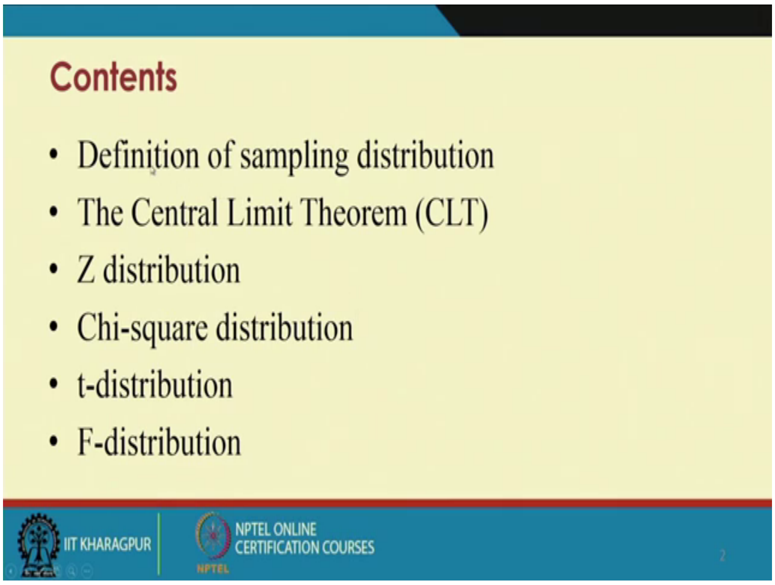
Design and Analysis of Experiments
Prof. Jhareswar Maiti
Department of Industrial and Systems Engineering
Indian Institute of Technology, Kharagpur

Lecture - 08
Sampling Distribution

Welcome to lecture 8 of design and analysis of experiment, DOE. So, today’s contents on sampling distribution, topic is sampling distribution.

(Refer Slide Time: 00:31)



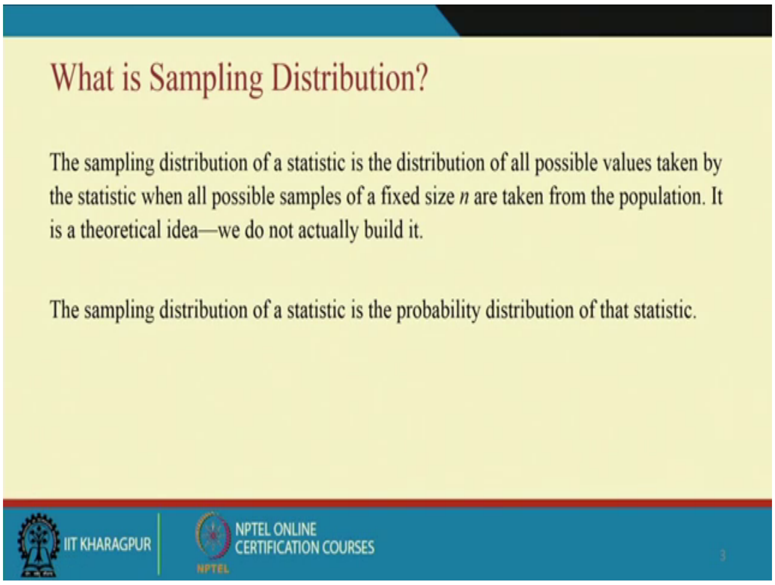
Contents

- Definition of sampling distribution
- The Central Limit Theorem (CLT)
- Z distribution
- Chi-square distribution
- t-distribution
- F-distribution

IIT KHARAGPUR
NPTEL ONLINE
CERTIFICATION COURSES

Contents are definition of sampling distribution, the central limit theorem, Z distribution, chi-square distribution, t-distribution and F-distribution.

(Refer Slide Time: 00:47)



What is Sampling Distribution?

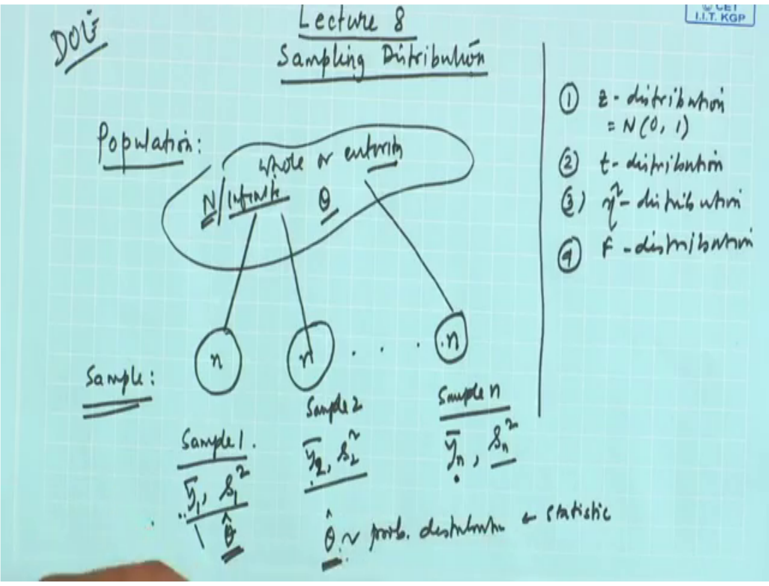
The sampling distribution of a statistic is the distribution of all possible values taken by the statistic when all possible samples of a fixed size *n* are taken from the population. It is a theoretical idea—we do not actually build it.

The sampling distribution of a statistic is the probability distribution of that statistic.

IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES
NPTEL
3

So, let us see what is sampling distribution. The sampling distribution of a statistic is the distribution of all possible values taken by the statistic when all possible samples of a fixed size n are taken from that population. It is a theoretical idea we do not actually build it. The sampling distribution of a statistic is the probability distribution of that statistic. So, let us explain a little more here.

(Refer Slide Time: 01:26)



I told you that the important concept line population and sample. For example, where this is my population this is the whole or entirety whole or entirety. Now when you take sample you take a portion of that whole or entirety of the population from different segments and the total collection will be something like this. Let this population may be finite or may be infinite, finite means the total number of items of interest will be n, infinite means it is you know what is infinite.

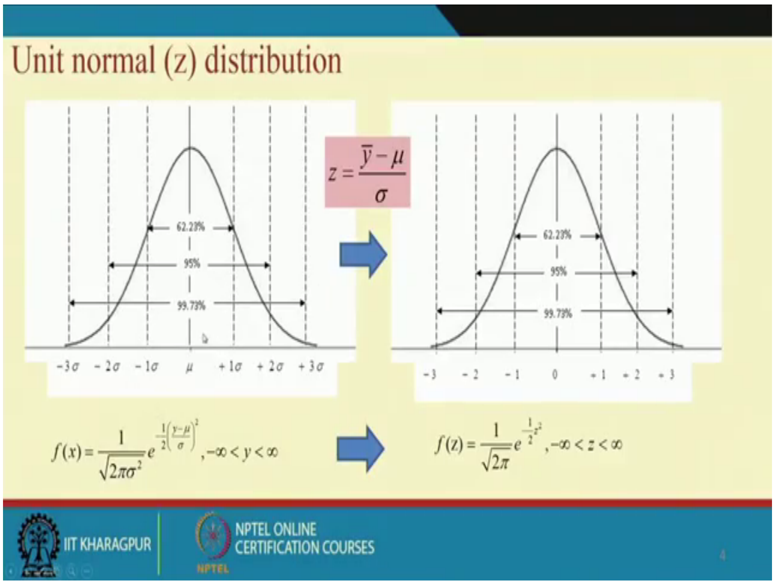
Now, suppose you have collected a sample of size small n, one time. So, let this is known as this is my sample one. So, you have collected a sample of size n from the population may be finite may be infinite theoretically we start with infinite population. Now, you compute statistics from that sample like the sample average y bar and your variance sigma s square as this sample is my first sample I can say this is y 1 bar s 1 square. Let collect another sample of size n which is sample 2 and here what happened y 2 bar and s 2 square these are the sample statistics of interest. In the same manner you can you can you can go for suppose n number of sample n with y n bar is the sample average here and s n square is the sample variance.

So, essentially we said that if you collect different sample from the same population the sample statistics will change the value will change. If the sample statistics is sample average that is y bar then y 1 bar y 2 bar and y 1 bar they will be different and what will be their values that is not known in advance and they follow certain distribution. So, in general if I say theta is the population parameter and theta cap is the sample statistic then we say the theta cap will have a probability distribution. So, this distribution of theta cap which is basically a statistic is known as sampling distribution. Theta cap may be the sample mean theta cap may be sample standard deviation of sample variance theta cap may be any other any other major statistic of interest.

So, let us see that what are the different sampling distributions. The first widely used one is Z distribution. We have discussed Z distribution earlier which is unit normal distribution which we say normally distributed with mean 0 variance 1, second one is t-distribution, third is chi-square distribution, fourth is F-distribution. These are very commonly used sampling distribution, and has lot of application in DOE subject as well as such in inferential statistics a lot of applications.

