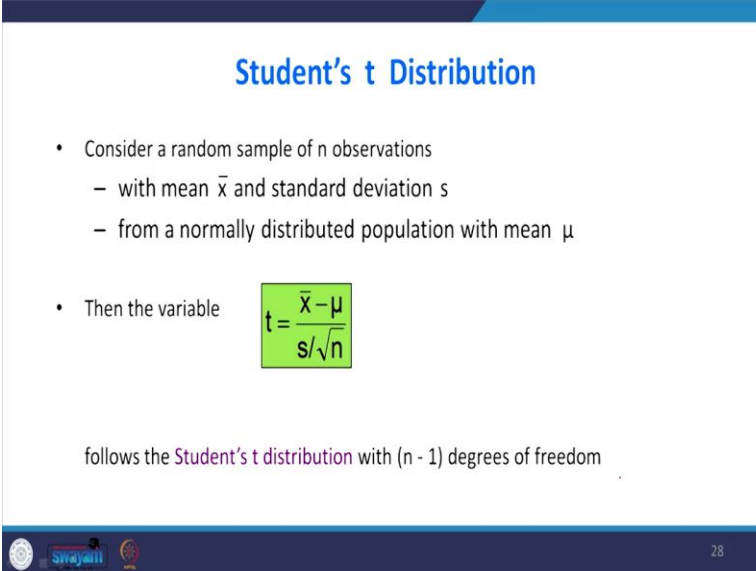


**Data Analytics with Python**  
**Prof. Ramesh Anbanandam**  
**Department of Management Studies**  
**Indian Institute of Technology – Roorkee**

**Lecture – 15**  
**Confidence Interval Estimation\_ Single Population – II**

Dear students in the previous class we have predicted the population mean with the help of sample mean where the condition was the Sigma square is unknown now know the Sigma square is known. Now we will see the next case where Sigma square is unknown then we will see how to predict the population mean.

**(Refer Slide Time: 00:49)**



**Student's t Distribution**

- Consider a random sample of  $n$  observations
  - with mean  $\bar{x}$  and standard deviation  $s$
  - from a normally distributed population with mean  $\mu$
- Then the variable 
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

follows the Student's t distribution with  $(n - 1)$  degrees of freedom

28

For this purpose you have to use student t distributions consider a random sample of  $n$  observations with the mean  $\bar{x}$  and standard deviation  $s$  from a normally distributed population with the mean  $\mu$  then the variable  $t$  is nothing but  $\bar{x} - \mu$  divided by  $s$  divided by root  $n$  you see that there is a connection between  $Z$ ,  $Z$  to be used to write  $\bar{x} - \mu$  divided by  $\sigma$  by root  $n$ . But in the  $t$  distribution what is happening the  $\sigma$  is unknown so we really we are going to use sample standard deviation.

The other thing is this  $n$  should be the smaller number it is less than 30. So, when will you go for

t distribution when Sigma is unknown when n is less than 30 then the variable  $t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$  follows the student distributions with  $n - 1$  degrees of freedom.

**(Refer Slide Time: 01:49)**

**Confidence Interval for  $\mu$  ( $\sigma^2$  Unknown)**

- If the population standard deviation  $\sigma$  is unknown, we can substitute the sample standard deviation,  $s$
- This introduces extra uncertainty, since  $s$  is variable from sample to sample
- So we use the  $t$  distribution instead of the normal distribution

29

Now we will see how to predict the confidence interval for MU when Sigma Square is unknown if the population standard deviation Sigma is unknown we can substitute the sample standard deviation  $s$  this introduces extra uncertainty since  $s$  is variable from sample to sample. So, we use the  $t$  distribution instead of the normal distribution.

