

Handling Large-Scale Unit Level Data Using STATA
Professor Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee
Lecture 09
Sample Size Determination-I

Welcome learner once again to the MOOC module on Handling large-scale unit-level data with STATA. So, this is our week 2 where we have been discussing collection of unit-level data, And, the title of today's lecture is on understanding sample size and its determination. In the last lecture also, we already started with sample size determination.

So, today we are going to complete, how the sample size is determined based on various context. And in the last lecture particularly, we discussed largely on what do you mean by sample size? And what are the different approaches of getting sample. And what are the different types of sample, we have already discussed, probability and nonprobability majorly. In a nutshell, let me mention that, in case of nonprobability sampling, usually the sample size is less. Whereas, in case of probability one we do require more sample. It is still very abstract, because a clear-cut idea of sample size is very very important. Without that there will be large number of questions on the researchers.

So, we have to statistically prove that sample size is of this much. And, further validity or precision of the work is expected. Without that, it is questionable. So, we have already discussed that, let me go into the deeper aspects of understanding sample size. Therefore, the lecture name is sample size determination. This is a continuation of the previous lecture.

(Refer Slide Time: 02:36)

The screenshot shows a web page titled "BASICS" for a sample size calculator. The URL is <https://select-statistics.co.uk/calculators/sample-size-calculator-population-proportion/>. The calculator is titled "Calculator" and has four input fields, each with an information icon (i):

- What margin of error do you need? (5%): 5%
- What confidence level do you need? (95%): 95%
- How big is the population? (10000): 10000
- What do you believe the likely sample proportion to be? (50): 50%

The result is displayed in a red box: "Your recommended sample size is 383".

We have already shown this website, if you just have a click on this link, you will be directed to the calculator for getting the sample size. Here what I wanted to refer, for all of you is, which indicators are important for sample size determination. How you could be able to assure yourself, that your sample size is perfectly fine.

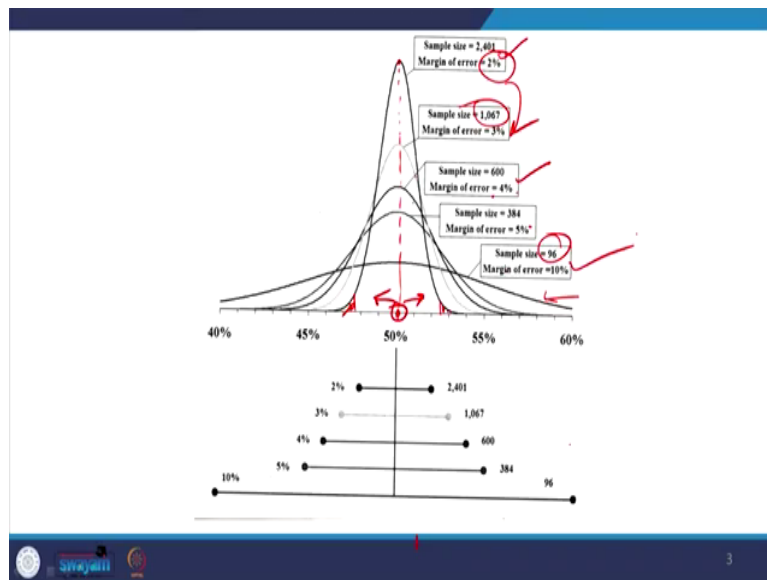
It may be the case that you have constrained your sample to a limited number. But that number might be attached with huge errors. So, errors may be of different variety, different types. So, here I am referring to margin error. Margin error and confidence, with which level of confidence, you can say that your margin of error is of this much. So, your result is of this much. So, in this particular example, we have said the margin of error, that you can manually set here at 5 percent level, which is usually acceptable, till 10 percent is usually acceptable and statistically proved in different literature and different papers.