So, very quickly we revisit the unit normal distribution.

(Refer Slide Time: 06:22)



Unit normal (z) distribution

62.23%
95%
99.73%
-3σ -2σ -1σ μ $+1\sigma$ $+2\sigma$ $+3\sigma$
62.23%
95%
99.73%
-3 -2 -1 0 +1 +2 +3

$$z = \frac{\bar{y}-\mu}{\sigma}$$

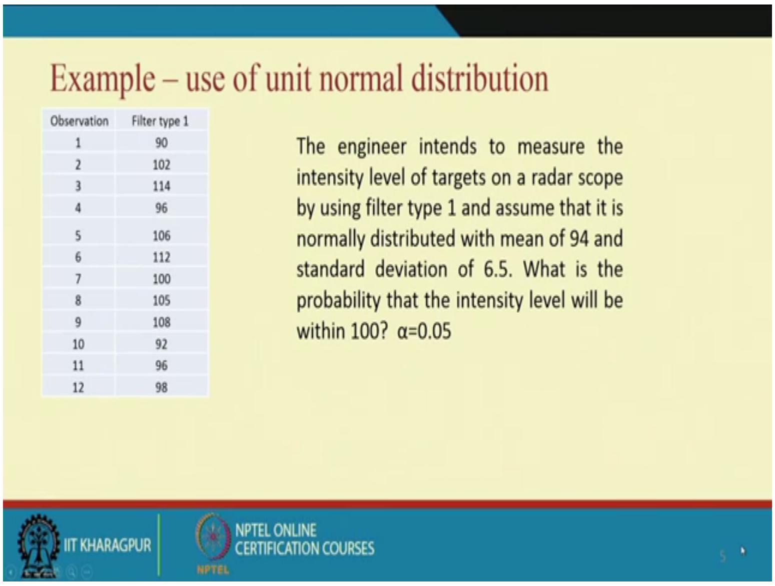
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}, -\infty < y < \infty$$

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, -\infty < z < \infty$$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, see this diagram here left hand side we are talking about pdf of y variable that is normally distributed and right hand side conversion to z. So, z equal to y bar minus mu by sigma and then you will get a pdf like this.

(Refer Slide Time: 06:48)



Example – use of unit normal distribution

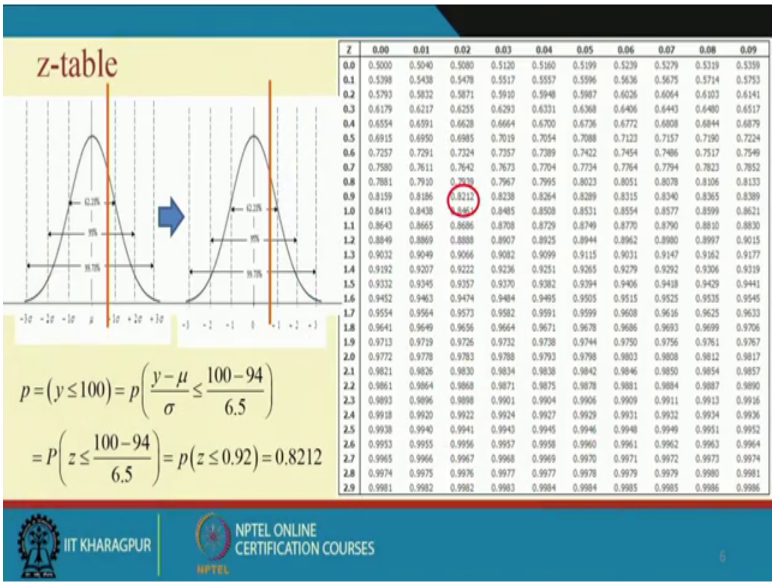
Observation	Filter type 1
1	90
2	102
3	114
4	96
5	106
6	112
7	100
8	105
9	108
10	92
11	96
12	98

The engineer intends to measure the intensity level of targets on a radar scope by using filter type 1 and assume that it is normally distributed with mean of 94 and standard deviation of 6.5. What is the probability that the intensity level will be within 100? α=0.05

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And last class I all I also have given you that how to use the unit normal distribution. So, another example, what we are saying that the engineer intends to measure the intensity level of targets on a radar scope using filter type 1 and assume that it is normally distributed with mean 94 and standard deviation 6.5. What is the probability that the intensity level will be within 100. Other way I can say that within 100 means may be less than equal to 100.

(Refer Slide Time: 07:29)



z-table

$$p=(y\leq 100)=p\left(\frac{y-\mu}{\sigma}\leq\frac{100-94}{6.5}\right)$$

$$=P\left(z\leq\frac{100-94}{6.5}\right)=p(z\leq 0.92)=0.8212$$

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9031	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9958	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

IIT KHARAGPUR

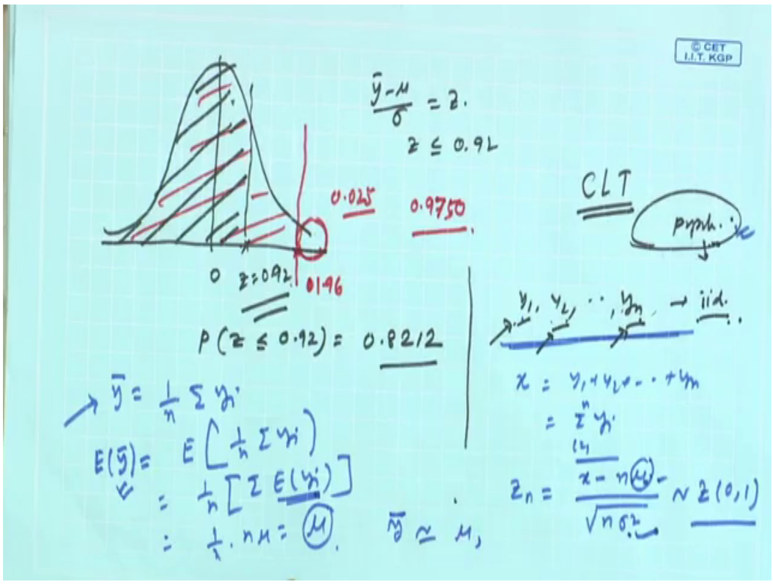
NPTEL ONLINE CERTIFICATION COURSES

NPTEL

6

So, we will use alpha 0.5, 0.05. So, theoretically suppose if we see through the distribution what happened? This mu and ultimately we are this is in the original scale and this is in the z scale. So, what you require to do you require to convert the y to z. So, y minus mu by sigma, now y is known 100 and your population mean is 94, population standard deviation is 6.5. So, ultimately if you put you got z and z equal to less than equal to 0.92 and this is the probability 0.8212. So, what happened here then?

(Refer Slide Time: 08:21)



$\frac{\bar{y}-\mu}{\sigma} = z.$

$z \le 0.92$

$P(z \le 0.92) = 0.8212$

$\bar{y} = \frac{1}{n}\sum y_i$

$E(\bar{y}) = E\left[\frac{1}{n}\sum y_i\right]$

$= \frac{1}{n}\left[\sum E(y_i)\right]$

$= \frac{1}{n}.n\mu = \mu.$ $\quad \bar{y} \simeq \mu,$

$y_1, y_2, \cdots, y_n \to$ iid.

$x = y_1+y_2+\cdots+y_n$

$= \sum_{i=1}^{n} y_i$

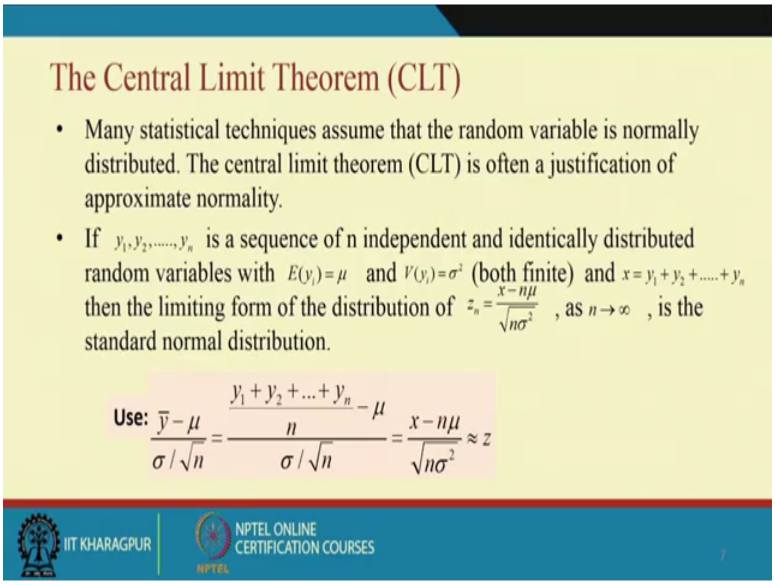
$z_n = \frac{x - n\mu}{\sqrt{n\sigma^2}} \sim Z(0,1)$

Actually when you convert it into Z distribution, this value will become 0 and you are ultimately y bar minus mu by this sigma this z value here this value is we are saying z less than equal to 0.92. So, somewhere here suppose this is the value z equal to 0.92. So, you are interested to know the area under the curve left to this value.