**(Refer Slide Time: 02:16)**

**Confidence Interval for  $\mu$  ( $\sigma$  Unknown)**

*(continued)*

- Assumptions
  - Population standard deviation is unknown
  - Population is normally distributed
  - If population is not normal, use large sample
- Use Student's  $t$  Distribution
- Confidence Interval Estimate:

$$\bar{x} - t_{n-1, \alpha/2} \frac{S}{\sqrt{n}} < \mu < \bar{x} + t_{n-1, \alpha/2} \frac{S}{\sqrt{n}}$$

where  $t_{n-1, \alpha/2}$  is the critical value of the  $t$  distribution with  $n-1$  d.f. and an area of  $\alpha/2$  in each tail

30

What is the assumption for the  $t$  distribution population standard deviation is unknown population is normally distributed with the population is not normal use very large sample the

student t distribution the confidence interval is  $\bar{X} - t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{n-1, \alpha/2} \frac{s}{\sqrt{n}}$  less than  $\bar{X} + t_{n-1, \alpha/2} \frac{s}{\sqrt{n}}$  yes daughter by root n so this also came from this  $t_{n-1, \alpha/2}$  by 2 this has come from this expression. So, when you readjust that this equations then we can get the lower limit upper limit for the population mean.

**(Refer Slide Time: 03:07)**

**Margin of Error**

- The confidence interval,
 
$$\bar{X} - t_{n-1, \alpha/2} \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{n-1, \alpha/2} \frac{S}{\sqrt{n}}$$
- Can also be written as  $\bar{X} \pm ME$ 

where ME is called the margin of error:

$$ME = t_{n-1, \alpha/2} \frac{\sigma}{\sqrt{n}}$$

31

So, if it is  $-\bar{X} - t_{n-1, \alpha/2} \frac{s}{\sqrt{n}}$  it is a lower limit if it is  $+\bar{X} + t_{n-1, \alpha/2} \frac{s}{\sqrt{n}}$  it is an upper limit. So, this can be written as like our previous Z distribution  $\bar{X} \pm ME$  this margin of error so this  $\mu$  is nothing but  $t \frac{\sigma}{\sqrt{n}}$  previously it was  $Z \frac{\sigma}{\sqrt{n}}$  now it is  $t \frac{\sigma}{\sqrt{n}}$ .

**(Refer Slide Time: 03:41)**

**Student's t Distribution**

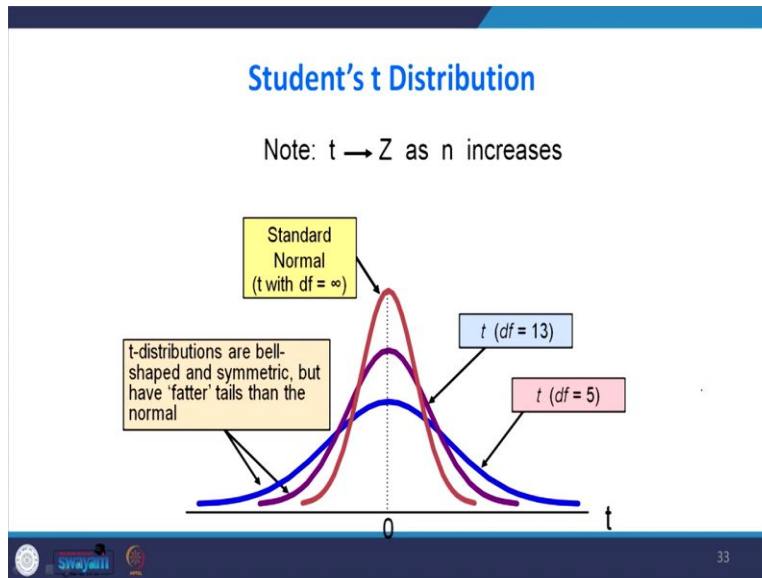
- The t is a family of distributions
- The t value depends on **degrees of freedom (d.f.)**
  - Number of observations that are free to vary after sample mean has been calculated

$$d.f. = n - 1$$

32

Student's t-distribution the t is a family of distributions because for every degree the degrees of freedom you will get you a different t distribution. The t value depends upon degrees of freedom number of observations that are free to vary after sample mean has been calculated nothing but degrees of freedom that is your  $n - 1$ .

