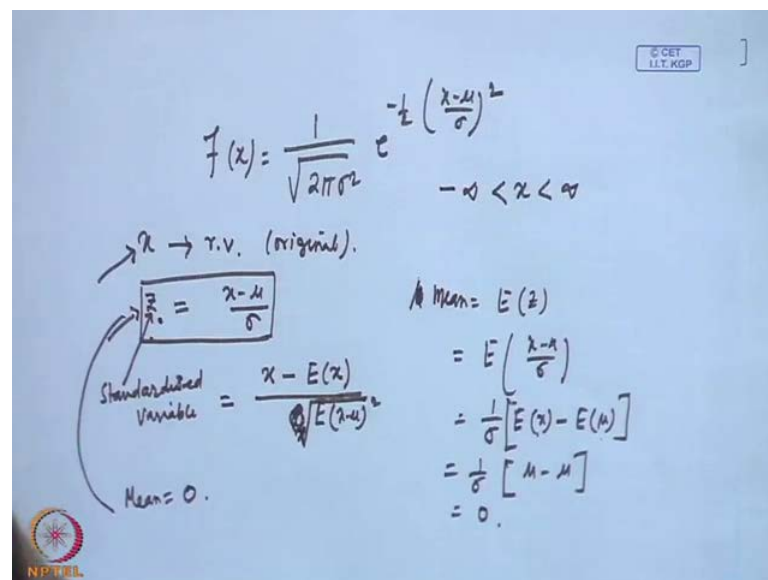
So, each parameter will have a distribution. How do I know that what distribution it follows? If you do not understand this you will face problem there. So, this is a I can say that one of the most fundamental concept and it should be, and you should be thorough about this concept. Now, let us concentrate on what is z distribution.

(Refer Slide Time: 12:58)



I am sure that you will not face it a difficult one, because all of you know that this left hand side, the figure of you see this is the normal distribution. The probability density function is $f(x)$ equal to 1 by...

(Refer Slide Time: 13:20)



I am writing here, once more $f(x)$ equal to one by root over 2 pi sigma square e to the power minus half x minus mu by sigma square. And x varies from minus infinite to plus infinite. Now, where x is a random variable this is the original one, original random variable, original one what you have observed, correct? Now, let us transform x in this

manner. Let z is a transform variable of x , which is x minus μ by σ . So, you have observed x , you are creating another variable z which we will be using, x as well as the population parameter like μ and σ . And this is this is what is known as the transform the standardized variable. This is known as standardized variable.

So, when a variable is subtracted by its mean and divided and the resultant quantity is divided by the standard deviation, that is known as standardized variable. So, that mean by standardized variable, we mean that the variable minus its expected value, that is a mean divided by that what I can say σ means. Basically, x minus μ expected value of this square root. You mean this square root of this, because we are considering standard deviation square root of this, correct? Now, what will be the mean value of z 0? How every guess, basically you can write in this manner. Suppose, what is mean? Mean is suppose if I write mean.

Mean is expected value of variable z . Here we are talking about μ_z so that mean expected value of x minus μ by σ , that mean 1 by σ expected value of x minus expected value of μ . And c expected value of constant is constant. So, expected value of x is μ minus μ , which is equal to 0 . So, that is why the z is a random variable whose mean value is 0 . What will happen to its standard deviation? You will get 1 in the same fashion.

(Refer Slide Time: 16:35)

$$\begin{aligned}
 V(z) &= E[(z - E(z))^2] \\
 &= E(z^2) \\
 V(x) &= \sigma^2 \\
 V(ax) &= a^2 \sigma^2 \\
 &= E\left[\left(\frac{x - \mu}{\sigma}\right)^2\right] \\
 &= \frac{1}{\sigma^2} \left[E(x - \mu)^2 \right] \\
 (0, 1) &= \frac{\sigma^2}{\sigma^2} = 1
 \end{aligned}$$

You can find out that, what is the standard deviation? Standard deviation means expected value, that is the variance. You find out first variance of z which is expected value of your z minus expected value of $E z$ that whole square. So, expected value of z already you got 0. So, this basically expected value of z square. So, it is expected value of z is nothing but x minus μ by σ square.

So, if the variance of x , suppose variance of x is σ square. The variance of $a x$, where a is constant, it will be $a^2 \sigma$ square. So, it will be σ square. So, that means you can write this one, 1 by σ square into expected value of x minus μ square. What is this expected value of x minus μ square? That is the σ square. So, your σ square by σ square, this is 1 . If this is the case, now you put you see that in this equation. Come back to this slide again, what we have put? We put z equal to x minus μ by σ square, our resultant equation. The probability density function for z which is 1 by root over $2 \pi \sigma$ square is 1 .

So, it is 2π into 1 , means 2π square root e to the power 1 by $2 z$ square. So, this is the conversion of any variable. For example, this normal variable to its unit normal distribution. So, z distribution is also known as unit normal distribution, because its mean value is 0 and standard deviation is 1 . What is the use? Why should we convert to unit normal distribution? The reason is the even if there are many variables, but once you standardized those any variables it will be a unit normal. So, you require only one normal distribution unit, normal distribution table and using that you are able to, what I can say use that table to different situation, even though the variable mean and standard deviation differs.

(Refer Slide Time: 19:33)

An example

Sl. No.	Months	Profit in Rs million	Sales volume in 1000	Absenteeism in %	Machine breakdown in hours	M-Ratio
1	April	10	100	9	62	1
2	May	12	110	8	58	1.3
3	June	11	105	7	64	1.2
4	July	9	94	14	60	0.8
5	Aug	9	95	12	63	0.8
6	Sep	10	99	10	57	0.9
7	Oct	11	104	7	55	1
8	Nov	12	108	4	56	1.2
9	Dec	11	105	6	59	1.1
10	Jan	10	98	5	61	1.0
11	Feb	11	105	7	57	1.2
	March	12	110	6	60	1.2

© Dr J Maiti, IEM, IIT Kharagpur

So, this our example. And if I consider profit, we are considering the profit and showing you that.