So, this value is, probability z less than equal to 0.92 this value is 0.8212. So, how do you get this one? See the table, see table our z value is 0.92. So, see this is z value go down get 0.9 and then add 0.02 this side. So, 0.9 to the 0.9 and 0.02, 0.92 and this value is 0.8212, 0.8212. So, suppose your z value is 1.96 then you see what happened this is 1.9 and you move to the right up to 0.06 this is 0.9750. So, it indicates that left to this value; that means, point this is 1.96 this value the area under the curve under the curve left to this point is 0.9, I think 9750.

So, I mean what is this, this area under the curve right to this one minus this means 0.025. So, this is the way you have to see read the table.

(Refer Slide Time: 10:32)



The Central Limit Theorem (CLT)

- Many statistical techniques assume that the random variable is normally distributed. The central limit theorem (CLT) is often a justification of approximate normality.
- If $y_1, y_2, \ldots\ldots, y_n$ is a sequence of n independent and identically distributed random variables with $E(y_i) = \mu$ and $V(y_i) = \sigma^2$ (both finite) and $x = y_1 + y_2 + \ldots\ldots + y_n$ then the limiting form of the distribution of $z_n = \frac{x - n\mu}{\sqrt{n\sigma^2}}$, as $n \to \infty$, is the standard normal distribution.

Use: $\dfrac{\bar{y} - \mu}{\sigma / \sqrt{n}} = \dfrac{\dfrac{y_1 + y_2 + \ldots + y_n}{n} - \mu}{\sigma / \sqrt{n}} = \dfrac{x - n\mu}{\sqrt{n\sigma^2}} \approx z$

Here we will discuss central limit theorem very important theorem, central limit theorem which is known also know abbreviated as a CLT, because what happened you will see there when you use the statistics in real life situations and although statistics whether what will be the distribution of the statistics whether it is normal or non normal, so that matters a lot. If it is normal things are easier because you have different kind of easily doing test normal distribution, normal leaky test, another test normal distribution related test. So, central limit theorem is a boon to such kind of situation because it gives you the approximate normality situation.

Let us understand what is central limit theorem. If y 1, y 2, y n is a sequence of n independent and identically distributed random variables. So, please understand the meaning. You have y 1, y 2 and y n these different observations. Observations are coming from a particular population, these population with respect to this random variable has certain distribution that may be normal maybe not normal, but some distribution is there. So, y 1, y 2, y n identical iid independent and identically distributed means when you take any observation it is not affected by the presence or absence of any other observations not related to y 2s picking up y 2 no way related to whether y 1 is already picked up or y n will be picked up later on and this is independent. Identically distributed means all y 1 y 2 y n every observations is will have this same probability distribution of that population distribution. So, if it is a normal population y 1 is also normally distributed y 2 is normal distributed y n is also normally distributed.

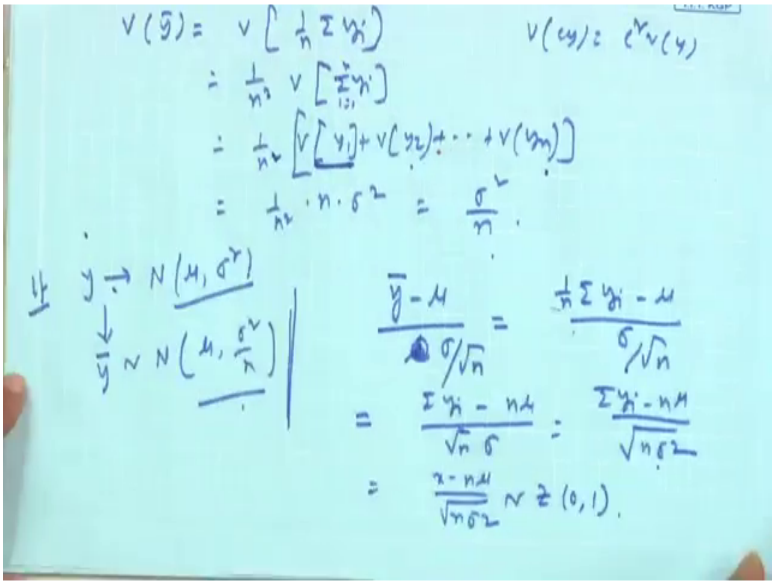
You may be thinking that if I collect y 1 y 2 data then this value one value only then, how these will be normally distributed, keep in mind you are not collected data you are thinking that you will be observing know n observations or n data points so; that means, what will be the y 1 value it can be any value. So, that is why it is probably it is probability, it has probability distribution and that to same probability distribution of the population.

If this is the situation now see that mean we are saying that expected value mu and variance sigma square for all observations and then if we create another quantity x which is sum of the observations then the limiting form of the distribution z n which is x minus n mu by root over n sigma square when n is very large each unit normal. So, let me explain. So, y 1, y 2, y n you have collected or you will be collecting you are creating another variable called x which is y 1 plus y 2 plus dot dot dot y n; that means, sum of y i i equal to 1 to n.

So, now you can create Z distribution where we are writing like z n which is sum total of all those things this is x minus n times mu divided by root over n sigma square where mu and sigma are the parameters related to this population and hence it is the expected value of y 1 y i is mu and variance of y i is sigma square. So, then this is unit normal. This is unit normal.

Now let us see that what is the use of this unit normal one important use you know, that was you calculated y bar which is one by n sum total of y i, y bar is a statistic, so it is a random variable. So, it has a expected value, it has an expected value. So, what will be the expected value of y bar? Let us see that then it will be expected value of 1 by n sum of y i which is 1 by n; that means, sum of the expected value of y i. So, expected value of y i is mean, the sum of mu that will be n mu, 1 my n into n mu equal to mu. So, that mean y bar is a random variable with mean. So, y bar is a random variable and it has if it has mean mu.

(Refer Slide Time: 16:10)



$$V(\bar{y}) = V\left[\frac{1}{n}\sum y_i\right) \qquad V(cy) = c^2 V(y)$$

$$= \frac{1}{n^2} V\left[\sum_{i=1}^{n} y_i\right]$$

$$= \frac{1}{n^2}\left[V[y_1] + V(y_2) + \cdots + V(y_n)\right]$$

$$= \frac{1}{n^2}\cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}.$$

If $y \rightarrow N(\mu, \sigma^2)$

$\downarrow$

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

$$\frac{\bar{y} - \mu}{\sigma/\sqrt{n}} = \frac{\frac{1}{n}\sum y_i - \mu}{\sigma/\sqrt{n}}$$

$$= \frac{\sum y_i - n\mu}{\sqrt{n}\,\sigma} = \frac{\sum y_i - n\mu}{\sqrt{n\sigma^2}}$$

$$= \frac{x - n\mu}{\sqrt{n\sigma^2}} \sim z(0,1).$$

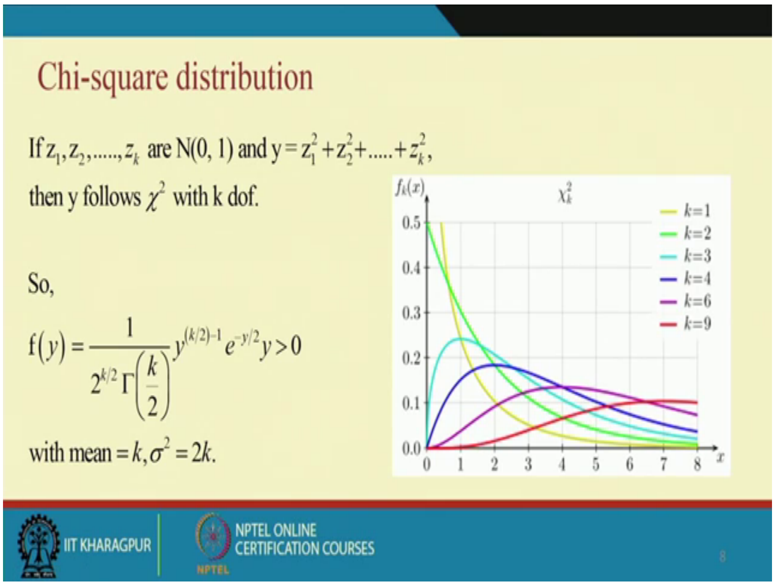
Now, what will be the variance of y bar? That mean variance of one by n sum total y i. So, I told you earlier variance of c y equal to c square variance of y. So, then from this you can write this is 1 by n square variance of sum total of y i. Again variance of sum total of y i is y is are independent. So, that will be sum up variability each variability.