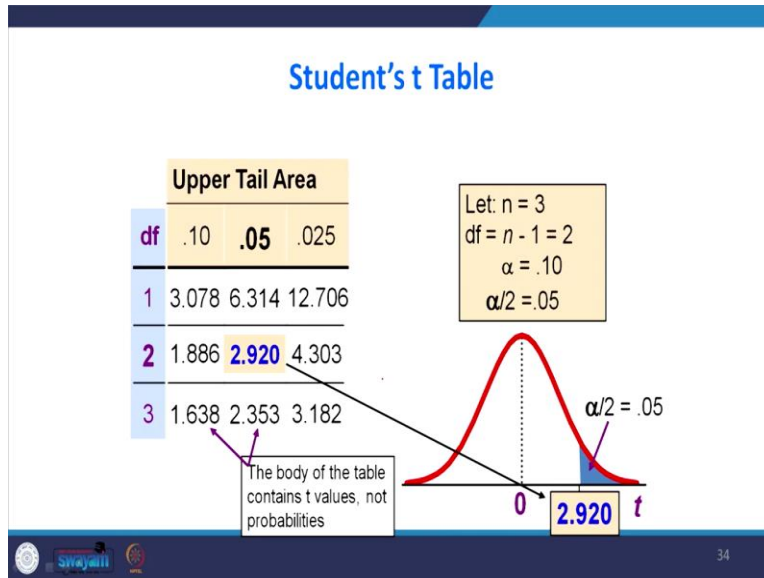
**(Refer Slide Time: 04:04)**



Look at this connection between t distribution and Z distribution, we start from the see the flatter one t distributions are bell shape and symmetric but have flatter tails than the normal. So, when the degrees of freedom EC initially 5 it is a flatter when the datas of freedom is 13 and so on you see when the degrees of freedom is infinity it is behaving like a Z distribution that is why in many **many** software packages see there would not be any option for doing Z test there will be option only for doing t test.

Because when the sample size increases for the t test so the behavior of Z distribution t distribution is same.

**(Refer Slide Time: 04:45)**



Then we look at the students of t table you said there is a difference between Z table and t table in Z table whatever value which is given is the area but in a t table you see that the area is given on the top say 0.05 the whatever value which is given inside the tea table is that is a critical value. For example if it is alpha by 2 is a 0.05 the corresponding t value is 2.920 so the body of the table contains t value not probabilities we should be very careful.

So, for example n equal to 3 and degrees of freedom is n - 1 to alpha equal to 5 then alpha by 2 0.05 so we are to see where the .05 in the column, column line when degree of freedom is 2 then we can see that is a 2.920.

**(Refer Slide Time: 05:41)**

### t distribution values

With comparison to the Z value

Confidence Level	t (10 d.f.)	t (20 d.f.)	t (30 d.f.)	Z
.80	1.372	1.325	1.310	1.282
.90	1.812	1.725	1.697	1.645
.95	2.228	2.086	2.042	1.960
.99	3.169	2.845	2.750	2.576

Note:  $t \rightarrow Z$  as n increases

A kind of a comparison between t values and Z values first we will go for this one resume so familiar for us when the confidence level is 95% see the Z value is 1.96 for different degrees of freedom you see that you see that when the degrees of freedom is 10 it is a 2.228 when is it 20 it is 2.086 see that when t equal to 30 the degrees of freedom is 2.0, so the value of t approaches Z when n increases you see that initially it is increasing so it is starting you know it is decreasing and finally it reaches 1.96.

This table explains whenever the degrees of freedom is increases we are getting Z is close to 1.96 means 96 for the t distribution.