(Refer Slide Time: 19:40)

Example – use of unit normal distribution

Assume the variable profit per month is normally distributed with mean of Rs 11 millions and standard deviation of Rs 1.5 millions. What is the probability that the profit per month will be delivered within Rs 12.5 millions?

$$\begin{aligned}
 p = (x \leq 12.5) &= p\left(\frac{x - \mu}{\sigma} \leq \frac{12.5 - 11}{1.5}\right) \\
 &= p\left(z \leq \frac{12.5 - 11}{1.5}\right) = p(z \leq 1) = 0.8413
 \end{aligned}$$

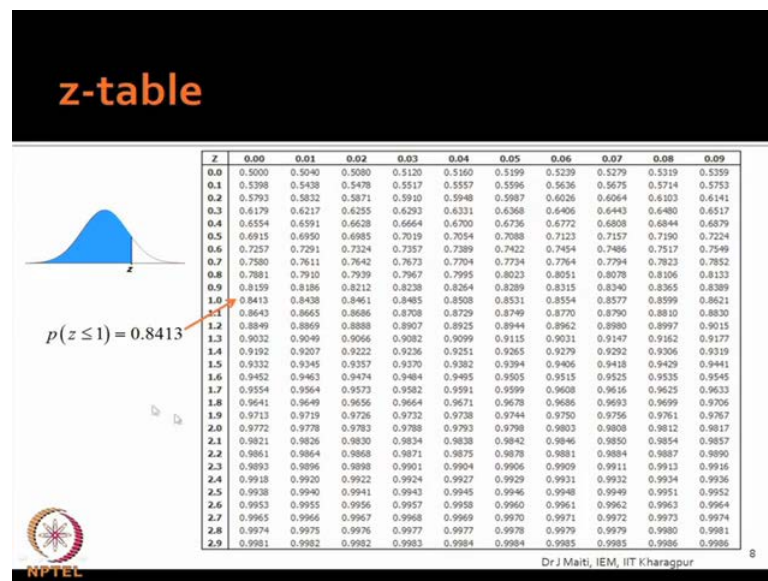
Dr J Maiti, IEM, IIT Kharagpur

What is the use of standard normal distribution here? Assume the variable profit per month is normally distributed with mean of rupees 11 millions and standard deviation of rupees 1.5 millions. What is the probability that the profit per month will be within 12.5 millions? How to go about it? You see the left hand side this figure, it is in the original variable. Right hand side is the unit normal one. Now, if you see this x axis in the bottom

one, you see what is the mean is the middle value. And every other one standard deviation, two standard deviation, three standard deviation, both side that demarcation is there.

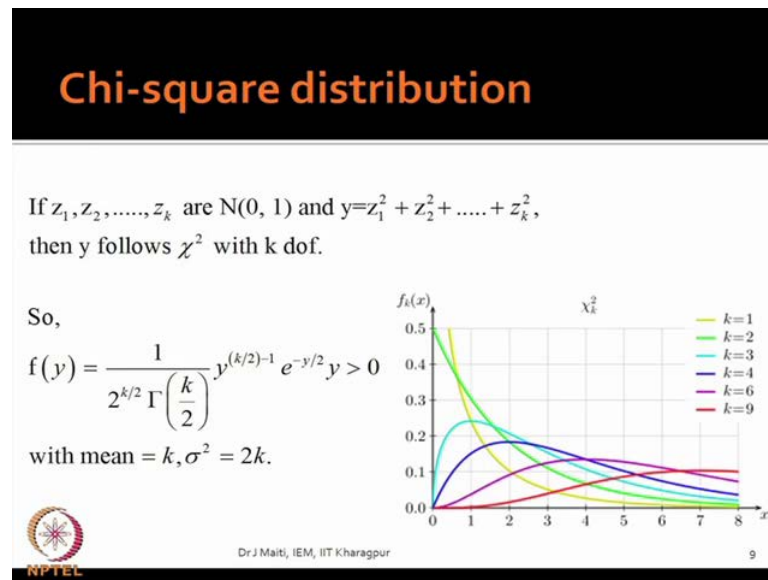
Now, as sigma equal to 1 that same will be now mean will be 0 and 1 into 1, that will be 1 that minus 1, minus 2, minus 3 like this. So, this line is what is the z line, because you are interested to know that your probability of profit, less than equal to 12.5 in rupees million. And if I convert into z value, it is coming 1. So, probability z less than equal to 1 that mean this 1, this is the z equal to 1. The left hand side values probability equal z less than equal to 1 will be the area under the normal distribution curve from minus infinity to that z value, because that is within this. Now, you have standard normal table. Now, you will, what you will do once you get the z value? You go for the table. So, our z value is 1.

(Refer Slide Time: 21:37)



You see that this is the shaded portion is the probability. This area is the probability, whose value is 0.8413. So, you are able to find out the probability that your profit will be within this.

(Refer Slide Time: 22:02)



Now, second distribution, sample distribution is chi square. Do have any idea that when you use chi square distribution you find out that if you go through the standard book like, very good book like Johnson and Richard book. You will find that everywhere may be when you talk about the constant there all statistical distribution, the chi square distribution. But why chi square distribution is used instead of t distribution, or instead of z distribution or instead of f distribution or instead of any other distribution. What is the basis that mean? We must know that, what is chi square distribution? How it is generated and when it will be used.

Now, you see this slide here what we are seeing that, if z_1, z_2, z_k , what is z ? z is normal distribution unit normal. So, you have collected suppose, k observations and those k unit that normal observations are that z_1, z_2, z_k and you are creating one variable, which is the sum of the normal variable, unit normal variable square. In what I mean to say here, then suppose you have collected n k data point 1, 2, 3 like k .

(Refer Slide Time: 23:34)

The slide contains handwritten notes on a light blue background. At the top right, there is a small box with the text '© CET IIT KGP'. In the bottom left corner, there is a logo for NPTEL. The main content is a table and several formulas:

i	x_i	$z = \frac{x_i - \mu}{\sigma}$	z^2
1	x_1	z_1	z_1^2
2	x_2	z_2	z_2^2
3	x_3	z_3	z_3^2
...
k	x_k	z_k	z_k^2

Below the table, there is a circled expression: $\sum_{i=1}^k z_i^2$. To the right of the table, there are several formulas and annotations:

- A circled formula: $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$. An arrow points from this formula to the word 'Variance'.
- A formula: $\bar{x} = \frac{1}{n} \sum x_i$. An arrow points from this formula to the word 'Mean'.
- A formula: $\hat{\sigma} = \frac{\bar{x} - E(\bar{x})}{\sqrt{V(\bar{x})}}$.

And you have x values for x_1, x_2, x_3 . Let it be x_k and you know z values, z equal to x minus μ divided by σ . Assuming that μ and σ are population parameter, then you are getting z_1, z_2, z_3 like z_k . Now, you are making z square, z_1 square, z_2 square, z_3 square, like this z_k square. Now, if you take a sum of these z_i , i equal to 1 to k , you get a quantity. This is also a statistic, this is the linear sum of the sum of the square of the variable values of the normal unit normal variable. These quantity that you have created, y as shown here it is y and any how you can change it to y , no problem.

So, these quantity follows chi square distribution, with how many degrees of freedom? k degrees of freedom. So, what is the essential learning? Here, our learning is suppose, I know that there is no, I know that there is a normal variable x , your collected data on it, you converted it to standard normal, each of the observations is squared and you have taken a sum and that sum you used for different purposes. That sum we will follow certain distribution, that distribution is chi square distribution. Remember, this one.