So that means, 1 by n square, 1 by n square variance of y 1 plus variance of y 2 like plus variance of y n. So, then this is nothing but 1 by n square and variance of y 1, y 2, y n all are same there is sigma square. So, n such sigma values will be added, n sigma square this (Refer Time: 17:07) sigma square by n. So, that mean if y is the original variable that is the that is basically characterizing the population, now for, and if it is normally distributed with mu and sigma square then if you sample then the statistics y bar it will also be normally distributed with mu and sigma square by n mu and sigma square by n. So, that is what is the interesting observations here.

Second thing is that suppose you create another quantity called y bar minus mu by let it be sigma let it be sigma by root n. So, can we not write down this 1 y by 1 by n sum of y i minus mu by sigma by root n. So, this can be, what is this? Sum of y i minus n mu divided by n will be multiplied if I multiply n both n then root n into sigma we can write down sum of y i minus n mu by root over n sigma square, then this sum of y i is nothing, but x earlier we have seen that sum of y i is x. So, sum of y i is x and that; that means, this is x minus n mu by root over n sigma square. So, as a result it is Z distribution with 0 1.

So, this is what is the, I say that huge of central limit theorem. So, that mean if you create such statistics like this y bar minus mu and which is basically following this concept then it is unit normal.

(Refer Slide Time: 19:30)



Chi-square distribution

If $z_1, z_2, \ldots\ldots, z_k$ are N(0, 1) and $y = z_1^2 + z_2^2 + \ldots\ldots + z_k^2$, then y follows χ^2 with k dof.

So,

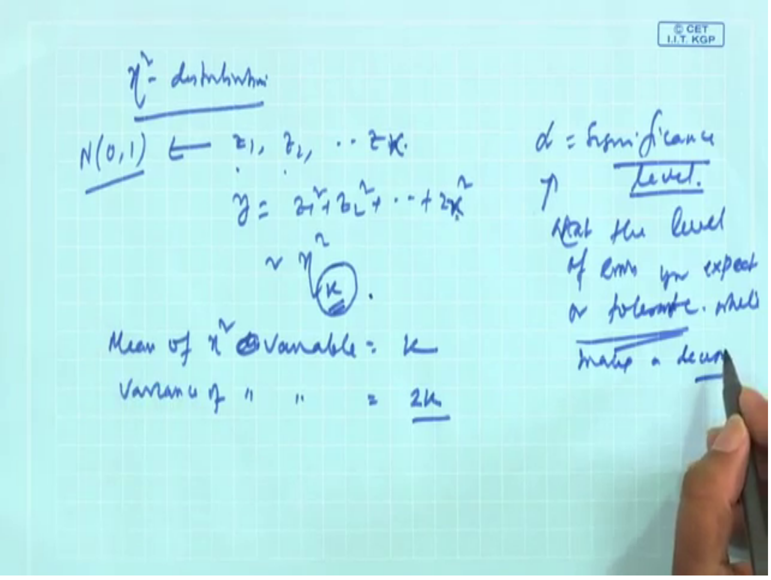
$$f(y) = \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} y^{(k/2)-1} e^{-y/2} y > 0$$

with mean $= k, \sigma^2 = 2k$.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, we will discuss another important distribution which is known as chi-square distribution, chi-square distribution.

(Refer Slide Time: 19:37)



χ^2- distribution

$N(0,1) \leftarrow z_1, z_2, \cdots z_K.$

$y = z_1^2 + z_2^2 + \cdots + z_K^2$

$\sim \chi^2_{\boxed{K}}.$

Mean of χ^2 variable = K

Variance of " " = 2K

α = significance level.

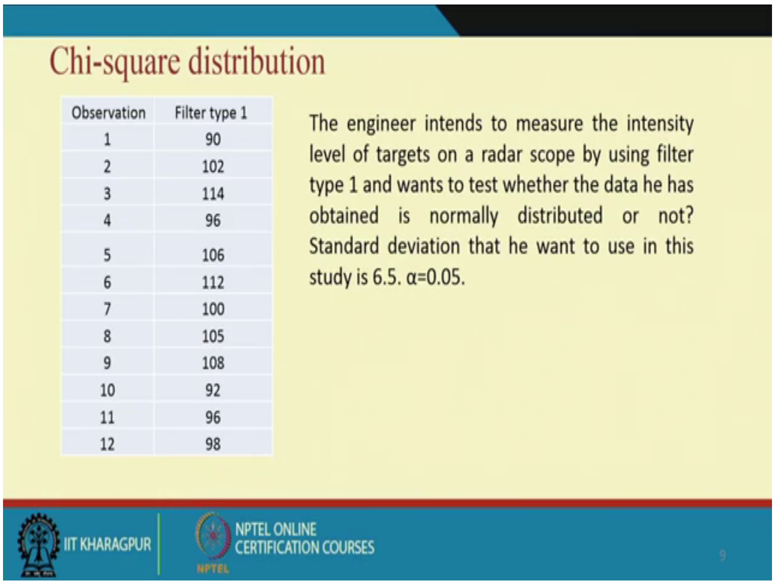
What the level of error you expect or tolerate. while making a decision

So, let us assume that you have unit normal distribution that mean the population is unit normal in that sense and you have collected then n observation z 1, z 2, z n, n observations. Now if you create one variable called y which is z 1 square plus z square plus plus z n square then this y is chi-square distributed with if it is n then suppose I write I want to write k. So, z 1 z 2 z k let it be because n we have use several times. So, let it be k. So, then it will be chi-square k that mean chi-square distribution with k degrees of freedom.

Now let us see the how the what is the pdf for chi-square distribution, see f y this is 1 by 2 to the power k gamma k by 2 y to the power k by 2 minus 1, e to the power minus y by 2, y by 2 and there will be comma y greater than y greater than 0. So, what is the mean of chi-square distribution? Mean m chi-square distribution is the degree of freedom k and variance is 2 times k. So, chi-square mean, mean of chi-square distribution chi-square variable if I write it is k and variance of chi-square it will be 2 times k. Degree of freedom is very very important concept because if degree of freedom changes the shape of the chi-square distribution also changes you will see here.

So, here we are we are assuming that this is chi-square. So, if k equal to 1 then that the shape of the distribution will be like this, if it is 2 then the green color, if it is 3 less green, if it is 4 this blue color, if it is 6 is maroon, if it is 9 this red see that changing of the shape of the distribution shape and size. So, degree of freedom is very very important concept all of you please remember when you use chi-square distribution. Please do not forget to mention that the degree of freedom also. If you say the y is chi-square distributed it is meaningless, if you say y is chi-square distributed we k degrees of freedom or n degrees of freedom then it is meaningful.

(Refer Slide Time: 22:31)



Chi-square distribution

Observation	Filter type 1
1	90
2	102
3	114
4	96
5	106
6	112
7	100
8	105
9	108
10	92
11	96
12	98

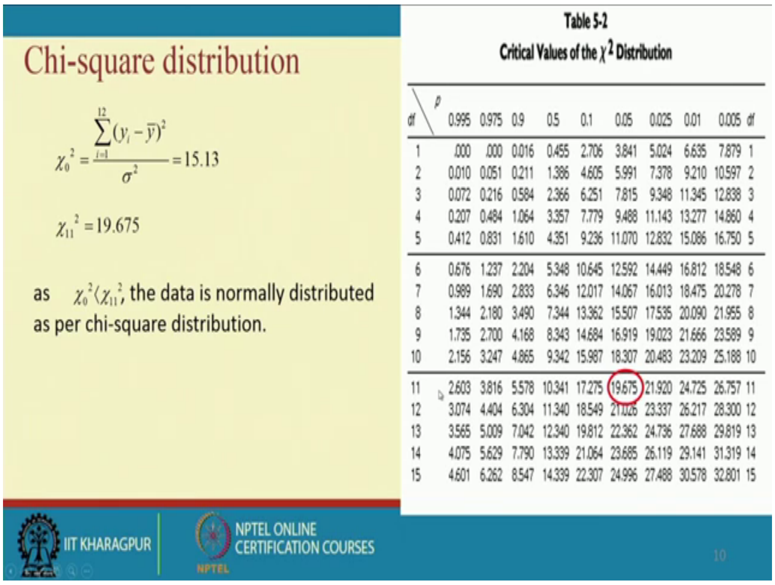
The engineer intends to measure the intensity level of targets on a radar scope by using filter type 1 and wants to test whether the data he has obtained is normally distributed or not? Standard deviation that he want to use in this study is 6.5. α=0.05.

IIT KHARAGPUR
NPTEL ONLINE
CERTIFICATION COURSES
9

Now, I will show you one example that use of chi-square distribution here. So, this example we have we have discussed several times that the radars intensity level on target for error in radar scope experiment and using filter type 1 this is the data 12 data points. Suppose you want to know whether the data is coming from normal distribution, normally distributed population or not and suppose you know the stand standard deviation is 6.5 use alpha equal to 0.05 alpha is the significance level, significance level. We will discuss what is significance level in hypothesis testing in more detail. For the time being you think that significant level means it basically talks about what level of errors the level of errors, level of errors you expect or you expect or you tolerate while making decision making a decision.

