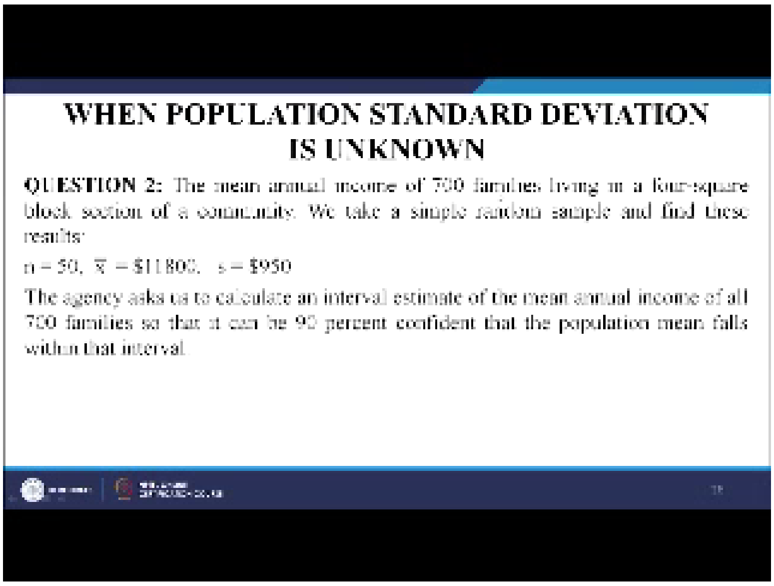
Marketing Research and Analysis-II
(Application Oriented)
Prof. Jogendra Kumar Nayak
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture - 22
Sample Size Calculation and Hypothesis Testing
(Problem Solving)

Welcome everyone we discussed about what is the alpha and that significance level alpha and the probability of the p-value right so and we understood its relationship and how to even calculate the p-value right and then reject or not reject hypothesis ok. So, today in this lecture we will be doing some problem solving. So, problem solving in terms of calculating the class interval the range between which the values fall and the sample size calculation and then we proceed if possible if time permits in this lecture for even hypothesis testing.

So, this case, the first case done was; this case is about population standard deviation is unknown. The last case was population standardization was known to us but now we are saying if the population standard deviation is unknown then how do we proceed? Let us take an example.

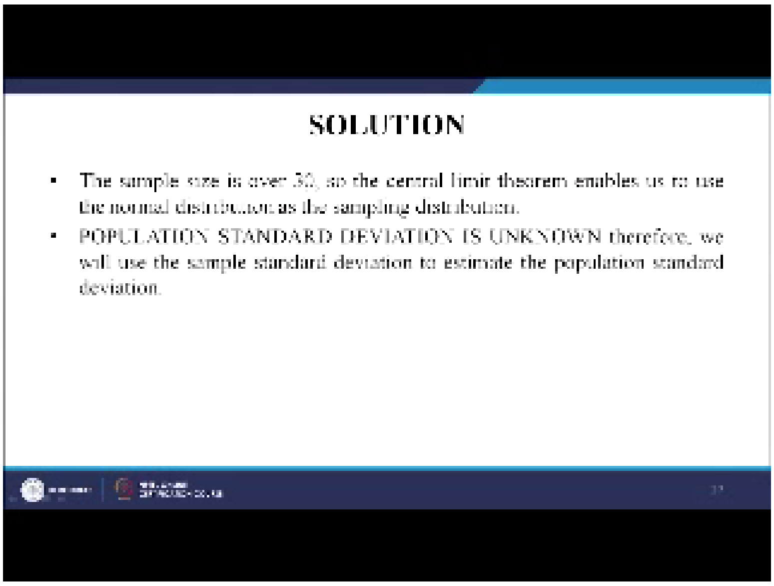
(Refer Slide Time: 01:29)



The mean annual income of 700 families living in four square block section of a community we take a simple random sample and find the results so what it is saying so the 700 families in your locality and we have taken just 50 people out of it at 50 sample. So, n is my 50 out the 700 I have taken 50 as my sample the sample mean was $\overline{x}$ = 11800 Dollars. The sample standard deviation this is sample standard deviation is 950. Here population we said the population standard deviation is unknown to us.