So, and the confidence level usually of 95, 90, and 99. So, let us be 95 level of confidence, and how big is the population. If your population size is known, if you have a finite population, then this link really works better that is if you set 10,000 here is the population. this will boil down to a sample size of 383, with the fact that you have certain pre-defined assumptions, that is, what is the sample proportion? How much proportion you are supposed to consider out of the total population? Who has the probability of how much? If that, the selection of those cases are of 30 percent somewhere some estimates are important.

If you do write estimates, then usually your sample size is going to be very good. But when you are highly confused about the population you do not know the near about probability of

their inclusion or exclusion. So, in that case, you have to assume 50 percent. So, 50 percent based on the 50 percent exemption show our result based on this calculator gives 383. So, 383 looks at 10,000 of the population. Now, it has narrowed down to 383. So, sampling is important. And what are the validity of this 383? How it is important? How do I believe that this is important? We have certain examples to validate. So, let us go by that.

(Refer Slide Time: 05:35)



Here is a picture we have already shown in the last lecture, but I did not explain much because of shortage of time. Here what I refer, given the distribution is normal, we are not even referring to non-normal distribution, we are assuming that normal distribution has symmetric format, but even if it is symmetric, but it might be leptokurtic, it might be platykurtic, or it might be mesokurtic.

So, there are different distribution. Higher the distribution within the normal distribution format, that means, higher the stretch, higher the distribution, some variations are expected, higher variations are expected. If it is leptokurtic, the mean is not deviated here, what I refer, let me pinpoint here to the 50 percent level.

So, I am just drawing the 50 percent or an orthogonal direction to the base, which falls at 50 percent. I am just drawing here for you for the clarity. If the distribution is of leptokurtic or the altitude of that is much higher, as compared to other. In case of platykurtic the distribution is like, this is the extreme form of kurtosis. So, if it is a leptokurtic in that case, the deviation from the central mean is very less.

So, since the deviation from the central mean is very less, the errors are expected to be very less. The margin error, that means, if you set certain alpha percentage, alpha as the type 1 error, as I already refer, what do you mean by it again? Rejecting a null hypothesis, when it is true. We have collected samples from a population, from a group, from the larger context.

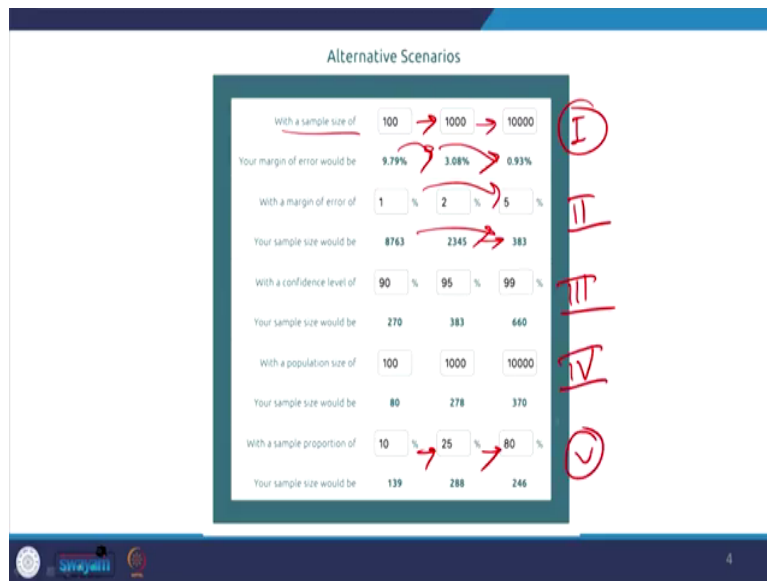
We have picked up some sample, based on the sample we determined that you are rejecting the entire population, what do you mean by that we are simply rejecting the entire bundle, based on the sample. It may be the case that, whatever you have picked up some errors are contained in your sample. So, our sampling process with a limited sampling direction may not be justifying the entire population, there must be repeated many samples.

See, if each of the sample are giving a central mean, having less errors, or less deviation from the central mean. All your distribution is nearby the central mean, your mean is not deviated much, that means your sample is correct. Where we are referring to precision, precision of the analysis, higher the precision higher is also the confidence.

