

Handling Large-Scale Unit Level Data Using STATA
Professor Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee
Lecture 10
Sample Size Determination-II

Welcome friends once again to the MOOC module of NPTEL on Handling Large Scale Unit Level Data with STATA, STATA is the software, we are referring and from the next week onwards, we are supposed to start with the use of STATA and STATA software will be introduced in the next week. So, have patience on it and you will certainly get the exposure of how to use STATA, without understanding the background of STATA it is not that good to immediately introduce STATA.

So, on third week onwards as per our schedule we have planned to take you to a journey through STATA. As a continuation to my previous lecture on Sample Size Determination. This lecture is titled as Sample Size Determination-2 and so, I do not have a specific introduction slide and I will start from the previous lecture.

(Refer Slide Time: 01:36)

**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.

The previous lecture content was on understanding sample size through two important parameters, one is precision rate or the margin error and second one is confidence level. Confidence level is very important. So far as tabulated value is concerned, what do you mean by tabulated value? Tabulated value is the standard value of the Z value or T distribution value or F distribution value mentioned in a tabular form and that has been already accepted

by larger researcher and presented in a systematic table. So, when an individual researcher do study on a particular context, have to take those tabulated value as the benchmark.

And benchmark levels should not be exceeded or undermined by any percentage. So, if we are deviating from the benchmark, there are some limits of division. So, the limits are given already either it is at 99 percent or 90 percent or 95 percent or any percent in between. So, we already discussed this and let me come back to the presentation on understanding sample size through the precision rate and confidence level. So, there are two approaches when the population size is known and the population size is not known.

What do you think, what is in reality true? In reality population size is not known, in many cases and in large number of surveys, we do not know the population size except a few, except a few experimental designs, we know this is our population and we know who are our stakeholders.

So, then only some planning can be made like in any Panchayat we know how many populations are there, if it is targeted at a village level only and you do an experiment on it, maybe on nutrition level your total population is known for sure within a village and total number of persons who are suffering from that particular disease also known if you are experimenting on a particular disease.

But what is going to be your outcome is not known. So, outcome we are not emphasizing here, we are referring to your population number and a sample number. If population is that much, then how much sample you are supposed to take? If population is not known still, we can also discuss about sample size. But just sample size, no single researcher in the world can be able to give you the idea of sample size. I think there are some misleading statement given by some faculty or the researcher in different platforms, like they simply say, the sample size should be 30 percent, sometimes they say it should be 10 percent, it should be 2 percent, there are different numbers they say.

And there are some misleading statements like, let me start with 30 percent, if it is 30 percent, still for a researcher covering 30 percent is not that easy. So, what for sure if it tends to more than 30, we generally refer not a T distribution it is a Z distribution. So, if it is less than 30 then I refer to the student T distribution, we will discuss some of those distribution later, but at this moment, I am just referring to you for your knowledge related to distribution and determining size. So, sample size there is no standard rule for sure, if somebody is telling you this is my population tell me how much sample should I take? For a PhD student, always it is

asked by the faculty or the board that how much sample you have taken and why you have taken?

So, in order to say why, you have to explain all those parameters or statistics I am referring, so, we require precision rate, we require confidence level, we require variance of the population, we require population size, four-five information are always required in order to derive the sample size, but there are some other techniques proposed by YAMANE and by Cochran, they suggested some alternative approaches given a limited information. So, Cochran and YAMANE is largely used by the researcher, we have that in our content.

So, as I mentioned, we have already explained this, I am not spending more time on it, let me just put a tick mark on it, just by emphasizing the fact that this is our margin error. And margin error is also called acceptable error, acceptable error is Z times the square root of the deviation, deviation upon square root of sample size. And this is written as e equals Z times sigma square root of N, you put this there. So, accordingly we can convert that equation and we will derive N, that I will discuss I am not putting more time on it.