What is the decision here? Here we are saying whether this data come comes from normal population or not. Based on this chi-square analysis you will accept yes it is coming from this, but there will be error this is what is the alpha. So, 5 percent error is accepted here.

(Refer Slide Time: 24:12)



Chi-square distribution

$$\chi_0^2 = \frac{\sum_{i=1}^{12}(y_i - \bar{y})^2}{\sigma^2} = 15.13$$

$$\chi_{11}^2 = 19.675$$

as $\chi_0^2 < \chi_{11}^2$, the data is normally distributed as per chi-square distribution.

Table 5-2

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

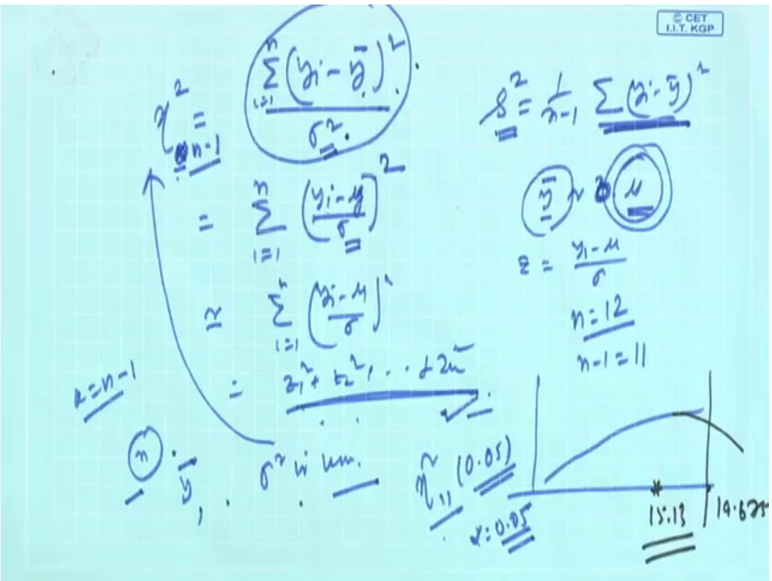
IIT KHARAGPUR

NPTEL ONLINE CERTIFICATION COURSES

10

So, now see the quantity what is the quantity we are using here for chi-square? We are using the quantity that y i minus y bar square divided by sigma square I equal to 1 to n. What is this y i minus this one? If you go to formula s square you will find out this will be 1 by n minus 1 sum total of y i minus y bar square.

(Refer Slide Time: 24:22)


© CET I.I.T. KGP
$\chi^2_{n-1} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sigma^2}$
$= \sum_{i=1}^{n}\left(\frac{y_i-\bar{y}}{\sigma}\right)^2$
$\approx \sum_{i=1}^{n}\left(\frac{y_i-\mu}{\sigma}\right)^2$
$= z_1^2 + z_2^2, \dots + z_n^2$
$k = n-1$
n
$\bar{y}$
σ^2 is known
$s^2 = \frac{1}{n-1}\sum(y_i-\bar{y})^2$
$\bar{y} \to \mu$
$z = \frac{y_i-\mu}{\sigma}$
$n = 12$
$n-1 = 11$
$\chi^2_{11}(0.05)$
$\alpha = 0.05$
15.13
19.675

So, now if you divide it by sigma square, this is basically s square the follower the variance sample variance there we have use this quantity is very important here. Now this can be rewritten like this y equal to 1 to n, y i minus y bar by sigma square. What is y i by minus y bar by sigma sigma square? So you will find out this is this is nothing, but z y, y bar if we are think that y bar is estimate of estimate of mu y bar is estimate of mu then this is this can be written approximately i equal to 1 to n y r minus mu by sigma square and you know the standardization z is y i minus mu by sigma. So, this is nothing, but the z one square plus z 2 square plus z n square. Here we are writing n z n square.

So, what happened then? That mean this is a quantity which is nothing, but sum of the can be approximate to sum of the square of the unit normal variable then it will be chi-square distribution. So, this is the huge. Now if you create a statistics in and then you find out that these statistics is nothing, but sum of the square of unit normal variable then that statistics will follow chi-square distribution.

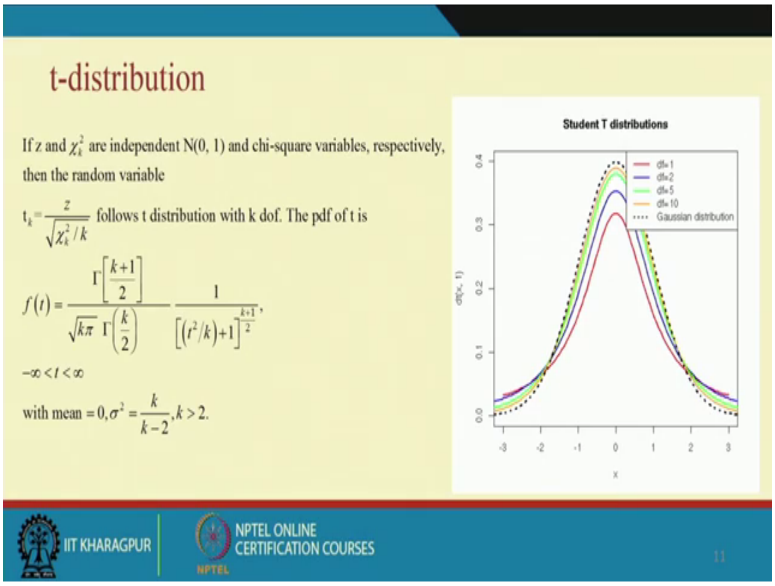
Now, what will be the degrees of freedom here what will be this k in this case? So, in this case k is not n the reason is you have not you have estimated mu the mean. So, estimated mu in terms of y bar, one degree freedom lost so that is why k is n minus 1. What does it mean? Suppose in your sample your sample size is n you have computed y bar you have come then you are basically creating chi-square in using this formula when sigma square is known sigma square is known. So, what and then ultimately what is happening y i minus y bar this quantity becoming chi-square as y bar is computed one degree is lost. So, this one will write chi-square n minus 1, chi-square n minus 1 is this quantity means this quantity follows chi-square distribution with n minus 1 degrees of freedom.

Anyhow in this example, you we have computed this chi-square 0 chi-squared 0 this is just notation used here this value is 15.13. Suppose I know that our n is 12. So, n minus 1 is 11 then the distribution here with k degrees of freedom k equal to 9 is something like this. So, then k equal to 1 here it will be something also suppose this something like this let it be like this, and then what we are doing here chi-square it is chi-square 11. So, if I come say that that alpha equal to 0.05 the detail will be told to later on and in that case what happen, if for you we want to find out what is the value of chi-square 11, 0.05.

So, you will see chi-square table then here in chi-square table there will be that rows represent and columns and the first column all degrees of freedom and in the all other columns these are the different probability values. So, our chi-square 11 with probability significant 0.05 this value is 19.675. So, chi-square 9 by sigma and our computed chi-square value is 12, it simply indicates that suppose the threshold value here suppose if I consider something like this, this one let it be 19.675 and the computed value is somewhere here 15.13. So, we it is coming from, it coming from normal population.

So, this is the huge although you require huge of the significance level and hypothesis testing knowledge, but I intentionally kept here so that you understand that where to use chi-square when your statistics will be chi-square. So, that is the fundamental.

(Refer Slide Time: 29:53)



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

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

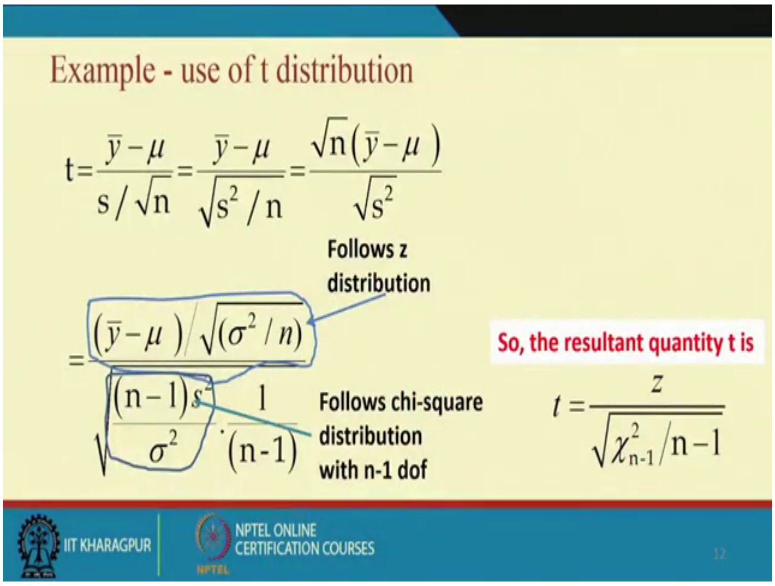
$-\infty < t < \infty$

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

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, another is important distribution is t-distribution. So, t-distribution is interestingly it is the ratio of z by chi-square k by k.