So, what is important here, so I was referring, how the distribution is spreaded. So, in this particular example, if it is leptokurtic, margin error is, look at margin of error is very very less because the alpha percentage is expected to be very very less because the distribution is very symmetric, it is symmetric but confined to a central mean. Whereas, when the, it is possible when the sample size is more. When the sample size is more, you have collected more sample, or the frequency of sample is also more. Number of samples is also more; you can taste the central mean. So, the error is expected to be very very less.

So, the margin of error is 2 percent only in that case. When the sample size is lesser as compared to the first one, your margin error increases, because the distribution is now getting less chaotic. And when the sample size is 600 margin error increases. So, the margin of error and sample size are inversely related to each other. And this is, when it is 96, sample sizes only 96, margin are increased to 10 percent. As per example, this is mentioned in the diagram here very clearly, you can follow it up and you will find out the correct interpretation.

(Refer Slide Time: 10:27)



From the same website, we can crosscheck the margin error as well as the role of confidence level. When the sample size is higher in the first example, your margin of error decreases. You can compare with the increase in sample size, as per the first set of examples. In the second case, with a margin error of 1 percent, you require an 8763 sample. When you have the freedom to go for a little higher margin of error, in some cases in order to restrict your budget, you may accept that 5 percent is fine, till 10 percent is generally acceptable.

So, you can stretch your margin error to little higher level, so that sample size you require is a little lesser. So, second one is just the opposite, but it is important so far as sample size is concerned. In the third case, we are referring to the confidence level, the confidence level matters how confidently we deal with our both the direction sample size.

In order to make your confidence level higher, you are supposed to collect more sample. So, samples are positively linked. And so, if your population itself is very high, you are bound to go for higher sample size. So, in all the case sample sizes referred. If with the same sample proportion of 10 percent, your sample size will be 139. If your sample proportion or probability of inclusion is 80 percent. If it is 20 percent, there are different brackets given we can calculate.

So, you can find different sample size based on sample proportions. So, we will take use of all those in our equation, the equation will clarify what sample proportion and why these rates are different. Somewhere it is higher somewhere it is lower, we will find out from a mathematical equation.

(Refer Slide Time: 12:50)

APPROACHES FOR DETERMINING THE SIZE OF THE SAMPLE

THE APPROACH BASED ON PRECISION RATE AND CONFIDENCE LEVEL

- to specify the precision of estimation desired and then to determine the sample size necessary to insure it.
- mathematical solution
- frequently used technique
- The limitation – it does not analyze the cost of gathering information.

One of the approaches for determining the sample size, is precisely based on the confidence level, as well as precision rate, the rate at which the precision is made to specify the precision of estimation desire and then to determine the sample size necessary to ensure it, we do require mathematical solutions, as well as frequently used techniques.

So, some mathematical solution is needed to justify the sample size correctly. And also, some frequent techniques, standard techniques are required like as I told you, there are some conventional methods, some convenient sampling method, purposive sampling method, snowball sampling method, non-probability method where we know certain frequent techniques.

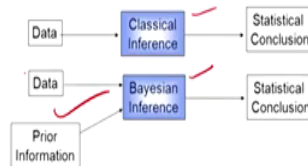
So, for those reasons, we can able to reduce the sample size, or our precision is expected to be much higher. And if we have hired a very good local investigator, having very good knowledge about the local areas, so you need not collect more sample because, the local area investigator having local knowledge could infer many better information.

So, what are the limitation in this case? The limitation is, it does not analyze the cost of gathering information. So far as precision rate in confidence is concerned, cost of the information is not discussed. When the cost is discussed, it gets complicated usually we refer in the context of Bayesian statistics. So, the Bayesian statistics usually take care of cost aspect also. Though, it is very complicated statistically.

(Refer Slide Time: 14:59)

THE APPROACH BASED ON BAYESIAN STATISTICS

- uses Bayesian statistics to weigh the cost of additional information against the expected value of the additional information.
- It is theoretically optimal
- it is seldom used because of the difficulty involved in measuring the value of information



Bayesian statistical analysis incorporates a **prior probability distribution** and **likelihoods** of observed data to determine a **posterior probability distribution** of events.