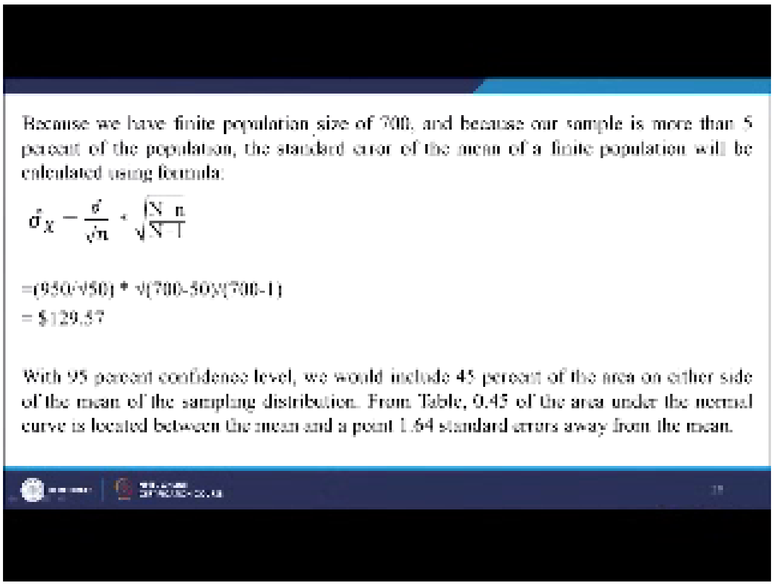
The agency is asking us to calculate an interval estimate of the mean annual income of all 700 families so that it can be 90% confidence. So, what is the level of significance? The level of significance is 1 - 90 that is equal to 10% level of significance, confident. The population mean falls within that interval. So, if I can know whether my population mean falls in the interval of upper limit and lower limit or not that is what I need to check so do we go.

(Refer Slide Time: 02:42)



Since the sample size is over 30 so it is you know it is following a normal distribution. The population standard deviation is unknown in this case highlighted ok. Therefore we will use the sample standard deviation to estimate the population whenever remember whenever the sample population standard deviation is unknown to you kindly use the samples standard deviation or in a real life case you can you make calculate the sample standard deviation and use it for the population standard deviation.

(Refer Slide Time: 03:16)



Now let us go to this case since we have a finite population there was one more thing in this case generally whenever we do any study we talk about; we do not talk about the size of the population or may not we may or we may not. When we are saying we may not that the case of infinite population we do not know what the population size. But in this case the sample size is very finite sample size and here it is only 700.

So, in such a condition what happens the σ_x or the standard error which is to be calculated is calculated in a slightly different way instead of only measuring through the $\sigma/\sqrt{n}$ which we earlier use to do we multiply it with another parameter which is $\sqrt{(N – n/ N- 1)}$. Now you see suppose this n in this case 700 ok had this 700 instead of 700 beans 7 70000 of 700000 right you would see there is hardly any difference that it makes on the Standard error.

But whenever the size of the population is small and the sample is a represent significant portion population there it makes an effect. So, now let us calculate the $\hat{\sigma}_x$.

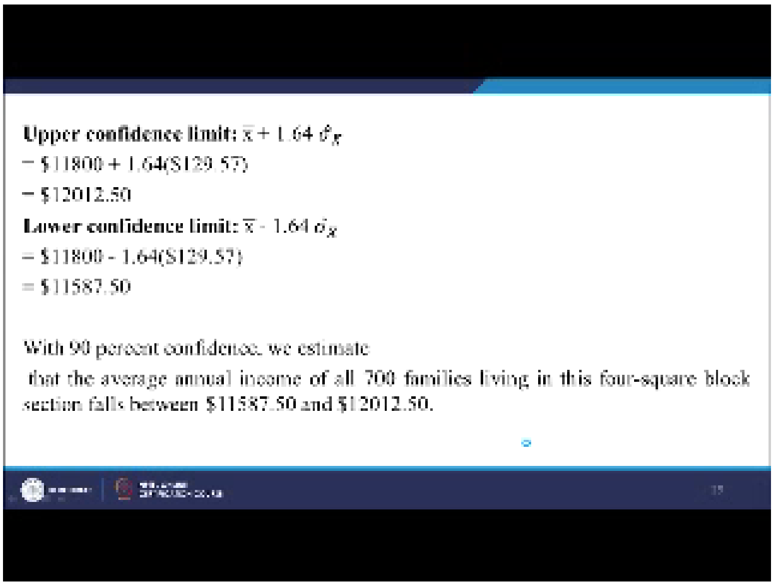
$\hat{\sigma}_x$ is equal to = $(950/\sqrt{50}) * \sqrt{(700-50)/(700-1)}$