**(Refer Slide Time: 06:44)**

### Example

A random sample of  $n = 25$  has  $\bar{x} = 50$  and  $s = 8$ . Form a 95% confidence interval for  $\mu$

- d.f. =  $n - 1 = 24$ , so  $t_{n-1, \alpha/2} = t_{24, 0.025} = 2.0639$

The confidence interval is

$$\bar{x} - t_{n-1, \alpha/2} \frac{S}{\sqrt{n}} < \mu < \bar{x} + t_{n-1, \alpha/2} \frac{S}{\sqrt{n}}$$

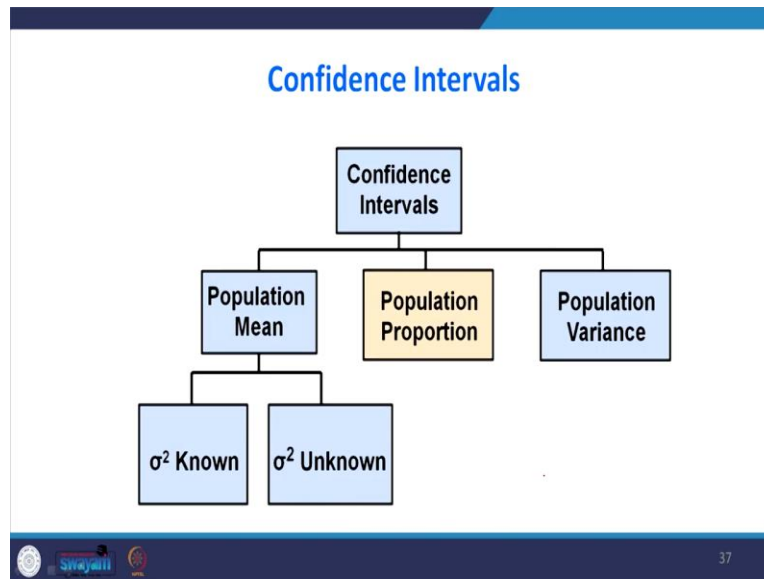
$$50 - (2.0639) \frac{8}{\sqrt{25}} < \mu < 50 + (2.0639) \frac{8}{\sqrt{25}}$$

$$46.698 < \mu < 53.302$$

Now we will see how to find out a confidence interval for your t distribution an example is a random sample of n equal to 25 as sample mean is 50 and sample standard deviation is 8 for me a 95% confidence interval for MU. The first one is we ought to go for degrees of freedom there are 25 so 25 – 1, 24 here confidence level is 95% so the significance level is 5% when they say it is a 5% because it is the upper limit lower limit we have to divide by 2 it is 2.5% when degrees of freedom is 24 alpha by 2 is 0.025.

When you look at the table the t value is 2.06 so you substitute X bar equal to 50 t equal to 2.06 yes is 8 and sample size is 25 so you are getting lower limit of forty 6.698 upper limit our upper limit of 53.302.

(Refer Slide Time: 07:51)



I will go to the next category finding the population proportion with the help of sample proportion.

(Refer Slide Time: 08:04)

**Confidence Intervals for the Population Proportion**

- An interval estimate for the population proportion ( P ) can be calculated by adding an allowance for uncertainty to the sample proportion (  $\hat{p}$  )

Confidence interval for the population proportion an interval estimate for the population proportion  $p$  can be calculated by adding and elements for uncertain uncertainty to the sample proportion that allowance is nothing but  $e$  were standard error.

(Refer Slide Time: 08:21)

## Confidence Intervals for the Population Proportion, p

(continued)

- Recall that the distribution of the sample proportion is approximately normal if the sample size is large, with standard deviation

$$\sigma_p = \sqrt{\frac{P(1-P)}{n}}$$

- We will estimate this with sample data:  $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Recall that the distribution of the sample proportion is approximately normal if the sample size is large then standard deviation is your Sigma P, Sigma P is root of P Q by n Q is nothing but 1 - P we will estimate this with the sample data. So, this is your sample standard deviation we can say standard deviation for sampling proportion root of P hat 1 - P hat greater by n.

**(Refer Slide Time: 08:50)**

## Confidence Interval Endpoints

- Upper and lower confidence limits for the population proportion are calculated with the formula

$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < P < \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- where
  - $z_{\alpha/2}$  is the standard normal value for the level of confidence desired
  - $\hat{p}$  is the sample proportion
  - n is the sample size
  - $nP(1-P) > 5$

To find out the lower limit upper limit of the population proportion we have to use the sample values because what will happen we may not know the population P value directly. If you know population P value what is the purpose of finding lower limit upper limit we know only the sample proportion so P cap - Z alpha by 2 root of P cap into 1 - P cap divided by n less than P P cap + Z alpha by 2 root of P cap into 1 - P divided by n.