(Refer Slide Time: 07:19)

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

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

Handwritten derivations:

$$e = z \frac{\sigma_p}{\sqrt{n}} \sqrt{\frac{(N-n)}{(N-1)}}$$

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

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

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

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

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

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

Let me proceed to the understanding once again on the finite population. If it is a finite population, if the population size is known as a ratio to the confidence interval, we are supposed to multiply the proportion of non-inclusion out of the total population, who are not included divided by who are the total population minus 1. If you are including that, the rationale behind this I have already discussed in the last lecture, but more statistical derivation can be followed from statistics book. Here, we are not referring those statistical

derivation, our purpose is to emphasize sample size in detail, and what parameter or statistics you are supposed to include for sample size.

And so, I can derive step by step for you, I know that, this is not easily understood, let me try to derive for you. So, Z is equal to σ times square root of n . Last class, I started deriving, but because of shortage of time, I could not able to emphasize. This is N minus 1, so it is square root. So, I have taken squaring up all those things, we can follow by squaring this Z square, σ square, so this will be capital N minus n divided by N minus 1.

We will do some further adjustment to it. If I simply multiple, take this out n is equal to this Z square σ square divided by N divided by e square or times N minus n divided by N minus 1 this is there. So, let me further follow by multiplying this. So, Z square σ square, so let me take N times Z square σ square divided by e square times N minus 1, minus N times small n times that is capital N , this is , small n times Z square σ square. So, this is e square.

So, what I will do, I will take small n to the left-hand side. this will be n , 1 plus Z square σ square upon e square. So, this is equal to capital N Z square σ square e square times N minus 1. So, this is the equation we have derived so far. What is left here? So, N times this is e square plus Z square σ square equal to N Z square σ square, e square N minus 1. So, more or less we have arrived into this position.

what is there for us? N equal to this. So, what I have mentioned? I have taken the square outside. e square is taken to this level to the left-hand side square is here. So, N time. So, it is 1 plus Z square σ square divided by e square. So, that is here clearly given. So in our equation it will be multiplied with Z e square. What we will do here? So, let me take n here, n is equal to capital N then Z square is power our equation then σ square is already given.

e square is also multiplied whole divided by N minus 1 times e square plus Z square σ square. So, these two are added. What is extra added here. So, e square is added. e square could have been taken square if I take out then what is left? If I take e square out of it, this will be then N minus 1 times e square. So, what I will do? e square is taken out, the numerator remains constant.

This is N minus 1 times another square is there and then plus Z square σ square. And the numerator N is there, Z square is there, σ square is there. So, e square is there, e square e

square cancelled out. So, we have derived the equation desired for sample size calculation. So, this is nothing but the equation as I just derived. So, finally, small n is equal to Z square times capital N , capital N we have already defined and sigma square N minus 1 e square and Z square and sigma square.

So, this is the final n . This is the equation derived and I presented before you for a systematic derivation, based on these we can apply for any calculation.

(Refer Slide Time: 14:22)

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.

Solution:

$N = 5000$
 $\sigma_p = 2$
 $e = 0.8$
 $z = 2.57$ (for 99% probability)

$$n = \frac{z^2 \cdot N \cdot \sigma_p^2}{(N - 1) e^2 + z^2 \sigma_p^2}$$
$$= \frac{(2.57)^2 \cdot (5000) \cdot (2)^2}{(5000 - 1) (0.8)^2 + (2.57)^2 (2)^2}$$
$$= \frac{132098}{3199.36 + 26.4196}$$
$$= \frac{132098}{3225.7796}$$
$$= 40.95 = 41$$

n = 41

As I already discussed this, if the sample size is given and somebody, some researcher wanted to estimate the true weight of the cereal containers for the sample size of 5000 on the basis of following information, if the variance is there, in the cereal container are baskets, each of 4 ounces on the basis of past records, some past records are taken and it is very difficult to estimate the present records because you do not know the present population and their estimation. So, some past estimation could be considered as your basis.