So this gives us a sample standard deviation or standard error of sorry not sample standard deviation, the standard error of 129.57 dollars.

So, with 95% confidence level or what is our objective to calculate the interval. So, with 95% confidence level we would include 45% of the area on either side that means it is two tailed we do not know which side. So, it can go either side of the side so we have divided this sorry this is not 95 size is a mistake this is 90% it is not 95. I think the question must 90% this is 90% sorry the typological error 90% confidence level 45% this side 45% this side ok. So, .45

area is from the Z table or the other Z table if you see if it is 1.64 standard error ok away from the mean.

(Refer Slide Time: 06:01)



The upper confidence level which means is equal to how much $\bar{x}$ sample mean + Z $\hat{\sigma}_{\bar{x}}$, Z Sigma the standard error that is giving us upper limit of this much 12112, 12012 sorry and 11587 right almost I am missing the 50th the decimals with 90% confidence we estimate that the average annual income of 700 families lies between this much and this much.

Upper Confidence limit: $\bar{x} + 1.64\,\hat{\sigma}_{\bar{x}}$

= 11800 + 1.64(129.57)

= 12012.50

Lower Confidence limit: $\bar{x} - 1.64\,\hat{\sigma}_{\bar{x}}$

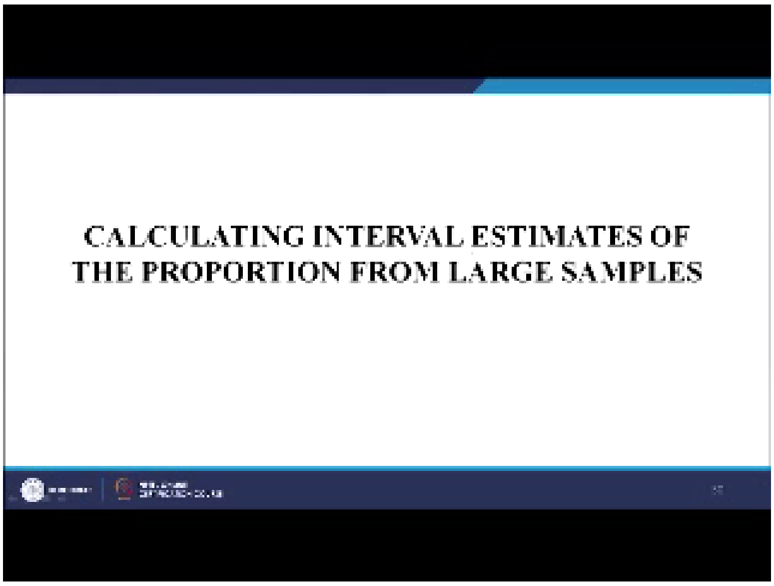
= 11800 - 1.64(129.57)

= 11587.50

So, this helps us some time to test the hypothesis now in our case what was the sample mean the sample mean was 11800. Does 11800 lie in this in obviously yes right so this is telling us the class interval.