So what is happening so with the help of our sampling proportions we can find out this value is a lower limit of our population proportion. This value is your upper limit of our sampling proportion you see that we with the help of sampling proportion very able to predict. There was a condition but the npq should be greater **greater** than 5 then only it can be approximated to normal distribution also.

**(Refer Slide Time: 09:52)**

**Example**

- A random sample of 100 people shows that 25 are left-handed.
- Form a 95% confidence interval for the true proportion of left-handers

41

An example a random sample of 100 people shows that 25 are left-handed for me a 95% confidence interval for the true proportion of left-handers, so this problem the P cap is 25 by 100 Z is 1.96 because 95% is confidence level all other P cap is given just you substitute this value and then you put plus decide minus you are getting the lower limit of population proportion is 0.1651. The upper limit of population proportion is 0.3349.

**(Refer Slide Time: 10:25)**

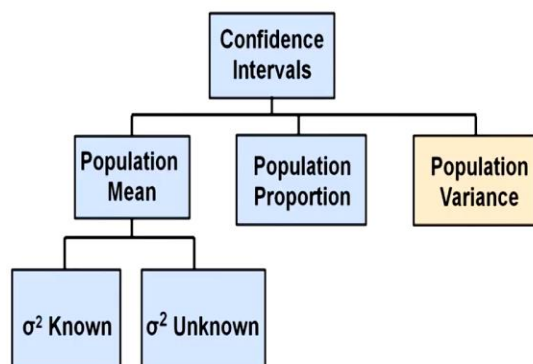
## Interpretation

- We are 95% confident that the true percentage of left-handers in the population is between  
16.51% and 33.49%.
- Although the interval from 0.1651 to 0.3349 may or may not contain the true proportion, 95% of intervals formed from samples of size 100 in this manner will contain the true proportion.

How to interpret this we are 95% confident that the 2% of left-handers in the population is between 16.51% and thirty 33.49% although the interval from 0.1651 to 0.3349 may or may not contain the true proportion 95% of intervals formed from the samples of size 100 in this manner will contain that is more important term. Another way you can say when you repeat this 100 times 95 times you can capture the true population proportion only 5 times you may not capture true population proportion.

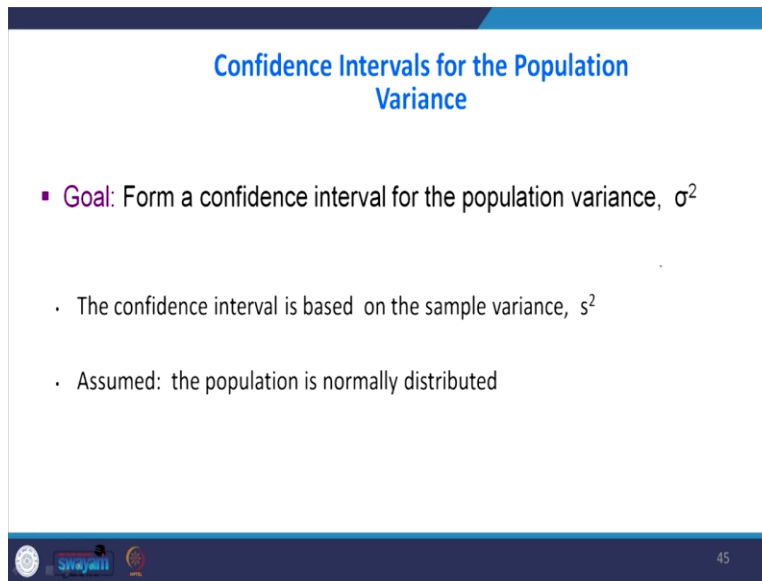
**(Refer Slide Time: 11:14)**

## Confidence Intervals



We will go to the last one how to predict the population variance. So, so far what we have seen we have predicted the population mean we have predicted the population proportion. Now we are going to predict population variance.