And if you allow a 0.8 or 8 percent, 0.8 ounces as your estimate to be within 0.8 inches of the true average weight with 99 percent probability, that means 0.8 is 8 by 10 is your margin of error, if you allow that means 80 percent is your margin error acceptable range, if that is the

case, if you apply in the formula you will get 41 is the sample size that we already discussed in the last lecture.

Now, as I mentioned how many indicators you want in a population size, you want to know standard deviation or the variance of the population. You want acceptable range then you want standard table and its value at 99 percent is given or even any other level 2.58 as the standard table value. And if you apply it you will get the value is 41.

(Refer Slide Time: 16:21)

□ 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 $e = z \cdot \sqrt{\frac{p \cdot q}{n}}$

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.

Handwritten notes in red:
 $\mu = P$
 $\sigma^2 = P \cdot Q$
 $P \pm z \sqrt{\frac{PQ}{n}}$

But those are four indicator 1, 2, 3, 4 indicators may not be available in reality. So, with the limited information how could you be able to derive it? for example, if the population or those information are not given, only probability of success and failure or success is given. Firstly, we have to specify the precision and confidence level in this context, the confidence interval for population proportion if it is denoted by P hat.

As if the sample mean P hat or the P is the population average, probability of success, P times Q as I already mentioned, this is the distribution. Usually these are referred in the context of population distribution, where probability of success and failure is given.

So, the mean of the poisson distribution is P, where is the variance is P times Q. So, the standard deviation is P times Q square root of N. So, if this is the context then what is the confidence interval then, this is the P plus-minus, the way we have written Z times, Z if you convert that distribution into Z values, how to convert? So, conversion through Z value is X minus mu by standard deviation if we do that, so Z value can be converted and times if we P

plus-minus Z times PQ, that is variance divided by we are multiplying with the standard deviation of the distribution, that defines the confidence interval of this kind of distribution.

So, our acceptable error or the margin error is going to be e is equal to Z times from this point Z times square root of P Q of N. So, here it is given sample proportion is P, Q is 1-P and N is the sample size.

(Refer Slide Time: 18:31)

$$n = \frac{z^2 \cdot p \cdot q}{e^2}$$

In case of finite population

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

So, then on the basis of same formula, we can define n the sample size is equal to Z square times P into Q divided by e square. But in case of finite population, if population is known, then you have to multiply with N minus 1 rest are same N divided by N minus 1 is there in the previous equation we have already derived that, finally N divided by N minus 1, content must have been there. So, here also N divided by N minus 1 is there. So, if we apply, we will get the result.

(Refer Slide Time: 19:16)

COCHRAN'S FORMULA FOR CALCULATING SAMPLE SIZE WHEN THE POPULATION IS INFINITE

Sample size to estimate a proportion

- ❑ A professor in a department is trying to determine the proportion of students who support gay marriage. She asks, "How large a sample size do I need?"
- ❑ Required steps:
 - ❑ Margin Error: 2.5% in a 2-tailed test, generally
 - ❑ confidence intervals: 95% or different,
 - ❑ Guesstimate of the proportion: $p=30\%$ who supports gay marriage
- ❑ The margin error is 1.96 times the SE (Cochran (1977))

$$ME = z \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

8

So, we are not mentioning much on it. On the basis of this calculation, where proportions are uniform, proportions are there. So, Cochran formula is based on this particular format. So, once this example is given, to estimate the sample size given a proportion. Suppose a professor in a department is trying to get the proportion of students who support gay marriage and she simply asked that how large a sample size do I need in order to understand gay marriage.

So, the required steps here to understand these question that how large the sample today should I get to understand proportion of gay marriage. So, the margin of error suppose is given as 2.5 percent that is in a two tailed test. Had it been a single test tail it could have been 5 percent, in a two tailed test it is divided if it is two tailed, then that error is divided into two parts divided by half. So, 2.5 and 2.5.