(Refer Slide Time: 30:04)



Example - use of t distribution

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

Follows z distribution

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

Follows chi-square distribution with n-1 dof

So, the resultant quantity t is

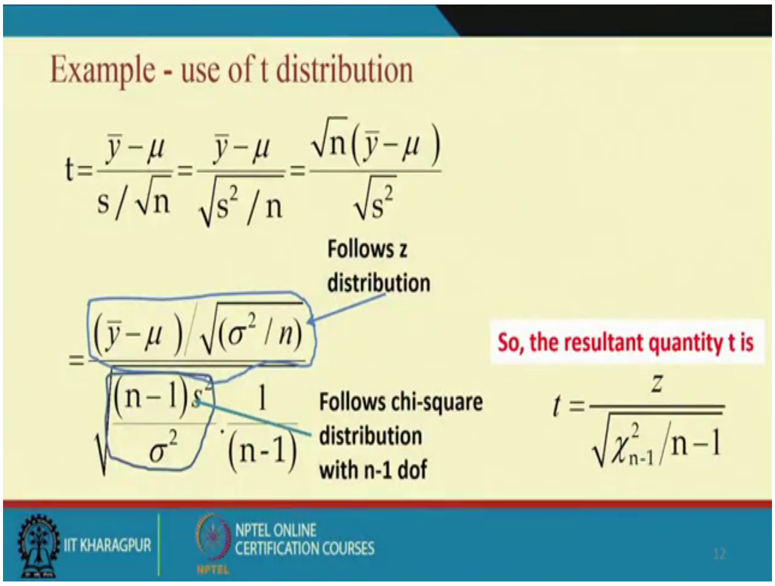
$$t = \frac{z}{\sqrt{\chi^2_{n-1}/n-1}}$$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Suppose you create a situation when your statistics is such that it is the ratio of a standard normal distribution and square root of chi-square distribution normalized by k divided by k. So, then this quantity, this quantity will follow t-distribution. So, t-distribution has this kind of probably pdf, its complicated pdf and you go through montgomery chapter 2 you will get this kind of formulation. And it is very similar to normal distribution and very important distribution in particularly, in statistics because applied statistics particularly because of all the parameter test everything will be based on this.

So, this is what is our, that you see that how t-distribution looks like. So, t-distribution after when sample size is large t-distribution will become almost normal distribution, equivalent to normal distribution.

(Refer Slide Time: 31:23)



Example - use of t distribution

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

Follows z distribution

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

Follows chi-square distribution with n-1 dof

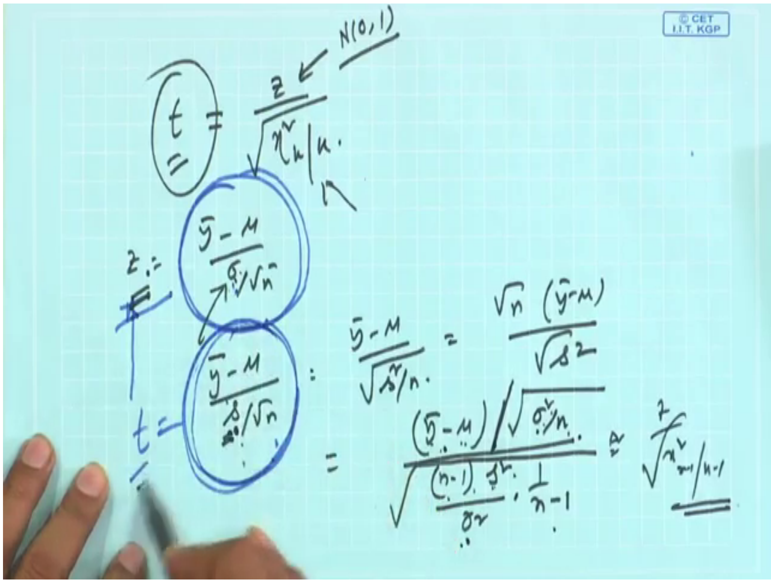
So, the resultant quantity t is

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

IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES

Now, let us see the use of t-distribution suppose in previous example we have used y bar minus mu by sigma by root n this is z from central limit theorem.

(Refer Slide Time: 31:28)

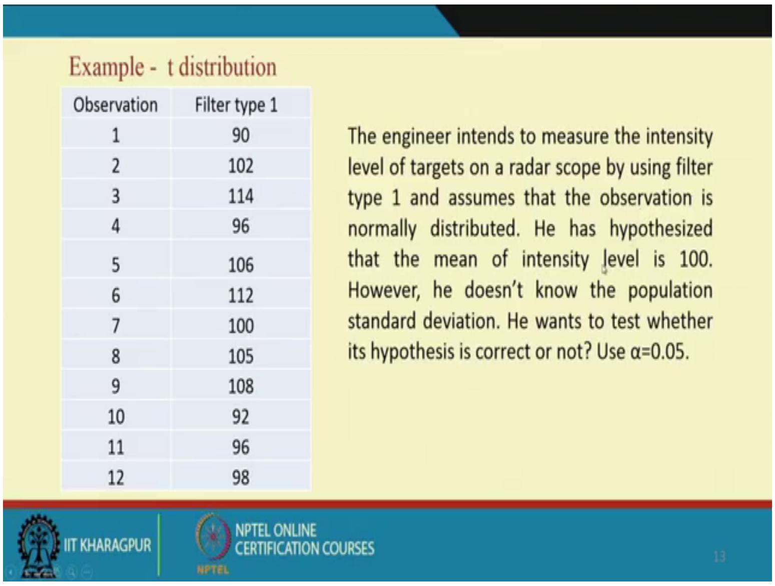

© CET I.I.T. KGP
N(0,1)
$t = \frac{z}{\sqrt{\chi^2_n / n}}$
$z = \frac{\bar{y}-\mu}{\sigma/\sqrt{n}}$
$t = \frac{\bar{y}-\mu}{s/\sqrt{n}} = \frac{\bar{y}-\mu}{\sqrt{s^2/n}} = \frac{\sqrt{n}(\bar{y}-\mu)}{\sqrt{s^2}}$
$= \frac{(\bar{y}-\mu)/\sqrt{\sigma^2/n}}{\sqrt{\frac{(n-1)s^2}{\sigma^2} \cdot \frac{1}{n-1}}} = \frac{z}{\sqrt{\chi^2_{n-1}/n-1}}$

And now what happens suppose sigma is not known, instead this you are using you creating a statistics which is something like this s by root n instead of sigma you are writing s then the sigma is constant, but s is a random variable. So, I can write this is nothing, but y bar minus mu by root over a square by n.

Now, this can be written like this root n y bar minus mu by root over of s square which further we can write like this that this will be y bar minus mu divided by root over sigma square by n divided by n minus 1, s square by sigma square into 1 by n minus 1 whole square root, whole squared root. So, this is sigma square and n these are constant forget about this, but what is y bar y bar minus mu by sigma if I this is by. So, what will happen? This sigma then it will be z and n minus 1 s square by sigma square, this is what is this, this is chi-square we have seen the chi-square, now chi-square and then into 1 by mu chi-square by n minus 1. So, this is nothing, but a z by root over chi-square here it is n minus 1 by n minus 1. So, it is t-distribution.

So, it is the interesting result that when you create this kind of statistics most of the time we use this, but if you do not know sigma you use s, if you know sigma this will be z distributed, if you do not know sigma then you will use a sample standard deviation this will be t. But again what happened as a as you see that if the sample size is large the ultimately converges with z. So, for large sample size under this situation also you will be able to use Z distribution, but if sample size is small then and then your statistics is this you have to use t-distribution.