Suppose, usually why normal distribution will start with normal distribution, but normal distribution plenty of things in the real world, most of the things can be converted to normally distributes. Most of the cases that is the starting point. Now, for the purpose of your analysis, purpose your model building. What is the purpose for which you want to use it there? What is required? You require the sum of the square of the variables values, then what we will do when you collect one sample? You have to have the distribution of

that, that is chi square distribution. Any question? Understood fully or not? Any question?

The general form of chi square distribution is like this, that $f(y) = \frac{1}{2^{k/2} \Gamma(k/2)} y^{k/2-1} e^{-y/2}$. And this is the case and the mean value of chi square is the k , which is degrees of freedom and variance is 2 times degrees of freedom. And this figure you see that, this is the probability density function for a chi square variable. Now, that the x is chi square, then what is happening here? 1 to 8 these are the values and ultimately you will be getting different shape of chi square density function.

When your k equal to 1 this is this as well as k equal to 2, it looks like exponential distribution, but slowly that shape will change. So, degrees of freedom plays an important role in chi square distribution. In z distribution, what is the degree of freedom required? No degrees of freedom. We have not discussed anything related to degrees of freedom in z distribution. When the unit normal distribution you table, you see there is no degrees of freedom column. So, that mean in z distribution it is basically not affected by the degrees of freedom available with the data set chi square distribution. When you talk about chi square distribution, please keep in mind the degrees of freedom is coming into consideration.

And chi square is nothing but the normal square. My x is normally distributed, I am taking the square linear square, linear sum of square of x , that is my chi square distribution. So, you may be wondering that where is the use? Now, see I told you we want to find out the distribution of sample statistic, getting me? Now, one of the sample statistic is a square variance, yes or no? Very much, what is the distribution of a square? How do we know what is the, what will be distribution of \bar{x} ? See, \bar{x} is $\frac{1}{n} \sum x_i$ normally distributed variable. So, \bar{x} if x is normally distributed, \bar{x} also follow normal distribution, by considering \bar{x} minus expected value of \bar{x} by variance of \bar{x} .

Then uninormal distribution I will not tell anything related to this computation. Later on I will tell you what is this from central limit theorem. You know what will be the distribution, but irrespective of what I mean to say here, that irrespective of the value that \bar{x} value, if you collect data from a normal distribution and \bar{x} will be normally distributed, getting me? Now, what will be the A square? How do what is A square?

A square is 1 by n minus 1 sum total of n equal to 1 to n x i minus x bar square. Can you find out any similarity here? What you have done, I say x is normally distributed, x bar is also normally distributed then you have made the squaring this. So, when normal variable is squared and you take summation and do little bit of manipulation using the population variance, what you will get? You will get standard normal z and summation of square of standard normal z. Is it not correct?


(Refer Slide Time: 30:53)

Example- use of chi-square distribution

$$\frac{(n-1)s^2}{\sigma^2} = \left[\left(\frac{x_1 - \bar{x}}{\sigma} \right)^2 + \left(\frac{x_2 - \bar{x}}{\sigma} \right)^2 + \dots + \left(\frac{x_n - \bar{x}}{\sigma} \right)^2 \right]$$

$$= z_1^2 + z_2^2 + \dots + z_n^2 \sim \chi_{n-1}^2$$

For the example data find out the distribution of the profit variance obtained through the 12 months data assuming population variance is 1.5.



Dr J Maiti, IEM, IIT Kharagpur

20

You see this slide, what I have shown here? A square is just it is nothing but the formula we have given $X_1 - \bar{x}$ square plus $X_2 - \bar{x}$ square plus $X_3 - \bar{x}$ square and sigma square is divided. This sigma is the population variance. Now, this quantity is coming like Z_1 square plus Z_2 square plus Z_n square. So, it is it is the sum total of normal squares, the variable value square. So, it is chi square distribution, that is why when we will go for interval estimation of variance we will use chi square distribution, because it should come to your mind that why x bar, for x bar we are using normal distribution, but for s square we are using chi square distribution.

That mean, the crux of the matter is here, this is the development, this is the development, fantastic. Then you may be wondering why the n minus 1? Already we have seen that while calculating the A square, one degree is lost, then same thing continues and ultimately our, this n minus 1 A square by sigma square, this quantity follows chi square distribution with n minus one degrees of freedom. Any question for

this? So, it is any question? No question? I think it is obvious now. So, keep in mind this one because many a times you will be using this type of derived units. You will square and then add and chi square is required, but you do not know what distribution you will be using, just follow this concept. And you follow chi square, some other case it will be other distribution.

This is the use for example, what the if we consider data then profit variance obtained through the twelve month data, assuming population variance is you can find out the distribution of the variance component computed from the twelve months data. It is all the uses of this chi square distribution will be revealed in subsequent lectures, but the concept remains same when we will use chi square distribution, this is the concept, okay?

(Refer Slide Time: 31:56)

Chi-square table

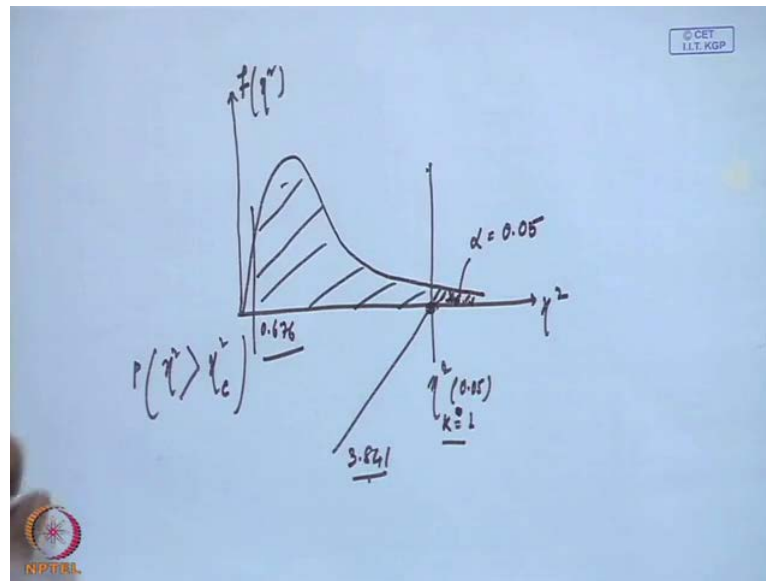
Critical Values of the χ^2 Distribution

df \ p	0.995	0.975	0.9	0.5	0.1	0.05	0.025	0.01	0.005	df
1	.000	.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879	1
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597	2
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838	3
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860	4
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750	5
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449	16.812	18.548	6
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278	7
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955	8
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589	9
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188	10
11	2.603	3.816	5.578	10.341	17.275	19.675	21.920	24.725	26.757	11
12	3.074	4.404	6.304	11.340	18.549	21.026	23.337	26.217	28.300	12
13	3.565	5.009	7.042	12.340	19.812	22.362	24.736	27.688	29.819	13
14	4.075	5.629	7.790	13.339	21.064	23.685	26.119	29.141	31.319	14
15	4.601	6.262	8.547	14.339	22.307	24.996	27.488	30.578	32.801	15

Dr J Maiti, IEM, IIT Kharagpur

Chi square table, when you have gone for Z table please keep in mind that in z table there is no degrees of freedom column. Chi square means there will be degrees of freedom. So, there the first column itself is degrees of freedom, 1 to 15 it will be, it will be infinite, because chi square value can go up to that level. Then there are different probabilities values here and for different probability values what will be your chi square value. So, how to use this table? Any idea?