And the confidence interval is 95 percent or any other different number some guesstimate are also there, guesstimate like usually in the society, there are 30 percent who support gay marriage not more than that, it cannot be 50 percent and 50 percent, very less percent of people who generally support gay marriage, if this information are there and, based on the margin error, 1.96 that is, if you have a confidence interval of 95 percent the Z value is of 1.96, I have already mentioned in the last lecture, the Z value based on the confidence interval, we can able to find out the Z value is 1.96.

That is 1.96 times the standard error. So, what is the standard error? So, standard error defined here, that is through the Cochran formula, which was proposed in 1977, where the margin

error as mentioned by Cochran is 1.96 that is Z times P into Q divided by N square root of N, that is standard deviation if proportions information are there. So, P Q divided by N square root is nothing but the standard deviation times the standard value that is 1.96.

So, based on this formula we can calculate the sample size. Sample size can be calculated what information are there only probability that is 30 percent if you are suggesting here, 30 percent people based on the guesstimates are known, there are many contexts like, related to gender issues, related to health issue, related to gender violence, related to epidemic, there are various context by which we can have lots of guesstimates for understanding the sample proportion.

So, two inputs are required, one is sample proportion and we require confidence level. So, if two intrusions there, we can estimate. another one is margin error, if three information are there, we can derive the sample size.

(Refer Slide Time: 22:55)

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

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

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

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

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

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 $n = 909$

So, based on that formula we have already derived. So, Z-score for different level is there 1.645, if you check the table, Z table these instruments are there, you just simply go to the Z table open that Z table. Usually, these are given in the back of the statistics books, check the 90 percent on the row and column. You check percentage are given 90 percent, 92 percent, 93 percent, 99 percent all percentage are given. And the column you will find out at what level you require.

So, how many sample sizes you have? Sample size is also given. Based on that the Z table can be suggesting you, use the standard Z value. So, the Z value at different level is, once is

calculated like these 1.96, 1.645 2.54 if it is there, N 0 that is sample size is equal to as per the formula. This is ME, marginal error, we have already defined as 2.5 percent that is 0.025. So, 1.96 is our standard value, Z value and this is probability of persons or the proportion of people who support gay marriage that is 30 percent. So, 1 minus P is 0.7, on the basis of that we can calculate the sample size.

So, for your sample size, you do require is 1291. Around you can stick to 1300, if you stick to 1300 that will be guaranteeing a 90 percent confidence level that it is confidently a researcher can claim that my result is going to be robust because my sample size has captured the true population with 90 percent confidence.

Based on that if you change your confidence level to 95 percent. So, we start with 95 percent, previous examples where our n was to be 1291, when you wanted to reduce your confidence level to 90 percent, your Z value is 1.64 on the basis of that your n is determined is 909. So, what do you mean by that? When you reduce your confidence level your sample size reduced. So, when you want to have higher confidence level you have to increase your sample size. So, this is clarified through the Cochran formula.

(Refer Slide Time: 26:05)

□ Having no idea of proportions
□ Assume 50%, i.e. $p = 0.5$

$$0.025 = 1.96 \sqrt{\frac{0.5 \times 0.5}{n}}$$

n = 1537

When you have no idea of sample proportion, but in that case, you have guesstimated or have certain proportion information, but in number of cases they are population and this proportion is not known. In that case, you have to guess 50 percent and 50 percent of success and 50 percent of failure. So, you have to assume that, if you do not know about probability of one

as against another one or successes against failure. So, like, how many supports the government there are many contexts.

So, if you have a distribution, which is person type, success versus failure, favor versus against, if these type of context and data is there, you have to guess. If guess is correct, then your estimation could be correct. If it is 50 percent, then n is going to be much higher that is 1537.

(Refer Slide Time: 27:06)

COCHRAN'S FORMULA FOR CALCULATING SAMPLE SIZE WHEN THE POPULATION IS FINITE

$$n = \frac{n_0}{1 + \frac{(n_0 - 1)}{N}}$$

n_0 is the sample size derived from equation (Cochran 1977) and N is the population size

A tax assessor wants to assess the mean property tax bill for all homeowners in Madison, Wisconsin. A survey ten years ago got a sample mean and standard deviation of \$1400 and \$1000.