So the Bayesian approach is used to weigh the cost of additional information against the expected value of the additional information. So, as I already mentioned, we do require cost of additional information, which is very very essential. And it is theoretically optimal, first important points, second, it is seldom used because of the difficulties involved in measuring the value of the information.

Look at the classical way of inference, and in the second one is Bayesian inference. So, in the statistical inference, we can also be able to conclude, and in this case also you can conclude systematically, but Bayesian inference does require prior information, this is quite important. So, the Bayesian statistical analysis incorporates a prior probability distribution and likelihood of observed data, to determine a posterior probability distribution of the events.

(Refer Slide Time: 16:11)

**DETERMINATION OF SAMPLE SIZE THROUGH THE APPROACH
BASED ON PRECISION RATE AND CONFIDENCE LEVEL**

□ **Sample size when estimating a mean:** The confidence interval for the universe mean, μ , is given by

$$\mu = \bar{X} \pm z \frac{\sigma_p}{\sqrt{n}}$$

acceptable error $e = z \cdot \frac{\sigma_p}{\sqrt{n}}$

$$n = \frac{z^2 \sigma^2}{e^2}$$

where
 N = size of population
 n = size of sample
 e = acceptable error (the precision)
 σ_p = standard deviation of population
 z = standard variate at a given confidence level.

Let us analyze the mathematical approach of understanding the precision rate and confidence level and so on. We can able to define the right value or number of the sample, the sample size can be determined based on the standard formula. let me proceed to an equation from the beginning, as we all know that, first of all, any data we have, our duty is to convert the data to a standard normal distribution, standard normal distribution looks like this.

(Refer Slide Time: 16:51)

$\hat{z}_i = \frac{\sum_{i=1}^n x_i - ne}{\sqrt{n}}$

c.I $(\hat{z}_i) = \bar{X} \pm z \frac{\sigma}{\sqrt{n}}$

acceptable error $e = z \cdot \frac{\sigma}{\sqrt{n}}$

$$n = \frac{z^2 \sigma^2}{e^2}$$

where
 N = size of population
 n = size of sample
 e = acceptable error (the precision)
 σ_p = standard deviation of population
 z = standard variate at a given confidence level.

So, let me draw here, it looks like this. So, why standard normal distribution is important? Because it simplifies the data. It centres the mean across the distribution, and accordingly, we can have lots of projection. And it simplifies the data and it makes the data symmetric. If the distribution is standard normal, then the data gets symmetric. So, how to get a standard

normal distribution? Usually denoted by Z . So, Z is equal to X minus μ divided by the standard deviation. So, X minus μ by the standard deviation.

If you have a certain finite population or infinite population, there are two approaches of understanding this. σ is quite important to read here because it is not just σ it is the actual formula of Z , Z_i is equal to sum of i varies from 1 to N , X_i and this is sum of μ divided by N .

So, sum of i varies from 1 to N , X_i minus μ divided by standard deviation. So when we take it, this is called \bar{X} and in sample mean minus population means, μ stands for population mean and the sum of the standard deviation is divided by N , square root of N , it is σ squared, σ square divided by N .

So, it will be equal to σ square root of N if you take the right approach of it, we can find out accordingly. if I just convert the equation in terms of mean μ is equal to \bar{X} because I have taken a sum of i varies from 1 to N divided by N into X_i . So it will be \bar{X} plus, if I am interested in calculating the confidence interval CI. I have already discussed that confidence interval are the intervals from the central mean. If we said the confidence interval at, this is 99 percent, 95 percent or it is at 90 percent confidence level. When it is 90 percent it covers 99 percent of the information.

And usually, the errors are very very less when the confidence level is very high the precision rate is also quite weightier, and the error margin error is very very less. In this case you need to define the confidence interval of the μ . So, it will be plus-minus of standard deviation because the distribution varies from the \bar{X} , \bar{X} plus 1 standard deviation, plus 2 standard deviations, plus 3 standard deviations depending upon the distribution we have. So, what we have? If I just simply multiply, here it will be Z_i times J times σ square root of N .