(Refer Slide Time: 34:04)



Example - t distribution

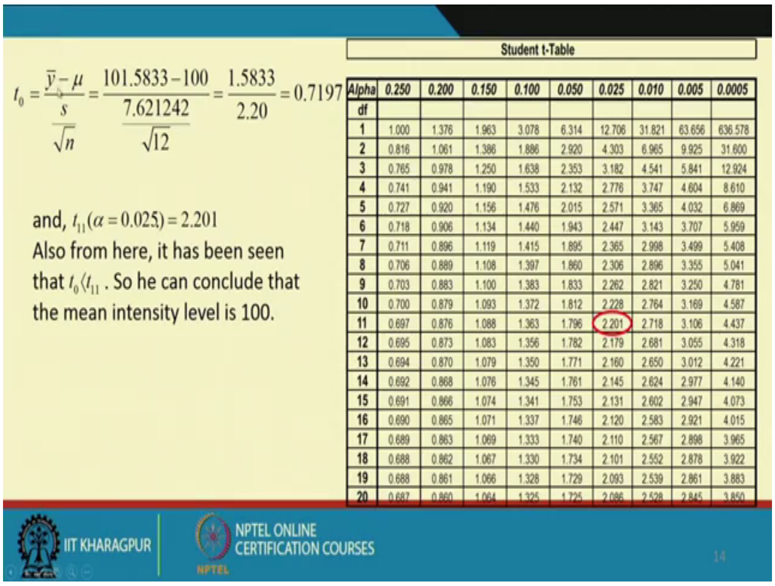
Observation	Filter type 1
1	90
2	102
3	114
4	96
5	106
6	112
7	100
8	105
9	108
10	92
11	96
12	98

The engineer intends to measure the intensity level of targets on a radar scope by using filter type 1 and assumes that the observation is normally distributed. He has hypothesized that the mean of intensity level is 100. However, he doesn't know the population standard deviation. He wants to test whether its hypothesis is correct or not? Use α=0.05.

IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES 13

So, this is the example, same example and here what happens sigma is not known the sigma not known we computed s from the data and then integer is computed like this and again from table we found that this computed value is less than the tabulated value.

(Refer Slide Time: 34:09)



$$t_0 = \frac{\bar{y}-\mu}{\frac{s}{\sqrt{n}}} = \frac{101.5833-100}{\frac{7.621242}{\sqrt{12}}} = \frac{1.5833}{2.20} = 0.7197$$

and, $t_{11}(\alpha = 0.025) = 2.201$

Also from here, it has been seen that $t_0 \langle t_{11}$. So he can conclude that the mean intensity level is 100.

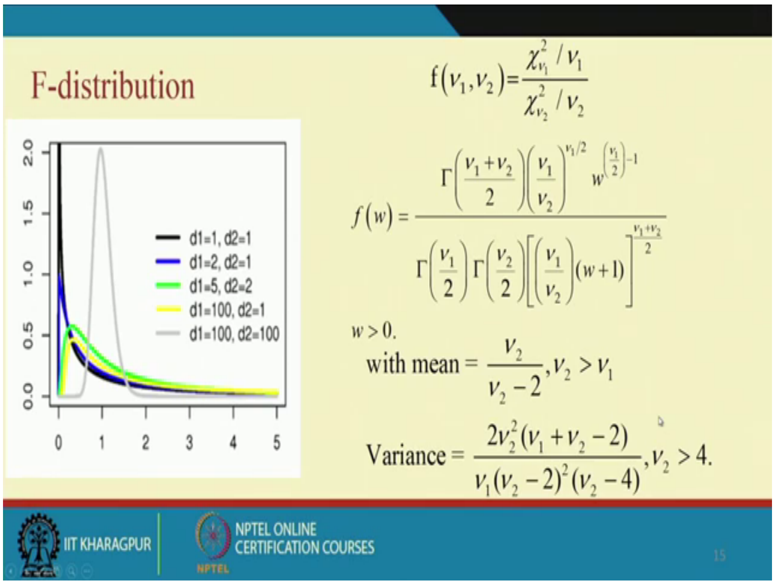
Student t-Table

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	2.528	2.845	3.850

IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES 14

So, this is basically we can say that the mean intensity level is 100 whether it is not far away from 100.

(Refer Slide Time: 34:33)



F-distribution

$$f(v_1,v_2)=\frac{\chi^2_{v_1}/v_1}{\chi^2_{v_2}/v_2}$$

$$f(w)=\frac{\Gamma\left(\frac{v_1+v_2}{2}\right)\left(\frac{v_1}{v_2}\right)^{v_1/2}w^{\left(\frac{v_1}{2}\right)-1}}{\Gamma\left(\frac{v_1}{2}\right)\Gamma\left(\frac{v_2}{2}\right)\left[\left(\frac{v_1}{v_2}\right)(w+1)\right]^{\frac{v_1+v_2}{2}}}$$

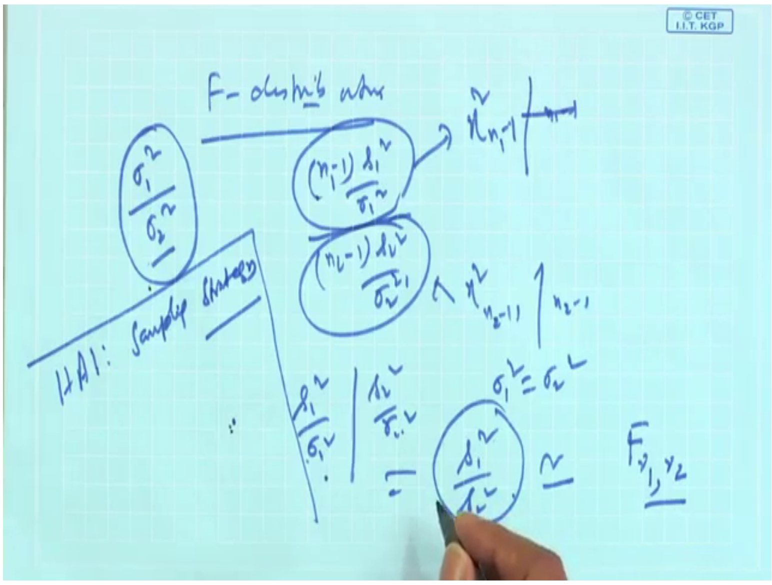
$w>0.$

with mean $=\frac{v_2}{v_2-2}, v_2>v_1$

Variance $=\frac{2v_2^2(v_1+v_2-2)}{v_1(v_2-2)^2(v_2-4)}, v_2>4.$

So, the last important standard sampling distribution is F-distribution, F-distribution.

(Refer Slide Time: 34:41)



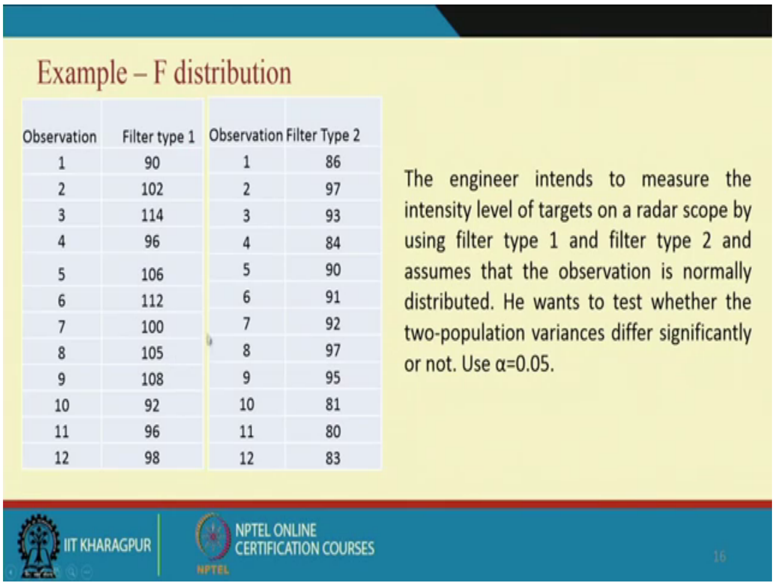
F-distribution
σ_1^2/σ_2^2
$(n_1-1)s_1^2/\sigma_1^2$
$\chi^2_{n_1-1}$
$(n_2-1)s_2^2/\sigma_2^2$
$\chi^2_{n_2-1}$
n_2-1
HA1: Sampling Strategy
$\frac{s_1^2}{\sigma_1^2} / \frac{s_2^2}{\sigma_2^2}$
$\sigma_1^2=\sigma_2^2$
$\frac{s_1^2}{s_2^2} \sim F_{v_1,v_2}$
© CET I.I.T. KGP

Suppose the example is suppose you are basically interested to compare the variance of two different populations, like population one variance is sigma one square population 2 variance is sigma 2 square. You will create a ratio and we have found out that n minus 1 s 1 square by sigma 1 square by suppose n 1 n 2 minus 1 s 2 square by sigma 2 square, if we create you we all know this one follow chi-square distribution with n minus n 1 minus 1 degrees of freedom this follow chi-square distribution with n 2 minus 1 degrees of freedom and if you divide this if I divide this by n 1 minus 1 and this by n 2 minus 1. So, ultimately this quantity will become s 1 square by sigma 1 square divided by s 2 square by sigma 2 square.

If we further assume that sigma 1 square equal to sigma 2 square then what will happen sigma one sigma 2 square will cancel out this result quantity will be s 1 square minus s 2 square. So, this kind of situation is nothing but the ratio of 2 chi-square variable normalized by their respective degrees of freedom, then this quantity follows F-distribution with suppose nu 1 the numerator degree of freedom and nu 2 denominator degree of freedom. So, what is this? Suppose you are having a statistics which is the ratio up to chi-square variable and which is weighted by their respective degrees of freedom and then the resultant quantity or the statistics this will follow normal sorry this will follow F-distribution.