How many tax records should be sampled for a 95% confidence interval to have a margin of error of \$100?

$$ME = t \frac{s}{\sqrt{n}}$$

$$n = \left(\frac{st}{ME} \right)^2$$

With $s = 1000, t = 1.96, ME = 100$, we get $n = 384$

Let us come to Cochran formula for calculating sample size, when the population is finite. In that case, we never refer population, the population size is nowhere referred, when the population size is referred the formula boils down to here, the sample size divided by that is small n divided by 1 plus small n minus one divided a capital N.

So, let us cite one example here. The example is a tax assessor wants to assess the average property tax bill for all home owners in one town. Let it be, Madison is given here, a survey 10 years ago got a sample mean and standard deviation of 1400 dollar and 1000 respectively, for their mean and standard deviation, mean is 1400 and the standard deviation is 1000. The taxpayer is trying to assess the mean property tax, for the all-home dwellers.

So, how many tax records should be sampled for a 95 percent confidence interval with a margin of error of 100. Look at margin error could be anything, if you set a margin error of a certain level, margin error 100 dollars, that is 100 dollars you are saying, if it is mean is 1200 dollar and your margin error is 100 dollars, may be acceptable. If margin error is more than that general is not acceptable.

So, where do you apply? Start with the formula, the margin error is equal to here since the sample size we are restricting to a particular distribution, we are referring to T distribution. And it is a small sample size and, in that case, T is written. So, in the place of Z and T is deliberately written and S is your standard deviation of the sample. And why S is referred? Because we are referring to T distribution, sigma could have been written if it is a Z distribution.

So, S by square root of N is called standard deviation, if it is T distribution. So, M is equal to S times T divided by a margin of error, it is whole square. So, all the information is given. We have a S is given. S is here as 1000, that is standard deviation is given 1000 and T is 1.96, at 95 percent confidence. I have already given you those values. 1.96 and margin error, is 100. So, based on this formula, we get N is equal to 384. So, Cochran formula could able to give us a clear result of 384, if very less infamous.

(Refer Slide Time: 30:14)

YAMANE (1967)

Sample sizes calculated by Yamane's formula

$$n = \frac{N}{1 + Ne^2}$$

(Handwritten: N = 450)

$$n_i = \frac{N_i}{N} * n$$

n = the sample size
N = the population size and
e = the acceptable sampling error (5%)

Sl. no. of schools	Population size, N	Sample size, n for 95% confidence level:		
		±5%	±7%	±10%
1	450	212	136	82
2	582	229	150	85
3	693	254	158	87
4	799	266	163	89
5	806	267	163	89
6	845	272	164	89
7	858	273	165	90
8	892	276	166	90
9	909	278	167	90
10	922	279	167	90
11	985	285	169	91
12	1009	287	170	91

Given. you have got examples for it, we have certain other formula when very limited information is given, called YAMANE formula, proposed in 1967, where the population information if it is there and margin error is there only population and margin error, two information is there you can able to find out the sample size. here the sample size is equal to capital N times 1 plus N e square.

So, what is that? Small n is the sample size capital N is the population size and E is the sample error usually considered to be 5 percent. here based on this YAMANE formula, we could get the information like, from the table, look at these population sizes given on the

second column, the first column, schools information's are there, first school, second school population size is there, sample size for 95 percent level of confidence with 5 percent error and that or 7 percent error or 10 percent error, there are different indicators are given.

If you simply enter the population that is 450 you can enter here, plus 1 plus 450 times the error that is 0.05. So, 5 percent, if you multiply that this boils down to a sample size. So, accordingly if you change the margin of error, you will get other sample size and this is considered to be the easiest formula. So far as sample size is concerned.