(Refer Slide Time: 11:35)



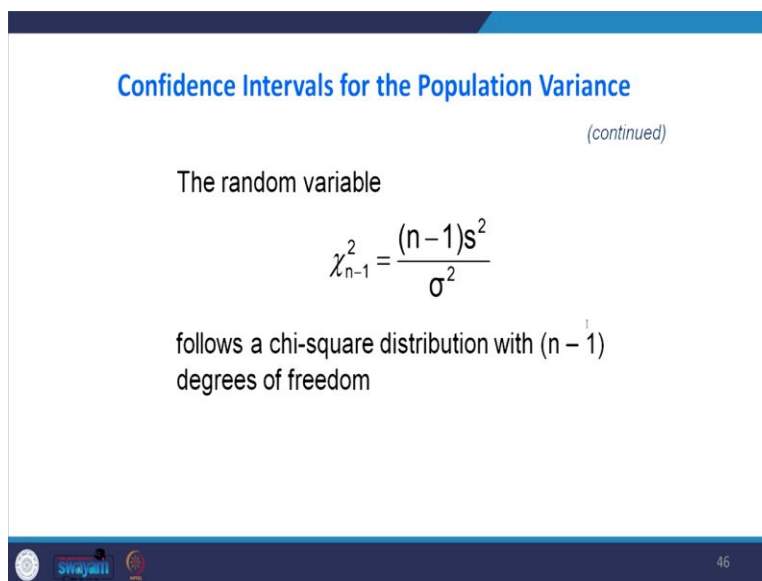
Confidence Intervals for the Population Variance

- **Goal:** Form a confidence interval for the population variance,  $\sigma^2$
- The confidence interval is based on the sample variance,  $s^2$
- Assumed: the population is normally distributed

45

The goal is to form a confidence interval for the population variance Sigma square. The confidence interval is based on the sample variance. So, what we are going to do with the help of sample variance we are going to predict the population variance interval. We are **we are** assuming the population is normally distributed.

(Refer Slide Time: 11:56)



Confidence Intervals for the Population Variance

(continued)

The random variable

$$\chi_{n-1}^2 = \frac{(n-1)s^2}{\sigma^2}$$

follows a chi-square distribution with  $(n - 1)$  degrees of freedom

46

We already we have seen that whenever there is a population is there if you take some sample from there when you plot the when you plot the sample variance that will follow Chi square distribution as I told you previously it will be like this. This will be  $n - 1$  s squared or y Sigma square. We are going to use this result when you readjust this right when you readjust this so

Sigma Square will be less than or equal to less than or equal to so, this will become Chi Square 1 - alpha by 2 here 1 - alpha by 2.

So, what is happening when you readjust **readjust** this equation when you readjust this equation for Sigma square you can find out the upper limit and lower limit of population variance.

**(Refer Slide Time: 13:13)**

**Confidence Intervals for the Population Variance**

The  $(1 - \alpha)\%$  confidence interval for the population variance is

$$\frac{(n-1)s^2}{\chi_{n-1, \alpha/2}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{n-1, 1-\alpha/2}^2}$$

47

Yes the same thing the  $1 - \alpha$  percentage confidence interval for the population variance is given by this one. You look at this the left hand side is alpha by 2 because what will happen when you look at the chi square distribution, we have given only the right side area when the right side area is alpha by 2, so what will happen here they will get to a bigger number. Suppose this was this value is over **over**  $1 - \alpha$  by 2.

So, here be a bigger number for example say 5 he will be smaller number when you numerator when you divide by bigger number will become smaller value that will become the lower limit of over variance. The numerator when you divide by a smaller value it will become bigger number that will become the upper limit of your population variance.

**(Refer Slide Time: 14:07)**

## Example

You are testing the speed of a batch of computer processors. You collect the following data (in Mhz):

Sample size        17  
Sample mean      3004  
Sample std dev    74

Assume the population is normal. Determine the 95% confidence interval for  $\sigma_x^2$



48

We will see you an example you are testing the speed of batch of computer processors you collect the following data, sample sizes 17 sample mean is 3004 sample standard deviation is 74 assume the population is normal determined 95% confidence interval for Sigma X bar square here Sigma square is nothing but lower limit upper limit of the sampling variance.