(Refer Slide Time: 20:29)

**DETERMINATION OF SAMPLE SIZE THROUGH THE APPROACH
BASED ON PRECISION RATE AND CONFIDENCE LEVEL**

□ **Sample size when estimating a mean:** The confidence interval for the universe mean, μ , is given by

$$\mu = \bar{X} \pm z \frac{\sigma_p}{\sqrt{n}}$$

acceptable error $e = z \cdot \frac{\sigma_p}{\sqrt{n}}$ where

$$n = \frac{z^2 \sigma^2}{e^2}$$

where
 N = size of population
 n = size of sample
 e = acceptable error (the precision)
 σ_p = standard deviation of population
 z = standard variate at a given confidence level.

7

So, let me move to the original equation it is clearly given here. When we have a sample size and the population is not known what is the total population, when we are estimating a mean the confidence interval for the universe means that is μ is equal to, we are estimate the confidence interval \bar{X} bar plus minus, we have already discussed this plus minus here, plus minus Z times sigma by square root of N this is I have already derived in the plane page. So, from here if I feed the equation. So, what is the error here?

Your population mean or the universal mean is centred around the sample mean that is \bar{X} bar with error of Z times sigma of square root sigma of 1 square root of N . So, this is the extent of error we have, the margin error or the precision rate we are going to find out through this.

This is correctly mentioned in this particular example. So, how we could do it? If we just adjust the equation. So, here is given Z times sigma open square root of N.

N is the size of the population and small n is the size of sample and e is acceptable error that is also called precision. So, N is the population but population N is not given as I already mentioned if it is not known, this is the formula. What is my sigma P here? Standard deviation of population is called sigma P. So, what is required in this equation, we should have certain information about the standard deviation, how the population distributed? What is the approximate value of the distribution of that particular area or the population?

So, standard deviation of the population not the sample, and Z standard variate or standard variate at a given confidence level. So, if it is further adjusted, this equation is given here, and if I calculate further it is, I will take e here, so square root of N is equal to J times sigma divided by e. So, N boils down to Z square, it is already given times sigma square divided by e square. So, e is calculated.

So, e is calculated and this is what the sample size small n stands for the sample size. So, sample sizes correctly estimated. If this information is there you have a standard Z table usually available. So, the Z table is available and some idea of population distribution is given and sample minimizing error is calculated and it is known, then we can find out the sample size.

(Refer Slide Time: 24:09)

In case of finite population, the confidence interval for the universe mean, μ , is given by

$$\bar{X} \pm z \frac{\sigma_p}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

acceptable error

$$= z \frac{\sigma_p}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

$$n = \frac{z^2 \cdot N \cdot \sigma_p^2}{(N-1)e^2 + z^2 \sigma_p^2}$$

Handwritten derivations:

$$\Rightarrow \mu = \bar{X} \pm z \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

$$\Rightarrow \sqrt{n} = z \frac{\sigma}{e} \sqrt{\frac{N-n}{N-1}}$$

$$\Rightarrow n = \frac{z^2 \sigma^2}{e^2} \frac{N-n}{N-1}$$

$$\Rightarrow \frac{n}{N-n} = \frac{z^2 \sigma^2}{(N-1)e^2} \Rightarrow n = N \left(\frac{z^2 \sigma^2}{(N-1)e^2 + z^2 \sigma^2} \right)$$

If we have a finite population and population is known, the correct equation for it is, we just need to subtract the net population after the total population. So, the net population divided

by capital N minus 1 that is total population, population minus 1 and 1 is deducted because of the degrees of freedom constant due to the sampling distribution. Seeing the sample is considered and sample due to the degrees of freedom reduced by 1 we can adjust with the degrees of freedom by dividing to it.

So, since it is divided, the total population is known and the probability of not getting selected with respect to the total population is given here. So that ratio is also multiplied here, the probability of not getting selected is also multiplied here and if that is there accordingly you can calculate the margin error or the acceptable error and accordingly we defined the sample size, how to get it?