(Refer Slide Time: 31:57)

$$n = \frac{N}{1 + Ne^2}$$

n= the sample size
N= the population size (1072756 and 83260 households in rural and urban areas of two districts respectively) and
e = the acceptable sampling error (5% = 0.05)

Two districts of Odisha (i.e. Balasore and Mayurbhanj) will be selected for personal interview. The rationale of the same is presented in the appendices. The total number of urban households in Balasore district is 47360 and 35900 in Mayurbhanj district. The total number of sample selected from urban areas (Yamane formula) are 398.0875 = 400 households and similarly another (397.87) 400 are selected from rural areas since the total number of rural households in these two districts is 1072756. Therefore, the total number of household selected for sample survey is 800.

Another example is given, two districts of Odisha, we know we can also cite by referring an example from a data, but I am deliberately skipping based on my own calculation, I have collected like, Bhalasore and Mayurbhanj selected for a personal interview, if two districts in Odisha are selected for personal interview, the rationale of the same is presented in appendix, we will provide some of that information in our appendices.

As a content for your preparation, at this moment, we are not including. So, total number of urban households in Bhalasore district, as per the census data, I am not including everything because number of slides is exceeding. So, 47,360 whereas in Mayurbhanj, it is 35,900 this information is clearly given so far as population is concerned.

The total number of samples collected from urban areas, as per the YAMANE formula are 3908.0875 which is roughly around 400 households and similarly for another 397 or 400 are selected from rural areas. Since, the total number of rural households in these two districts is

107275. If I apply this number, I can get the proper calculation through YAMANE formula and this table will provide it certainly for your reference in your final preparation.

(Refer Slide Time: 33:31)

The image shows two presentation slides. The top slide is titled "DETERMINATION OF SAMPLE SIZE THROUGH THE APPROACH BASED ON BAYESIAN STATISTICS". It describes a procedure for finding the optimal sample size 'n' and lists three steps: finding EVSI, approximating sample cost, and comparing EVSI to sample cost to find ENG. It includes the formula $(EVS_i) - (\text{Cost of sample}) = (ENG)$. The bottom slide continues the discussion, noting that this approach is rarely used in practice because it is cumbersome.

DETERMINATION OF SAMPLE SIZE THROUGH THE APPROACH BASED ON BAYESIAN STATISTICS

The procedure for finding the optimal value of 'n'

- ❑ Find the expected value of the sample information (EVS_i)* for every possible n;
- ❑ Also workout reasonably approximated cost of taking a sample of every possible n;
- ❑ Compare the EVS_i and the cost of the sample for every possible n. In other words, workout the expected net gain (ENG) for every possible n as stated below:

For a given sample size (n):

$$(EVS_i) - (\text{Cost of sample}) = (ENG)$$

❑ Above the optimal sample size, that value of n which maximizes the difference between the EVS_i and the cost of the sample, can be determined.

The computation of EVS_i for every possible n and then comparing the same with the respective cost is often a very cumbersome task, this approach is rarely used in practice.

So, what is the takeaways? So far, as we know different calculus is concerned. Here, we are also suggesting something called Bayesian statistics, those who wanted to go for advanced analysis, there are some variance, stochastic variance information is required, estimate a stochastic variance, variance stochastic information. If it is there, EVSI is there, this approach is applied, but this approach is largely applied when we are interested to estimate the cost, cost related to the sample size determination.

And here the sample size EVSI that is little complicated. So, we are not including it, because it is not our purpose of emphasizing, I am just mentioning for your reference, so, EVSI minus

cost of sample is generally referred so far as sample size is concerned. But this is very very useful when you are very much concerned for cost estimation. So, what are thumb rules? So far as estimation is concerned. So far as sample size is concerned, the same thumb rules are, so in the first purpose of understanding the thumb rules is that how accurate your estimate is.