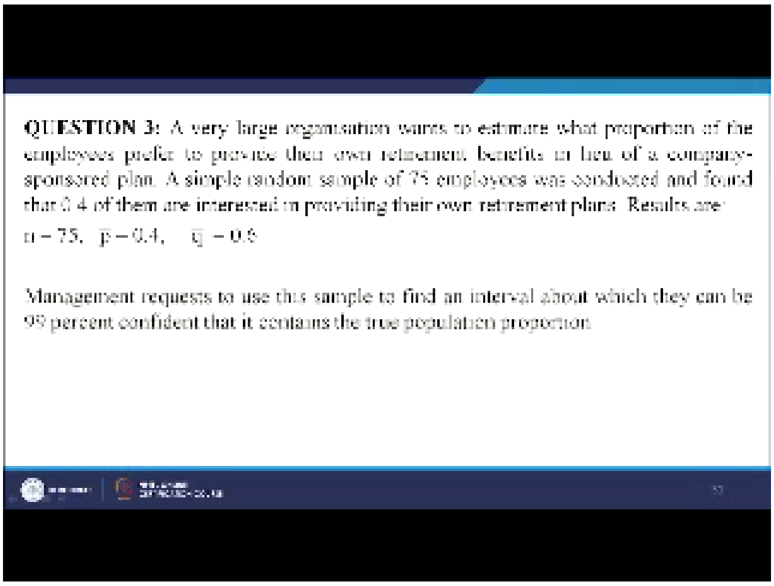
Now let us take another case, last case was the sample population deviation was unknown to you and even you had a finite population.

(Refer Slide Time: 06:58)



Now let us take this case when the interval estimates of the proportion from large samples so you are dealing with means now we are dealing with proportions. Now what is proportion and means as we all know that proportional basically the ratios basically right.

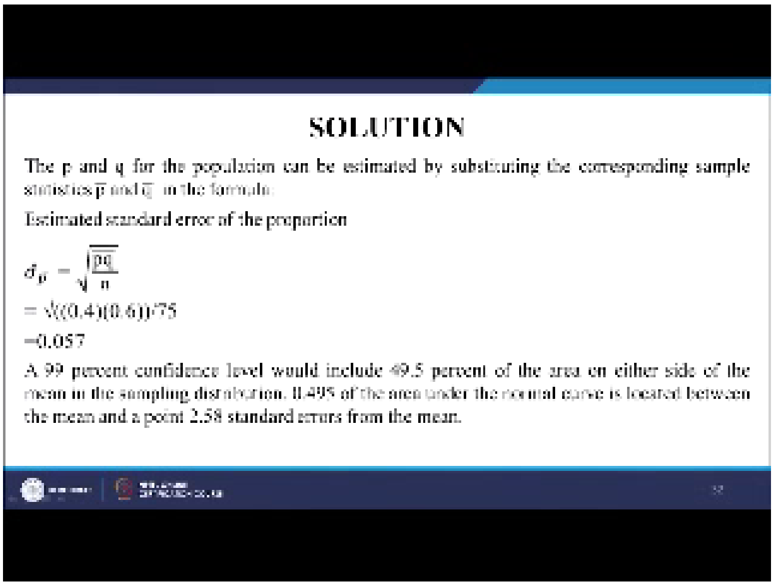
(Refer Slide Time: 07:13)



So, very large organisation wants to estimate what proportion of the employees prefer to provide that own retirement benefits in lieu of a company sponsored plan. There is a choice that the people can have their own retirement benefits or they can take the companies ok. So, simple random sample of 75, n = 75 was chosen and found that p proportion of success right was 0.4, 0.4 of them are interested in providing own retirement plan.

So, if p= 0.4 so we know q = 1 - p so that is equal to 0.6 ok. The management request to use this sample to find an interval about which they can be 99% confident that it contains the true population proportion now 99% is confidence level that you need to check about.