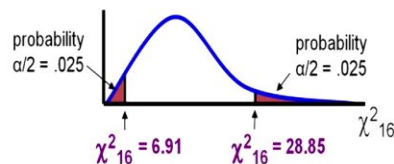
**(Refer Slide Time: 14:37)**

## Finding the Chi-square Values

- $n = 17$  so the chi-square distribution has  $(n - 1) = 16$  degrees of freedom
- $\alpha = 0.05$ , so use the the chi-square values with area 0.025 in each tail:

$$\chi_{n-1, \alpha/2}^2 = \chi_{16, 0.025}^2 = 28.85$$

$$\chi_{n-1, 1-\alpha/2}^2 = \chi_{16, 0.975}^2 = 6.91$$



49

So, n equal to 17 then chi square distribution has the  $n - 1$ , 16 degrees of freedom when alpha equal to 0.05 because it is we are finding upper limit lower limit we got 2 divided by 2 so 0.025 so when it is alpha by 2 it is 28.25 so what will happen this is the right side limit when you want to know the left side limit you do in the chi square table when area equal to  $1 - 0.025$  that area

you have to find out that probability when the degrees of freedom is 16 so corresponding value is 6.91.

**(Refer Slide Time: 15:21)**

**Calculating the Confidence Limits**

- The 95% confidence interval is

$$\frac{(n-1)s^2}{\chi_{n-1, \alpha/2}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{n-1, 1-\alpha/2}^2}$$
$$\frac{(17-1)(74)^2}{28.85} < \sigma^2 < \frac{(17-1)(74)^2}{6.91}$$
$$3037 < \sigma^2 < 12683$$

Converting to standard deviation, we are 95% confident that the population standard deviation of CPU speed is between 55.1 and 112.6 Mhz

50

So when you substitute this value 17 - 1 s square is 74 so this value is chi square value when it is alpha by 2 chi-square value if it is 1 - alpha by 2 you are finding the lower limit is 3037 and upper limit is 12 683 converting the standard deviation we are 95% confident that the population standard deviation of CPU speed is between when you take square root of this between 55.1 and 112.6.

**(Refer Slide Time: 15:55)**

**Finite Populations**

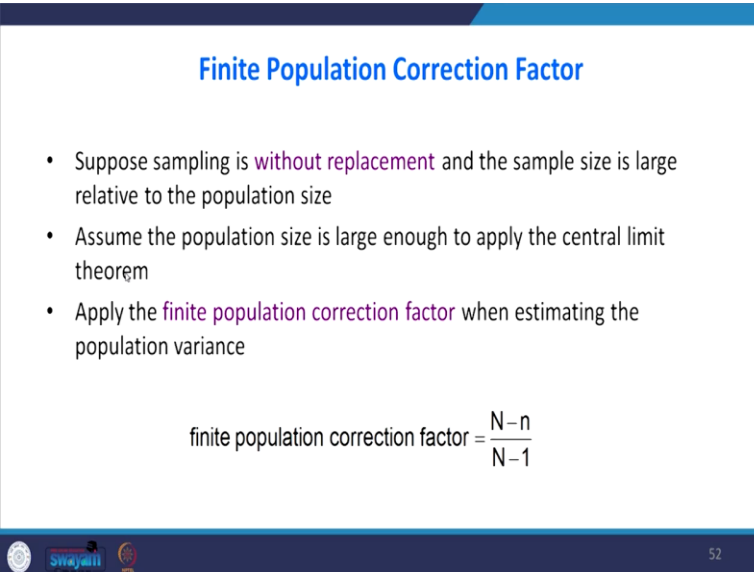
- If the sample size is more than 5% of the population size (and sampling is without replacement) then a **finite population correction factor** must be used when calculating the standard error

51

So, far we have assumed that the infinite population sometime there is a finite to population. Finite population is when the when the when the number of element in the population is small, if the sample size is more than 5% of the population size and sampling without replacement then the finite population correction factor must be used in calculating standard error. So, we have to add this correction factor when we go for a finite population.