(Refer Slide Time: 34:28)



For example, let the chi square distribution PDF is like this chi square, this is function of that chi square PDF. Let the distribution look like this. Now, I want to know what is the chi square value for the probability of this side? Let it be alpha which is 0.5, from this table can u find out this value, chi square value, what will be this chi square? 0.05 understand?

Because these things you will be requiring later on. You have to see this chi square table frequently. What I mean to say, I know the probability right hand side probability here, for which I want to know what will be the chi square value. If you are giving like this, chi square with probability 0.05, you cannot calculate. You cannot find out the value from here which degrees of freedom is it, 1 degree of freedom or 15 degrees of freedom or 120 degrees of freedom.

So, another quantity should be here which is degree of freedom. So, suppose k equal to 1 then what is this value? k equal to 1, probability is 0.05. So, your chi square value is 3.841. So, my this value is 3.841. Suppose, you require a value where the probability, so what is this? This is this one is probability, that chi square value is this chi square. This probability that chi square value this value, this value that will be greater than some value, getting me?

So, let any value we have, we had basically k equal to 1 some value this is that value. We will say that chi square computed value probability that this will be chi square. Chi

square is computed, this value this will be greater than this is the value, this side less than this will be other side. Now, if your degree of freedom is 6 and you want the probability that the value is that 9.995, then your value chi square value will be 0.676 which will be somewhere here.

(Refer Slide Time: 37:50)

t-distribution

If z and χ_k^2 are independent $N(0, 1)$ and chi-square variables, respectively, then the random variable


$$t_k = \frac{z}{\sqrt{\chi_k^2 / k}}$$

follows t distribution with k dof. The pdf of t is

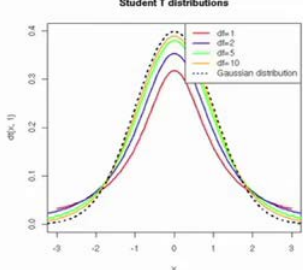
$$f(t) = \frac{\Gamma\left[\frac{k+1}{2}\right]}{\sqrt{k\pi} \Gamma\left(\frac{k}{2}\right)} \frac{1}{\left[\left(t^2/k\right)+1\right]^{\frac{k+1}{2}}},$$

$-\infty < t < \infty$

with mean $= 0, \sigma^2 = \frac{k}{k-2}, k > 2$.

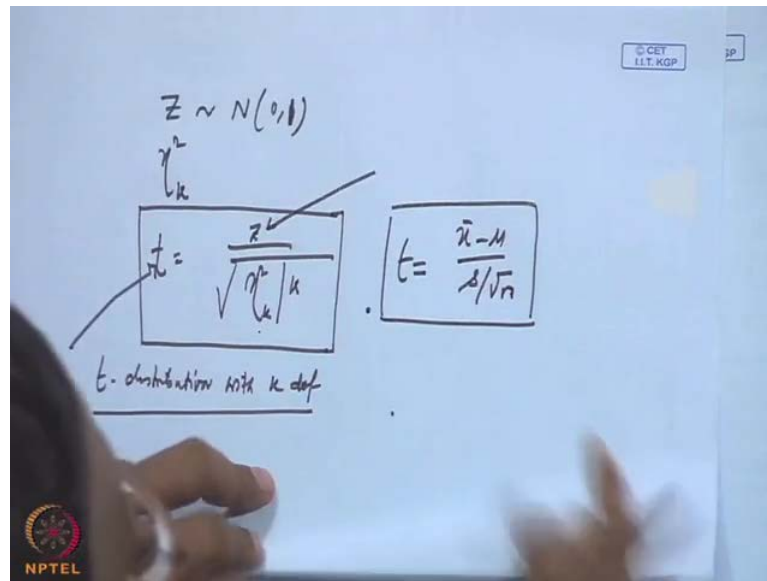


Dr J Malvi, IEM, IIT Kharagpur



So, you are talking about the total probability this side. Now, let us see t distribution, t distribution, when will you use t distribution? First you have to understand when we will be using t distribution, then we will see the uses of t distribution. Then again I will show you the table, how to use table to find the different critical value for t distribution. You see here, if z and x_k are independent normal and chi square variables.

(Refer Slide Time: 38:29)



So, we are now considering two variables, one is a normal variable that is unit normal that is z , another one we are talking about chi square. Let it be chi square with k degrees of freedom, fine? Now, you create another variable which is t , which is x by square root of chi square k divided by k . Now, in your development process suppose when you are developing a model, if you find that you have created some variable which is of this form, that normal by square root of chi square divided by its degrees of freedom. Then this quantity that t will follow t distribution with k degrees of freedom. t distribution with k degrees of freedom, that means t distribution case also degree of freedom will come.

Degrees of freedom is very important for chi square distribution, for t distribution also. And you see it is the same of t distribution, it is similar to normal distribution. When your this side that degree of freedom will be infinite, that is it exactly matches with normal distribution. What is the mean value of t distribution? You are not getting? See z what is the mean value of z 0 mean, that t value divided by tangent 0 is coming. So, what is happening here? The t distribution mean value is 0 and standard deviation value, if square root of k by k minus 2 and k must be greater than equal to 2 and you see that this one here in the diagram itself, the mean value is 0.

(Refer Slide Time: 40:56)

Example- use of t distribution

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} = \frac{\bar{x} - \mu}{\sqrt{s^2 / n}} = \frac{\sqrt{n}(\bar{x} - \mu)}{\sqrt{s^2}}$$


Follows z distribution

$$= \frac{(\bar{x} - \mu) / \sqrt{(\sigma^2 / n)}}{\sqrt{\frac{(n-1)s^2}{\sigma^2} \cdot \frac{1}{(n-1)}}}$$

So, the resultant quantity is

$$t = \frac{z}{\sqrt{\chi_{n-1}^2 / n - 1}}$$

Follows chi-square distribution with n-1 dof



Dr J Maiti, IEM, IIT Kharagpur

13

What is each? What is its use? Which one this? There you basically it speaks what is the parameter of t distribution. What is the parameter of normal distribution, mu and sigma square. What is the parameter of t distribution, that is the case. What is here? You have see everywhere k is 3. You see the function gamma k plus 1 by 2 k. Here k, everywhere k is there. So, that means the parameter of this distribution, t distribution k degrees of freedom and you are getting, that is why mean is 0 and sigma square in terms of k. Let it be.