You first convert that equation, how to convert that equation? We are interested in calculating, this is μ , is equal to \bar{X} plus minus Z sigma square root of n . We are multiplying the proportion of non-inclusion and that is square root of capital N minus n divided by N minus 1.

So, our error term, margin error is explained by this. I am interested in calculating small n . So, small n can be taken to this side. So, first step is simply take a square root of n here and your Z is here, it is sigma divided by e , this is N minus small n to N minus 1 the square root is given here. The N is defined as then Z square sigma square divided by e square.

So, all the terms are avoided with square root this will be N minus 1. Still there are small n contents on the right-hand side. So, if I simply take here, so what I will do, I will multiply this with N , I can take like these capital N minus n , J squared. So, rest will be their Z square sigma square divided by capital N minus 1 times e square.

So, this can be presented there, you simply subtract. So, you take multiply right hand by capital N minus n . So, next step will certainly be N is equal to, So, N times of this, I am just multiplying this and minus small n times of this and this is here. What we will do? You take small n to the left-hand side and start solving it you will certainly get the final equation is Z squared times N multiplied by this and this is the final equation. Once we have defined the correct N when the finite population is there, we can able to solve many equations.

(Refer Slide Time: 28:25)

Illustration

Determine the size of the sample for estimating the true weight of the cereal containers for the

universe with $N = 5000$ on the basis of the following information:

1) the variance of weight = 4 ounces on the basis of past records.

(2) estimate should be within 0.8 ounces of the true average weight with 99% probability.

1 example is before us, how to determine the sample size for estimating the true value of the cereal containers for the universe with a population of 5000? On the basis of the following information the variance of the weight is 4 ounces and then the estimates should be within 0.8 ounces of the true average weight with 99 probabilities.

So, what is given the e given Z value is given Z , Z the confidence level is given, confidence level is of 99 percent and e value that is 0.8 and the variance is also given, variance not the standard deviation. So, variance here is 4. So, all the information is given we can simply pour this information and find out the correct sample size.

(Refer Slide Time: 29:18)

Solution:

$N = 5000$

$\sigma_p = 2$

$e = 0.8$

$z = 2.58$ (for 99% probability)

$$n = \frac{z^2 \cdot N \cdot \sigma_p^2}{(N-1)e^2 + z^2 \sigma_p^2}$$

$$= \frac{(2.57)^2 \cdot (5000) \cdot (2)^2}{(5000-1)(0.8)^2 + (2.57)^2 (2)^2}$$

$$= \frac{132098}{3199.36 + 26.4196}$$

$$= \frac{132098}{3225.7796}$$

$$= 40.95 \approx 41$$

n = 41

□ Z-score is 1.645 for 90%, 1.96 for 95%, 2.58 for 99%

$$n_0 = \frac{z^2 pq}{e^2}$$

$$0.025 = 1.96 \sqrt{\frac{0.3 \times 0.7}{n}}$$

$$\frac{0.3 \times 0.7}{n} = \left(\frac{0.025}{1.96}\right)^2 = .0001627$$

$$n = \frac{0.3 \times 0.7}{.0001627} = 1291$$

So we would need a sample of about 1300 students at 90% confidence interval, for which $z = 1.645$

$$0.025 = 1.645 \sqrt{\frac{0.3 \times 0.7}{n}}$$

we can quickly find **n = 909**

population N is given, the population variances is 4. So, the standard deviation will be 2. Margin of error we have already defined that is 0.8 was given and the Z value for the 99 percentage, 99 level of confidence. If you go by the Z table, the Z table will give you a value of 2.57 there are another one, let me just show you here. for the Z table. For 99 level it is 2.58 for 95 percent is 1.96. This is a standard Z value if the confidence level is given. So, about 90 percent confidence level it is of 1.645.