(Refer Slide Time: 34:54)

THUMB RULES

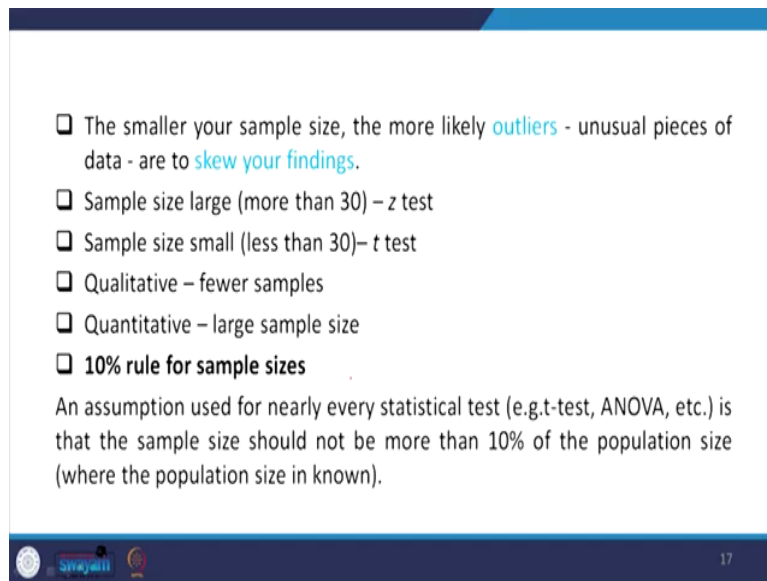
- ❑ The larger the sample size, the more **accurate** your estimates (*estimate of the true population mean*)
- ❑ **Sampling error** is **inversely related** to the size of the sample
- ❑ The size of **S.E.**, depends upon the sample size to a great extent and it varies inversely with the size of the sample.
 - ❑ *If double reliability is required i.e., reducing S.E. to 1/2 of its existing magnitude, the sample size should be increased four-fold.*
- ❑ The **central limit theorem** assures that the sampling distribution of the mean approaches normal distribution as the sample size increases.
 - ❑ *This fact holds especially true for sample sizes over 30.*

16

So, the Thumb rule of sampling error is inversely related to the size of the sample, as we already know, the general thumb rules, but these are required. The size of the standard error is also important that depends upon the size to a great extent and it varies inversely with the size of the sample. And so far as theory is concerned central limit theorem is important, we suggest that, this ensures that the sampling distribution of the mean approaches to the normal distribution as the sample size increases.

So, when you are increasing the sample size, approaching towards the population mean, this flat hole is especially true for sample size over 300. So, when sample size exceeds 30 which generally referred as not a T distribution it is a Z distribution. So, or a normal distribution, but it is up to the researcher but generally T distribution is required because we generally narrow down by our approaches narrow down our all the data to a T format. So, what are the thumb rule here?

(Refer Slide Time: 36:17)



- The smaller your sample size, the more likely outliers - unusual pieces of data - are to skew your findings.
- Sample size large (more than 30) – z test
- Sample size small (less than 30)– t test
- Qualitative – fewer samples
- Quantitative – large sample size
- 10% rule for sample sizes**

An assumption used for nearly every statistical test (e.g.t-test, ANOVA, etc.) is that the sample size should not be more than 10% of the population size (where the population size is known).

Thumb rule is the smaller your sample size more likely outliers are there. So, unusual pieces of data and their outliers are more. And so, the findings are very skewed. And so, sample size, if it is more Z test is applied and Z test is more fitted, when it is small, less than 30, T test is more fitted. Qualitative case we require fewer samples, in case of quantitative estimation we require large samples. So, usually a standard 10 percent rule is applied in many cases when your population is homogeneous.

An assumption used for nearly every statistical test that it is T-test, ANOVA, etcetera, is that the sample size should not be more than 10 percent of the population size, when the population size is known, when it is not known then that is not also standard. I think we already discussed some of the thumb rules, were already discussed some standard formula we discussed with certain examples, some calculations I think those will be very useful for the learners and friends take a note of all those things you might be in a position to prepare for your quizzes in every week for assignments, and these will be very very useful.

With this let me stop here, we will see you in the next class. Thank you.