So, you will know the parameter from the PDF only in t distribution, PDF all the all other values are constant, t defines the random variables. Otherwise, for other things you see that pi is the constant for and that is that. By other way only k is there everywhere, gamma distribution is beginning, getting me? Now, come back to this use. What is the use of when do you use t distribution? Now, let us see that the t is like this, that x bar minus mu by s by square root of n. You have created this type of composition. How it will come or how it will, why it will come?

(Refer Slide Time: 43:11)

A hand is writing on a whiteboard. The text on the board is as follows:

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum x_i \\ E(\bar{x}) &= E\left[\frac{1}{n} \sum x_i\right] \\ &= \frac{1}{n} [E(x_1) + E(x_2) + \dots + E(x_n)] \\ &= \frac{1}{n} [\mu + \mu + \dots + \mu] \\ &= \frac{n\mu}{n} = \mu\end{aligned}$$

There is a small logo in the top right corner that says "© CET I.I.T. KGP" and an NPTEL logo in the bottom left corner.

How it will come? I am showing you one thing that come. I am showing you one thing that when we talk about the distribution of \bar{x} . So, \bar{x} is $\frac{1}{n}$ sum total of x_i , then I ask you what is the expected value of \bar{x} . You said that expected value of $\frac{1}{n}$ sum total of x_i . So, that mean $\frac{1}{n}$ expected value of x_1 plus expected value of x_2 plus expected value of x_n . What is the expected value of x_1 ? All are μ , if x is normally distributed every observation will be correspondingly distributed to the expected, respective mean. So, it is basically $n\mu$ by n , that is μ . So, we say that expected value of sample average is μ . Now, what will be the variance component?

(Refer Slide Time: 44:13)

A hand is writing on a whiteboard. The text on the board is as follows:

$$\begin{aligned}V(\bar{x}) &= V\left[\frac{1}{n} \sum x_i\right] \\ &= \frac{1}{n^2} [V(x_1) + V(x_2) + \dots + V(x_n)] \\ &= \frac{1}{n^2} [\sigma^2 + \sigma^2 + \dots + \sigma^2] \\ &= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}\end{aligned}$$

Below this, the distribution is given as:

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \rightarrow z = \frac{\bar{x} - E(\bar{x})}{\sqrt{V(\bar{x})}}$$

Then, the standard normal variable is defined as:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}}$$

There is a small logo in the top right corner that says "© CET I.I.T. KGP" and an NPTEL logo in the bottom left corner.

If I want to know variance of \bar{x} , that means this is nothing but variance of $\frac{1}{n} \sum_{i=1}^n x_i$. So, as I said that variance if you add $\frac{1}{n^2}$ then variance of x_1 plus variance of x_2 plus variance of x_n . x_1, x_2, x_n , all are normal random variables with variance σ^2 . So, $\frac{1}{n^2}$ then σ^2 plus σ^2 , like plus σ^2 which will $n \sigma^2$ by n^2 , that is $\frac{\sigma^2}{n}$.

So, that means using that if I create it, create a that we will \bar{x} , I told u that if normally distributed with mean of these. Mean will be μ and $\frac{\sigma^2}{n}$. Now, you are creating a z variable here. If you create a z variable here, so z is nothing but \bar{x} . This is the normal variable minus its expected value of \bar{x} divided by square root of variance of \bar{x} . If any variable random variable, any variable is subtracted by its mean and divided by the standard deviation, it is z standard value.

So, what is this? \bar{x} minus expected value of \bar{x} is μ . We have already proved, we have already proved and your variance is $\frac{\sigma^2}{n}$. So, it is $\frac{\sigma^2}{n}$ by variance is $\frac{\sigma^2}{n}$. This is the quantity or other way I can say z is \bar{x} minus μ by σ by \sqrt{n} . So, this will follow z distribution.

(Refer Slide Time: 46:29)

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Sample Std. dev.
Population Std. dev.

Now, here in t case what is happening here it is t case you will see that what we are saying, suppose \bar{x} minus μ is there, but σ is not there known. You are using s in z case \bar{x} minus μ by σ by \sqrt{n} . This σ is population standard deviation, but this is not known. Instead you are using sample standard deviation in the t

case. So, s is a random variable, but here σ is a constant. Now, depending on the situation ultimately what level there are different conditions.

So, if sample size is very high large then the same this quantity can be in unit normal also, but most general case is this quantity is t . Why, because this quantity follows t distribution. How can we justify that this quantity follows t distribution? If you want to justify this then I just written this one, the same thing I have written in this manner, \bar{x} minus μ divided by σ square by n . And then the square within square root component is n minus 1 square by σ square 1 by n minus 1, just manipulation. What is that the numerator and denominator is manipulated with some constants, then what is this \bar{x} minus μ by σ square by n . This is z , already we have seen this is z top portion. What is the bottom portion?

Earlier I have shown you that n minus 1 s square by σ square follow chi square distribution. You go back you see what is this one, chi square distribution. So, your resultant variable is that z value, z random by the square root of chi square by its degrees of freedom. So, it is t distribution that is the use. Why you will use t distribution in this case? So, you see this formula and appreciate it, because if you understand this huge problem will be solved. That is the use, follows chi square this one. So, the resultant quantity is this. Now, you require to use chi square table.

(Refer Slide Time: 49:06)

t-table

Student t-Table

Alpha	0.250	0.200	0.150	0.100	0.050	0.025	0.010	0.005	0.0005
df									
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.656	636.578
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.600
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.526	2.845	3.850

Dr. Manoj K. J. Nair, IIT Kharagpur

NPTEL

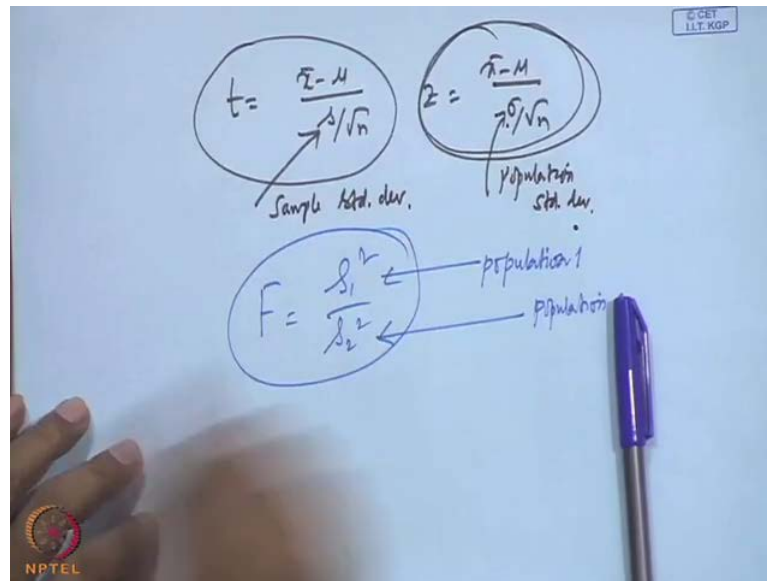
14

This is our chi square table. So, please keep in mind in chi square table also there will be degrees of freedom. Degrees of freedom coming into consideration, because the parameter is degrees of freedom, because the parameter of the distribution is degrees of freedom. I think you will be able to find out, suppose if I say that if my t distribution with 11 degrees of freedom and what is the probability for that means that 0.0025. If we consider then you will be getting a value of 2.228, but if you see the z distribution for the same thing for 0 to 5 there is no need of any degrees of freedom. 1.96 will be the z value.