(Refer Slide Time: 08:19)

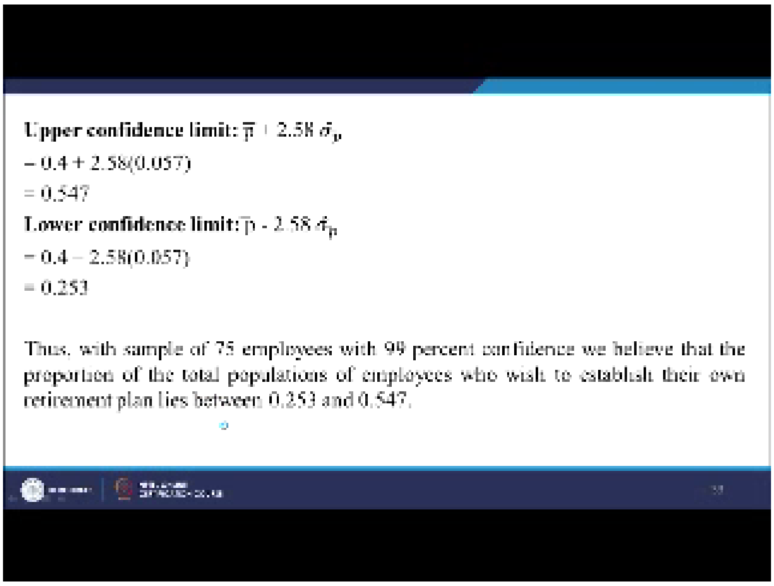


So, as it says, if p and q for the population can be estimated by substituting the corresponding formula. So, that means your; it was earlier $\hat{\sigma}_{\bar{x}} = \hat{\sigma}/\sqrt{n}$. In case of means but here what we are doing sigma = pq, so, what we are taking is this formula is changing to this one is coming up to this much right. $\hat{\sigma}_{\bar{p}} = \sqrt{(\bar{p}\,\bar{q})/n}$

Now the standard error of proportion this is called estimated standard error of proportion this is equal to 0.4 * 0.6 / n, n was 75.

So, this gives us 0.057 now what is he asking at 99% confidence level that means 99% confidence level again has to be split into two parts so 49.5, 49.5 right. So, this 49.5, this is 49.5 on either side and 49.5 the Z value is 2.58 standard error from the mean.

(Refer Slide Time: 09:32)



So, let us see how much is the confidence limit you can also calculate on your own and so that is equal to proportion of success + 2.58 into the standard error of proportion so that is equal to. 0.057 p was 0.4 let us replace this with. So, the upper limit is coming to 0.547 and the lower limit is coming to 0.253.

Upper Confidence Limit: $\bar{p} + 2.58\ \hat{\sigma}_{\bar{p}}$

= 0.4 + 2.58(0.057)

= 0.547

Lower Confidence Limit: $\bar{p} - 2.58\ \hat{\sigma}_{\bar{p}}$

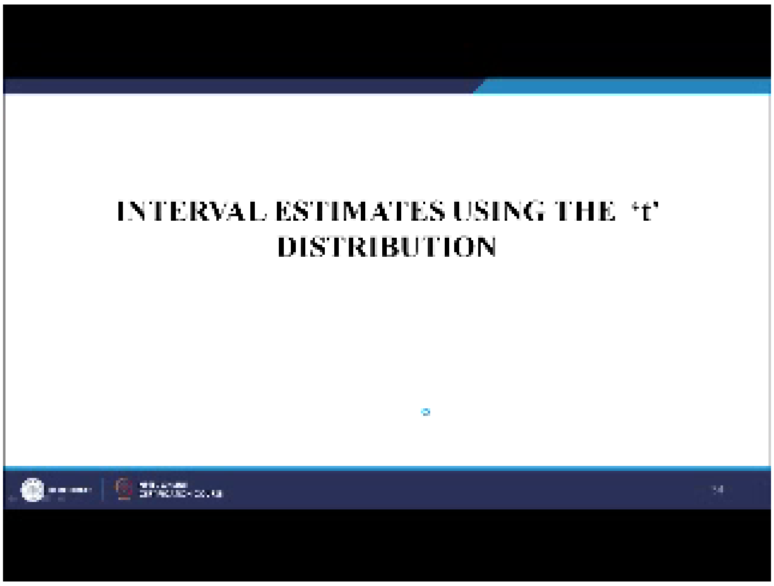
= 0.4 – 2.58(0.057)

= 0.253

If you see what is the sample of 75 employees with 99% confidence? We believe in the proportion of the total population of employees who wish to establish their own retirement plan lies between these two values.