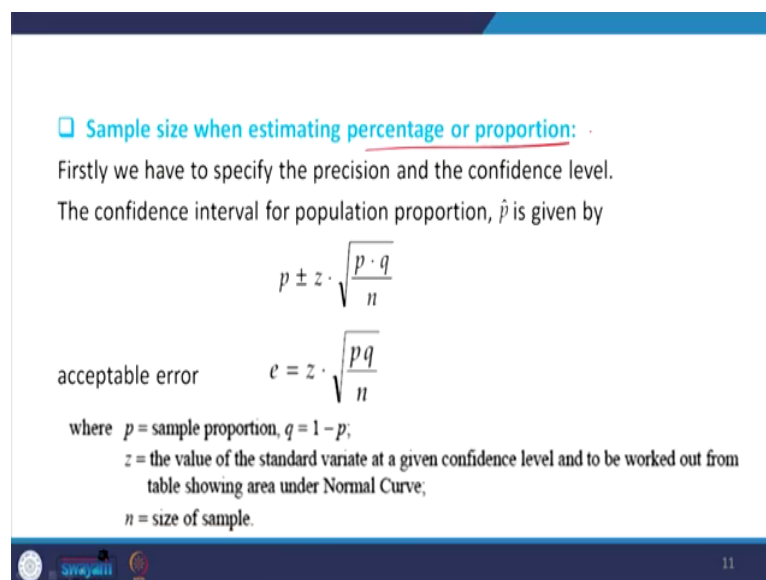
How could we able to understand these correctly? If we interpret 1.645 standard Z value or a different level 1.645, 1.96 and 2.584, 99 percent level of confidence. If I use that for 2.58 or 2.57 is given it is 2.58. So, for Z of 2.58 if I interpret every word. This is 2.58 capital N is 5000, sigma square is 4 that is sigma is 2, and N minus 1, e value is 0.8 then rest are given.

So, if you just put all the values these boils down to a sample size of 40.95 that is precisely 41.

How could I interpret this out of 5000 population size? We can derive the sample size or sample sizes determined is 41 it is minimized, but so, many information must have been given there are some certain standard determining, like in a margin error, some estimates are required, Z value, confidence level are generally set at 90 percent or 95 percent and 99 percent level.

So, standard Z table value has to be remembered or should be checked in the table, these are from a variance is very tricky, it is not that easy to determine. So, if some information regarding the population variance and population size is important, so far as this method is there to apply.

(Refer Slide Time: 32:06)



□ **Sample size when estimating percentage or proportion:**

Firstly we have to specify the precision and the confidence level.
The confidence interval for population proportion, \hat{p} is given by

$$p \pm z \cdot \sqrt{\frac{p \cdot q}{n}}$$

acceptable error $e = z \cdot \sqrt{\frac{pq}{n}}$

where p = sample proportion, $q = 1 - p$;
 z = the value of the standard variate at a given confidence level and to be worked out from table showing area under Normal Curve;
 n = size of sample.

So, this method is a little complicated. Some new methods are suggested based on other information. What is important to be noted here, is that we are supposed to follow certain techniques when we have certain other information although information are not given only sample proportions are given, percentages are given, or proportions are given, if only limited information is given some specific answers or equations are, therefore, referred to what do you mean by that like certain population proportion information is given either success or failure, probability of success and failure information is given to you.

So, in that case the distribution is not normal, it is not a normal distribution, when a population is not normal, you cannot convert that population to a normal distribution and it

will be a poisson distribution type P and Q that is probability of success and probability of failure $1 - P$ that is Q is given. So, if it is a poisson distribution then it is not that easy to apply the same rule we have derived.

So, follow the standard, created a given confidence level and to be worked out from the table swing area under normal curve, when you convert that through normal curve, it is very challenging, and accordingly a different formula is applied. If P and Q is there, standard table or the standard Z distribution is tricky to determine, if that is determined then we can define accordingly. So, I think I have already exceeded the limit of 30 minutes today, what I will do, I will keep the rest of the contents for the next lecture.

And because we have very interesting part, let me just have an overview of it will carry forward this content in the next class because those who do ground-level research, do require this information for better calculation. So, these are still there, we will discuss in the next class. Let me stop here. Thank you.