Then come to f distribution, please know that is developed by that I think I just and the it is developed while he was a student so far. I know this one student that is distribution is here sample variance, but population variance is still there that $n - 1$ square by sigma square. But question is that when we compute the t there is no population variance. When you compute t there is no population variance, s is there. There everything can be calculated.

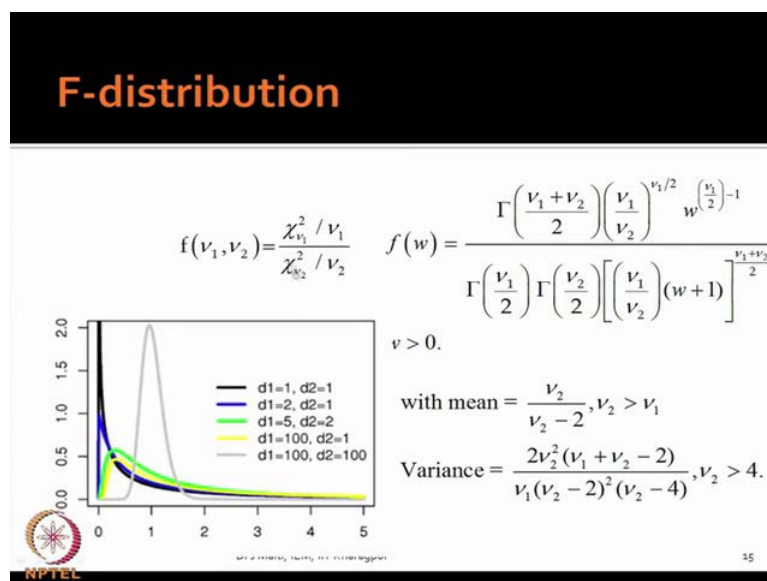
Now, I want to know what is the distribution of this and we found out that the distribution is t distribution, f distribution. Now, come to a ratio measure, what I am trying to we will see here. Suppose, you have two population for the same variable, population 1 and population 2 are characterized by the same variable, and you have computed. You have taken sample, from both the sample you have computed your standard deviation or variance for population 1 as well as population 2. You want to compare whether variability in population 1 is different from the variability in population 2 or not or they are basically equal.

(Refer Slide Time: 51:56)



That mean you are basically creating a variable, which is suppose s 1 square by s 2 square. This is coming from population 1, this is coming from population 2, whether two population are having same variance or not in many models, we assume that the population variance are equal. For example, when you do anova even in manova case also we will be seeing that one of the condition for sometimes. We use that population variances are equal, this ratio we want to require to know in anova. You will be finding out the use of this anova, must use of these two regression also is the same.

(Refer Slide Time: 52:51)




So, the ratio when you are finding some quantity, you are finding out which is basically the ratio of two chi square variable divided by their respective degrees of freedom. Then the quantity basically this is what I meant to say that if this is the w, this w if we create like this the quantity follows f distribution with two degrees of freedom. Numerator degrees of freedom and denominator degrees of freedom, why you see the distribution here mu 1 and mu 2, everywhere mu 1, mu 2 is there. Rest of gamma function it is in other cases. So, getting me?

So, this distribution is characterized by numerator degrees of freedom and denominator degrees of freedom, and when do you use f distribution? When you derive any quantity, which is the ratio of two chi square variables and definitely that ratio weighted ratio. This ratio means weighted ratio of two chi square variable, where weight is nothing but the degrees of freedom 1 by that degrees of freedom.

(Refer Slide Time: 54:19)

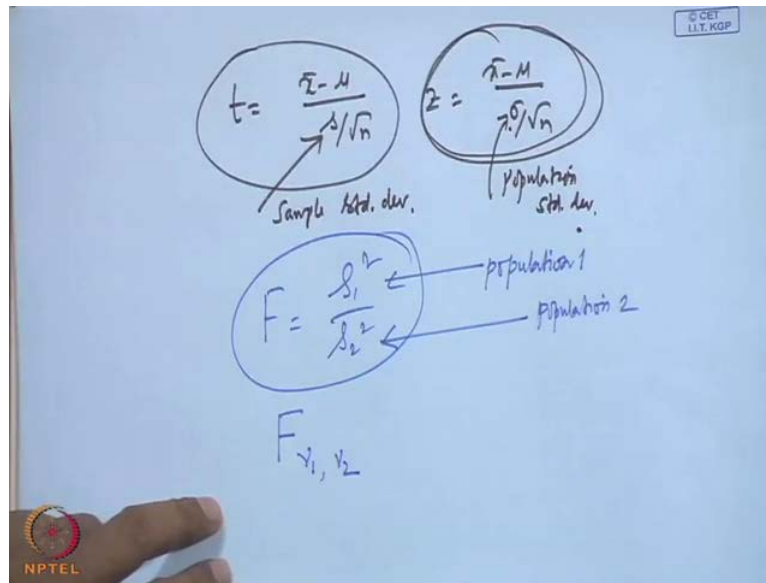
Example- use of F distribution

$$\frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2} = \frac{\chi_{n_1-1}^2(y) / n_1 - 1}{\chi_{n_2-1}^2(y) / n_2 - 1} = w \sim F_{n_1-1, n_2-1}$$


Dr J Maiti, IEM, IIT Kharagpur
16

You are getting me? For example, what I say let us see the use see s 1 square, s 2 square from two population. That population variance is sigma square, sigma 2 square I can write this like this, because all of this we have already seen that n minus 1 is s 1 square by s n minus 1 square by sigma square follows chi square distribution. So, that means this quantity will be chi square divided by degrees of freedom, the denominator point it will be again be chi square divided by the respective degrees of freedom, correct? So, this is w and this one is f distributed.

(Refer Slide Time: 55:01)



So, in f distribution please keep in mind when we talk about f distribution, there will be numerator degrees of freedom, denominator degrees of freedom.