F distribution is a very complicated one in terms of pdf and it also depends on the degrees of freedom f degree numerator and denominator degrees of freedom changes then F-distribution shape and size will also change.

(Refer Slide Time: 36:57)



Example – F distribution

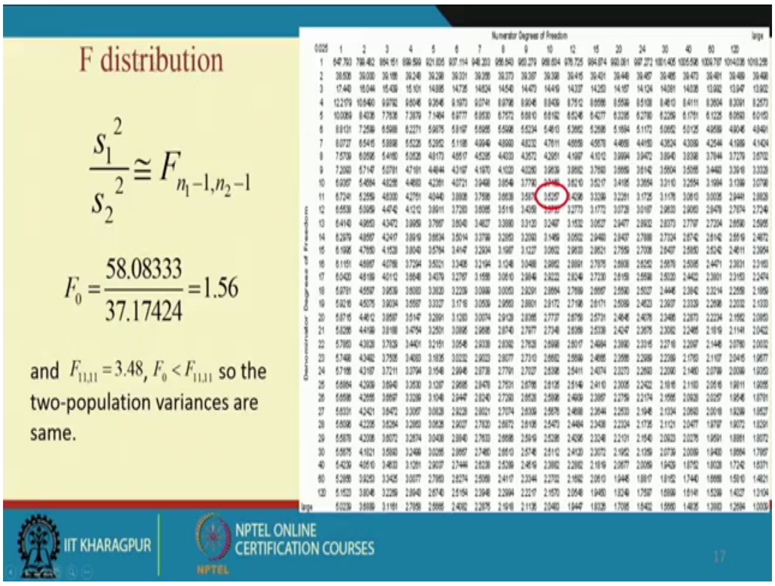
Observation	Filter type 1	Observation	Filter Type 2
1	90	1	86
2	102	2	97
3	114	3	93
4	96	4	84
5	106	5	90
6	112	6	91
7	100	7	92
8	105	8	97
9	108	9	95
10	92	10	81
11	96	11	80
12	98	12	83

The engineer intends to measure the intensity level of targets on a radar scope by using filter type 1 and filter type 2 and assumes that the observation is normally distributed. He wants to test whether the two-population variances differ significantly or not. Use α=0.05.

IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES 16

So, let us see the example. So, here what happened we have done the experiment earlier with filter type 1 and filter type 2 and suppose we got 12 observation experimental data in first case and 12 another case about we want to know that whether 2 population variance differ significantly or not. So, you can use F-distribution and with such development s 1 square by s 2 square there will be normally f distributed with n 1 minus 1 and n 2 minus 1 degrees of freedom n 1 n 2 both same 12, so 11 11 degrees of freedom.

(Refer Slide Time: 37:21)



F distribution

$$\frac{s_1^2}{s_2^2} \cong F_{n_1-1,n_2-1}$$

$$F_0 = \frac{58.08333}{37.17424} = 1.56$$

and $F_{11,11} = 3.48$, $F_0 < F_{11,11}$ so the two-population variances are same.

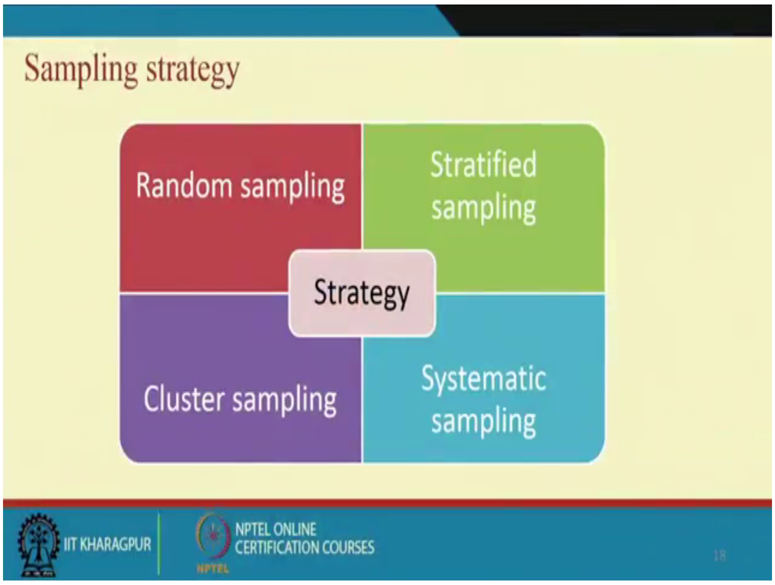
Numerator Degrees of Freedom
Denominator Degrees of Freedom
0.025
large
IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES
17

So, from there sample you compute these 2 values s 1 square s 2 square, it is 1.56. Now, from theory you see that whether this value, this, what is the value theoretical value from table you will get and then you compare the 2 values. If the computed value is less than the tabulated value you accept that the variances are not different they are same.

Now, how to see this distribution F-distribution? Keep in mind here for every probability your level of significance there is a probability value alpha or alpha by 2 alpha value. So, what happened? The numerator degrees of freedom will be across the column and denominator degrees of freedom will be across the rows and as it is 11 11 you see that 11th row and you are sorry 10th row and no; 11 11, 11 and 11. This side 11 and this side also 11 because any 12, n minus 1 will be 11 not 10. So, 11 11, but in the column side in this table 11 is not available. So, you can what you can do? You can find out the value for 10 11 nu 1 and 10 nu 2, 11 nu 1 and 12 nu 2 and then take the average of the two that will give you this tabulated value 3.48.

For the time being you understand that if the tabulated value is more than the computed value from the sample variances then the two variances are not different, what is the error or whether the decision is correct or not all those things will be discussed later.

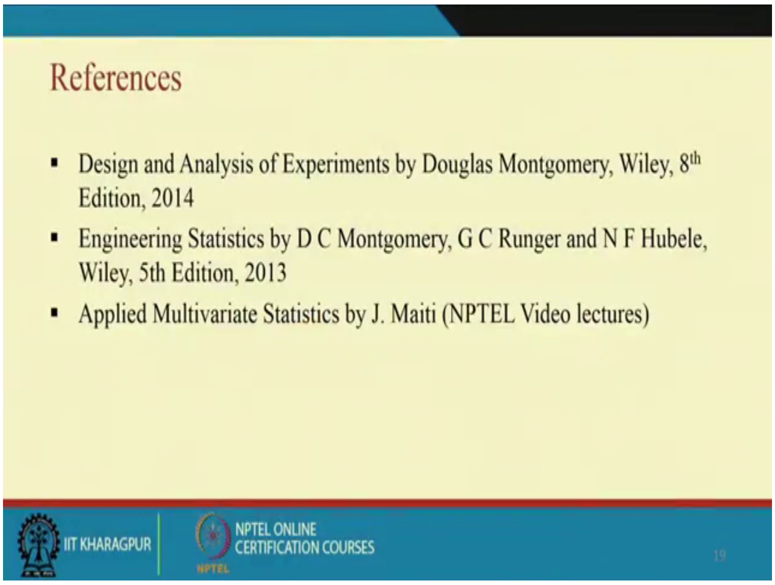
(Refer Slide Time: 39:41)



Sampling strategy

So, I think you must go through sampling strategy because it is sampling strategy is very important one and I want you to do some kind of home study, home assignment study on sampling strategies. The sampling strategies can be random sampling, stratified sampling, cluster sampling and systematic sampling. There can be some other variant of sampling strategies also, but for the time being you must you have gone through random sampling, but what is stratified sampling, what is cluster systematic sampling, why random sampling is used in this case. So, all those things, all those things you please go through and it is your home assignment, yes home assignment. So, home assignment 1 will be sampling strategy it is basically this student you will submit that what are the, this 4th type of sampling strategy when it is used and something like this.

(Refer Slide Time: 40:41)



References

- Design and Analysis of Experiments by Douglas Montgomery, Wiley, 8^{th} Edition, 2014
- Engineering Statistics by D C Montgomery, G C Runger and N F Hubele, Wiley, 5th Edition, 2013
- Applied Multivariate Statistics by J. Maiti (NPTEL Video lectures)

IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES
19

So, with this I finished this sampling distribution. Please keep in mind very important thing is that it is a very important concept sampling distribution is nothing, but the probability distribution of a statistic. There can be different kind of sampling distribution depending on the statistics of interest for example, there will be z statistics which statistics which probability Z distribution, statistics that will follow t-distribution, statistics that will follow chi-square distribution, statistics that will follow your F-distribution.

Please keep in mind you must know when which kind of distribution will be used because these distribution probability distribution will be used in hypothesis testing, in confidence interval, in estimation particularly. And these are the references, I have taken some amount of this material from my earlier lecture, applied multivariate statistical modeling and it is available in the NPTEL video models, video lectures.

Thank you very much.