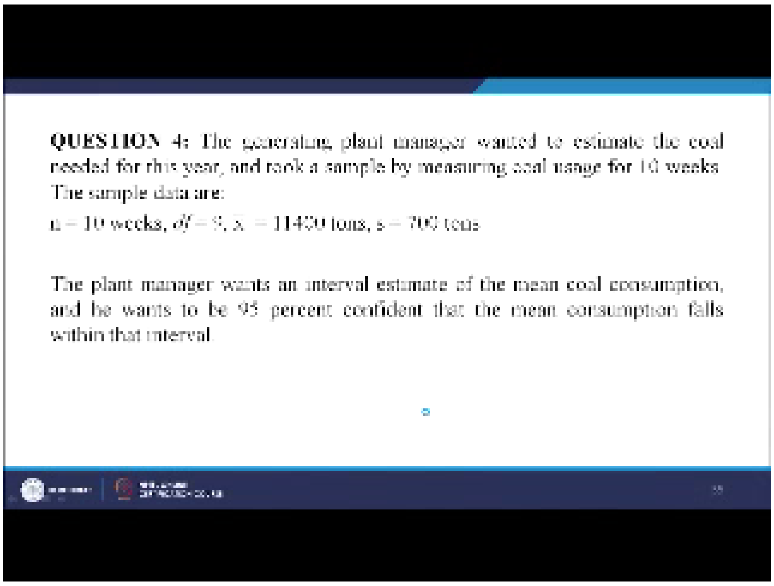
So, if you if you go back what was it saying the if you look at this, so if you look at the problem now from here it wanted to check at 99% level whether the people want their own retirement plans are not? Now yes it is saying that the retirement plan the people who wish to establish their own retirement plan lies between these two values. So, this is way of calculating the class interval when you are talking about proportions ok. Now interval estimate using the t-distribution earlier we were using the Z distribution.

(Refer Slide Time: 10:54)



Now we are talking about using a t distribution because that means the sample size now what we are going to use a talk about is less than 30 and not greater than 30.

(Refer Slide Time: 11:07)



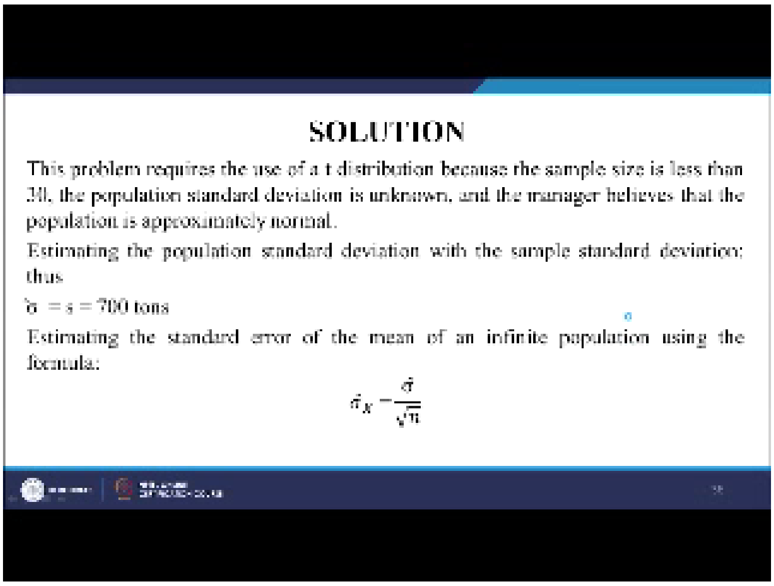
Now let us see the plant manager wanted to estimate the coal needed for this year. The plant manager wants to estimate the amount of coal needed for the year. They took a sample by measuring coal uses for 10 weeks n = 10 so we said whenever n is equal to less than 30 it is the case of t distribution. Now degree of freedom what is this word degree of freedom why it is important? It is important because degree of freedom helps us to understand what are the number of free parameters available?