(Refer Slide Time: 55:11)

F-table

		Numerator Degrees of Freedom															Significance Level		
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
Denominator Degrees of Freedom	0.025	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
	1	647.763	750.482	804.151	856.560	907.895	957.114	984.333	995.643	998.279	999.834	999.725	999.874	999.981	999.999	1000.000	1000.000	1000.000	1000.000
	2	38.509	39.000	39.186	39.248	39.289	39.311	39.326	39.337	39.347	39.356	39.364	39.371	39.378	39.384	39.389	39.393	39.396	39.398
	3	17.442	18.044	18.430	18.701	18.885	18.995	19.044	19.064	19.076	19.084	19.090	19.094	19.097	19.099	19.101	19.102	19.103	19.104
	4	12.217	12.840	13.142	13.346	13.484	13.564	13.604	13.624	13.636	13.644	13.649	13.652	13.654	13.655	13.656	13.657	13.658	13.658
	5	10.009	10.632	10.934	11.138	11.276	11.356	11.396	11.416	11.428	11.436	11.440	11.442	11.443	11.444	11.444	11.445	11.445	11.445
	6	8.813	9.436	9.738	9.942	10.080	10.160	10.199	10.219	10.231	10.239	10.243	10.245	10.246	10.247	10.247	10.248	10.248	10.248
	7	8.027	8.650	8.952	9.156	9.294	9.374	9.413	9.433	9.445	9.453	9.457	9.459	9.460	9.461	9.461	9.462	9.462	9.462
	8	7.570	8.193	8.495	8.699	8.837	8.917	8.956	8.976	8.988	8.996	8.999	9.001	9.002	9.003	9.003	9.004	9.004	9.004
	9	7.203	7.826	8.128	8.332	8.470	8.550	8.589	8.609	8.621	8.629	8.633	8.635	8.636	8.637	8.637	8.638	8.638	8.638
10	6.937	7.560	7.862	8.066	8.204	8.284	8.323	8.343	8.355	8.363	8.367	8.369	8.370	8.371	8.371	8.372	8.372	8.372	
11	6.724	7.347	7.649	7.853	8.000	8.080	8.119	8.139	8.151	8.159	8.163	8.165	8.166	8.167	8.167	8.168	8.168	8.168	
12	6.558	7.181	7.483	7.687	7.834	7.914	7.953	7.973	7.985	7.993	7.997	7.999	8.000	8.001	8.001	8.002	8.002	8.002	
13	6.414	7.037	7.339	7.543	7.690	7.770	7.809	7.829	7.841	7.849	7.853	7.855	7.856	7.857	7.857	7.858	7.858	7.858	
14	6.297	6.920	7.222	7.426	7.573	7.653	7.692	7.712	7.724	7.732	7.736	7.738	7.739	7.740	7.740	7.741	7.741	7.741	
15	6.196	6.819	7.121	7.325	7.472	7.552	7.591	7.611	7.623	7.631	7.635	7.637	7.638	7.639	7.639	7.640	7.640	7.640	
16	6.115	6.738	7.040	7.244	7.391	7.471	7.510	7.530	7.542	7.550	7.554	7.556	7.557	7.558	7.558	7.559	7.559	7.559	
17	6.040	6.663	6.965	7.169	7.316	7.396	7.435	7.455	7.467	7.475	7.479	7.481	7.482	7.483	7.483	7.484	7.484	7.484	
18	5.971	6.594	6.896	7.099	7.246	7.326	7.365	7.385	7.397	7.405	7.409	7.411	7.412	7.413	7.413	7.414	7.414	7.414	
19	5.916	6.539	6.841	7.044	7.191	7.271	7.310	7.330	7.342	7.350	7.354	7.356	7.357	7.358	7.358	7.359	7.359	7.359	
20	5.871	6.494	6.796	6.999	7.146	7.226	7.265	7.285	7.297	7.305	7.309	7.311	7.312	7.313	7.313	7.314	7.314	7.314	
22	5.800	6.423	6.725	6.928	7.075	7.155	7.194	7.214	7.226	7.234	7.238	7.240	7.241	7.242	7.242	7.243	7.243	7.243	
24	5.746	6.369	6.671	6.874	7.021	7.101	7.140	7.160	7.172	7.180	7.184	7.186	7.187	7.188	7.188	7.189	7.189	7.189	
26	5.700	6.323	6.625	6.828	6.975	7.055	7.094	7.114	7.126	7.134	7.138	7.140	7.141	7.142	7.142	7.143	7.143	7.143	
28	5.661	6.284	6.586	6.789	6.936	7.016	7.055	7.075	7.087	7.095	7.099	7.101	7.102	7.103	7.103	7.104	7.104	7.104	
30	5.627	6.250	6.552	6.755	6.902	6.982	7.021	7.041	7.053	7.061	7.065	7.067	7.068	7.069	7.069	7.070	7.070	7.070	
40	5.529	6.152	6.454	6.657	6.804	6.884	6.923	6.943	6.955	6.963	6.967	6.969	6.970	6.971	6.971	6.972	6.972	6.972	
60	5.439	6.062	6.364	6.567	6.714	6.794	6.833	6.853	6.865	6.873	6.877	6.879	6.880	6.881	6.881	6.882	6.882	6.882	
120	5.352	5.975	6.277	6.480	6.627	6.707	6.746	6.766	6.778	6.786	6.789	6.791	6.792	6.793	6.793	6.794	6.794	6.794	
∞	5.278	5.901	6.203	6.406	6.553	6.633	6.672	6.692	6.704	6.712	6.716	6.718	6.719	6.720	6.720	6.721	6.721	6.721	

Dr J Malvi, IEM, IIT Kharagpur

So, when you go to see the f distribution table, you have to see that two distribution, two different degrees of freedom. For example, if you are interested to know for 5 numerator degrees of freedom and 3, that denominator degrees of freedom with a probability 0.025, then you find out this

value. You will be getting and this value 0.7, that is 7.76. So, this way it will be used and central limit theorem is the final one for our.