Then finite population rhythm when the sample size is more than 5% and being you go for without replacement.

**(Refer Slide Time: 16:34)**



The slide is titled "Finite Population Correction Factor" in blue text. It contains three bullet points: "Suppose sampling is without replacement and the sample size is large relative to the population size", "Assume the population size is large enough to apply the central limit theorem", and "Apply the finite population correction factor when estimating the population variance". Below the bullet points is the formula: "finite population correction factor =  $\frac{N-n}{N-1}$ ". At the bottom left, there are logos for "Swayam" and "UPEA". At the bottom right, the number "52" is displayed.

Suppose sampling is without replacement and the sample size is large relative to the population size we should go for finite population correction factor. Assume the population size is large in our to apply the central limit theorem. So, apply the finite population correction factor when estimating the population variance, so, this factor  $N - n$  total by  $n - 1$ . So,  $N$  is population size is  $n$  is sample size.

**(Refer Slide Time: 17:06)**

## Estimating the Population Mean

- Let a simple random sample of size  $n$  be taken from a population of  $N$  members with mean  $\mu$
- The sample mean is an unbiased estimator of the population mean  $\mu$
- The point estimate is:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

Let the simple random sample of  $n$  be taken from the population of  $n$  members with  $\mu$  the sample mean is unbiased estimator of the population mean  $\mu$  then the point estimator is  $\frac{1}{n} \sum X_i$  there is no problem for sample mean when you are going for sample variance we have to add this correction factor that is this correction factor has to be added. If the sample size is more than 5% of the population size and unbiased estimator for the variance of the sample mean is  $s^2$  by  $n$  you have to multiply this.

**(Refer Slide Time: 17:42)**

## Finite Populations: Mean

- If the sample size is more than 5% of the population size, an unbiased estimator for the variance of the sample mean is

$$\hat{\sigma}_{\bar{x}}^2 = \frac{s^2}{n} \left( \frac{N-n}{N-1} \right)$$

- So the  $100(1-\alpha)\%$  confidence interval for the population mean is

$$\bar{X} - t_{n-1, \alpha/2} \hat{\sigma}_{\bar{x}} < \mu < \bar{X} + t_{n-1, \alpha/2} \hat{\sigma}_{\bar{x}}$$

So  $100(1-\alpha)\%$  confidence interval for the population mean is  $\bar{X} \pm t_{n-1, \alpha/2} \hat{\sigma}_{\bar{x}}$  square.

**(Refer Slide Time: 17:51)**



## Estimating the Population Proportion

- Let the true population proportion be  $P$
- Let  $\hat{p}$  be the sample proportion from  $n$  observations from a simple random sample
- The sample proportion,  $\hat{p}$ , is an unbiased estimator of the population proportion,  $P$



55

So, this is applicable for population proportion also when the population proportion is population when we are going to predict the population proportion when the sampling proportion is larger and the population proportion is finite then you have to add another correction factor. Let the true population proportion  $P$  let  $\hat{p}$  be the sample proportion from  $n$  observation from the simple random sample.

The sample proportion  $\hat{p}$  is unbiased estimator of the population proportion  $P$  so here also we have to add this  $\frac{P}{n-1}$  as a correction factor all others are remaining same.

**(Refer Slide Time: 18:37)**

## Lecture Summary

- Introduced the concept of confidence intervals
- Discussed point estimates
- Developed confidence interval estimates
- Created confidence interval estimates for the mean ( $\sigma^2$  known)
- Introduced the Student's  $t$  distribution
- Determined confidence interval estimates for the mean ( $\sigma^2$  unknown)



57

Now we will summarize what we have done so far. In this class we will summarize what we have done so far in this lecture we have created a confidence interval estimate for the proportions then we have created a confidence interval estimate for the variance of a normal distribution. For each proportion and variance estimations we all you taken a numerical example to solve the problem to understand this concept of parameter estimation, thank you.