(Refer Slide Time: 55:46)

Central limit theorem (CLT)

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

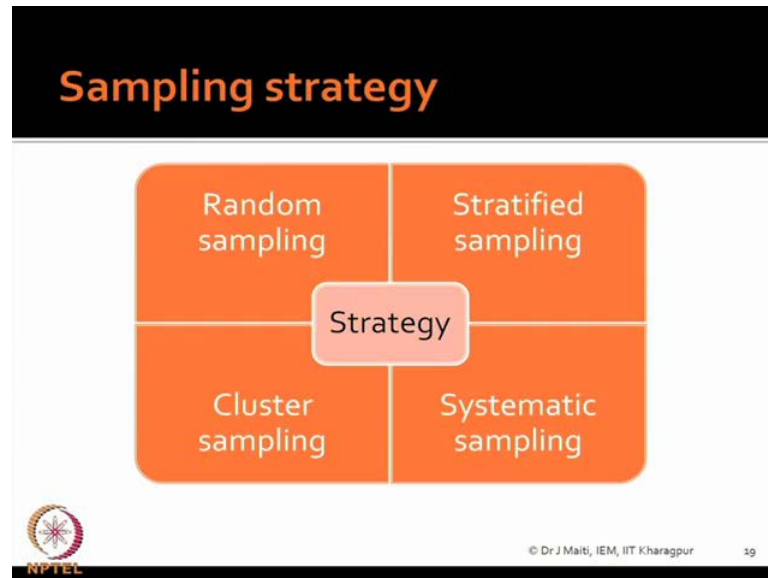
$$\begin{aligned}
 E(\bar{x}) &= E\left(\sum_{i=1}^n \frac{x_i}{n}\right) = \frac{1}{n} E(x_1 + x_2 + \dots + x_n) \\
 &= \frac{1}{n} [E(x_1) + E(x_2) + \dots + E(x_n)] \\
 &= \frac{1}{n} [\mu + \mu + \dots + \mu] \\
 &= \frac{n\mu}{n} \\
 &= \mu
 \end{aligned}$$

$$\begin{aligned}
 \sigma_{\bar{x}}^2 &= V(\bar{x}) = V\left[\frac{1}{n} \sum_{i=1}^n x_i\right] \\
 &= V\left[\frac{1}{n} \{x_1 + x_2 + \dots + x_n\}\right] \\
 &= \frac{1}{n^2} [v(x_1) + v(x_2) + \dots + v(x_n)] \\
 &= \frac{1}{n^2} \cdot n\sigma^2 \\
 &= \frac{\sigma^2}{n}
 \end{aligned}$$


Dr J Maiti, IEM, IIT Kharagpur
18

Today again we will discuss. Next class we will discuss the sampling strategy. Today we will finish by this central limit theorem says, if you sample from normal population or other population when the sample size is large. Then the distribution of x bar, the statistical distribution that is sampling distribution of x bar will be normal. x bar is normally distributed with mean mu, variance sigma square by n. So, there are many sampling strategy, first one is random sampling, stratified sampling, cluster sampling, systematic sampling.

(Refer Slide Time: 56:32)



There are some other sampling, convenient sampling, all those things basically we talk about the sample statistic. You have collected data, how you have collected data strategy, means what method you have adopted while collecting the sample random means. You will randomize the collection procedure in such a manner that each and every observation is equally likely to come. Stratified sampling is sometime required, suppose you want to see that at different age groups what is the pattern.

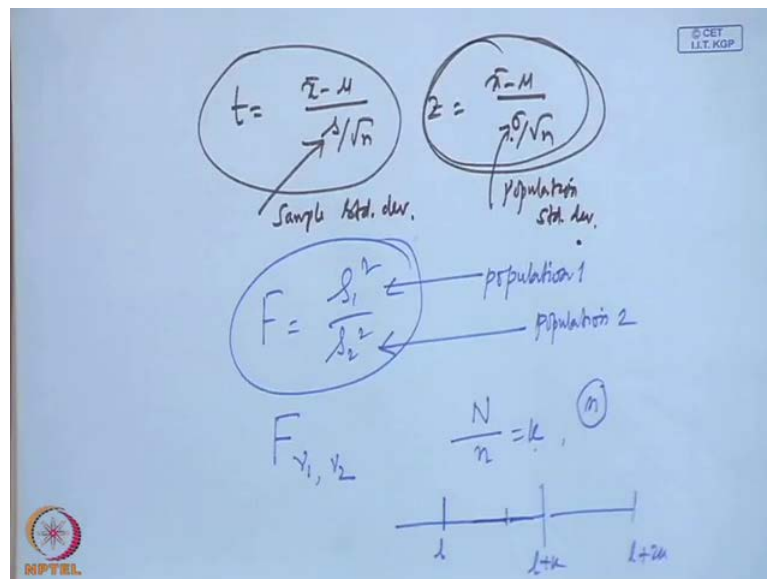
Suppose, for a particular suppose blood pressure pattern then you will take young people, middle people, old people, three strata you will create and every strata you randomly select observation cluster sampling. So, what cluster sampling is? Suppose, you think that our exit poll this time for panchayat election in Bengal, there are in West Bengal. There are so many district, each district is a cluster.

Suppose, you cannot go for every district and collect sample from each district, you can randomize the cluster. So, that means you will select based on some randomization experiment. You will select some of the select suppose you select one district based on randomization and every voters on that district is sample. Then that is single stage cluster sampling.

Now, it may so happen that you may go for 3, 4 or little more district at random, and again each of the district you will collect data from individual selecting randomly, getting me? First one is district, number of district is there, you select one randomly, one

district sample everybody that is single stage cluster sampling. Second is first randomize the selection of the district, take few district and again in each district you randomize the selection of individual, that is two stage cluster. And now, if you again in district level you can go for city also, there one more randomized sampling is possible. So, that multistage cluster it will come last one is systematic sampling systematic sampling, means you basically follow an order. For example, suppose the tenth day you will set a target for example.


(Refer Slide Time: 59:18)




Suppose, the k th observation you will collect in one sample, correct? Suppose, the population size is N, population size is N, sample size is small n so that and you want to suppose this quantity is k. You want to collect a sample of k. What you will do from the first ten observation? You may be at the eleventh point you observe, then you go on adding k. The second observation will be after you will select the l plus k th item, then l plus 2 k that way you will collect the l sample, getting me? So, these are the strategies and you go through some books and you know that as I told you that t distribution, and f distribution will be having the use.

(Refer Slide Time: 01:00:02)

Father of F-distribution




RA Fisher (1890-1962)
English Statistician



George W. Snedecor
American Mathematician

Dr J Maiti, IEM, IIT Kharagpur




And f distribution all of you know that R A Fischer, an English statistician, he has contributed a lot in agricultural sector. And George W Snedecor, american mathematician, they are the pioneer in developing the distribution. The next class I will bring.

(Refer Slide Time: 01:00:26)

References

- D C Montgomery (2001). Design and analysis of experiments. Wiley, Fifth Edition, 684p.
- Aczel A D (2010). Complete business statistics. Tata McGraw Hill, Sixth Edition, 820p.

© Dr J Maiti, IEM, IIT Kharagpur



That student distribution, that is you can remember throughout, you know Fischer and Snedecor they have basically developed the f distribution, Snedecor and Fisher, R A Fisher is one only montgomery and this Aczel A D complete business statistics. This

both books are very much available and multivariate statistics, Johnson and Richard that all that I told you in the beginning. So, next class we will discuss estimation particular confidence interval. So, I think again it will be on Tuesday, coming Tuesday, 7 o'clock evening. Tuesday 7 to 9.

Thank you very much.