That means what suppose if you want to understand in real life example let us say I tell you that the average of 10 people the weight average weight of 10 people is equal to let say 500 ok kg sorry the total weight of 10 people is equal to 500 kg. I am saying I have given you the let us say the weights of 9 people A B C D I am giving 9 people do I need to give you the 10th persons weight? It is not required correct because you know 500 minus weight of the 9 people, 9 people’s weight so that will automatically tell me the 10th person's weight.

So, that is what we say degree of freedom is actually in sense in research you need to be the word is uses parsimonious you need to be as parsimonious as possible that means you should be using minimum resources to get your result right. So, that is why if in this case n was 10 and degree of freedom will automatically be n -1 because of 10th one is not required to be given. So, that is what is the degree of freedom. Sample mean is this much tons the sample standard deviation and here also the sample standard deviation is given to you is 700 tons.

The plant manager wants an estimate interval estimate of the mean consumption and he wants to be 95% confident right.

(Refer Slide Time: 13:09)

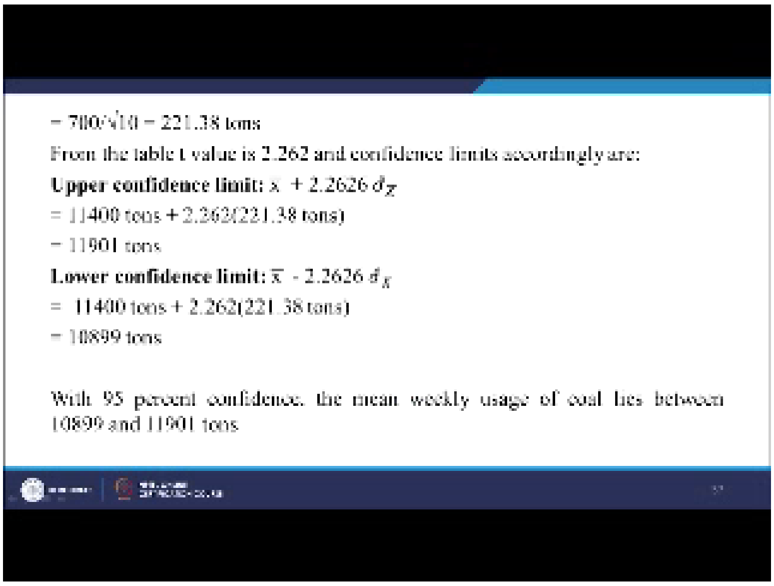


So, 95% confidence against split up right so let us see this problem requires the use of a t distribution because it is less than 30 as discussed ok. Estimating the population standard deviation with the samples standard deviation does Sigma estimated sample $\hat{\sigma}$ this is estimated one this sign always please remember this hat is it indicate estimated value estimated value. So, is equal to 700 tons. Now estimate the standard error of the mean of an infinite population using the formula what?

Estimated standard error is equal to estimated sample standard deviation which is the sample standard deviation divided by root over of n.

$$\hat{\sigma}_X = \frac{\hat{\sigma}}{\sqrt{n}}$$

(Refer Slide Time: 13:56)



Now what was the estimated standard deviation of the sample standard deviation 700 given by $\sqrt{n}$ so n was 10 so that give us a standard error of 221.38 tons

$= 700/\sqrt{10} = 221.38$ tons

from the t table remember that t table and the Z table is only one difference. The Z table is less flatter right and the t-distribution is more flatter in comparison to the Z so that this is t this is Z. So the only difference between t and Z is the flatness of the curve.

Now if you increase the sample size if you increase the sample size t starts growing up right till it reaches a point of the Z distribution that is why this is the reason why you do not see the Z distribution Z you know in any software or anywhere you do not see a Z test. You always see a t test because the t test is does not is naturally after sufficient sample size it behaves like a normal distribution.

So, let us come back to this case the upper confidence limit is

Upper confidence limit: $\bar{x} + 2.2626\, \hat{\sigma}_{\bar{x}}$

= 11400 tons + 2.262 (221.38 tons)

= 11901 tons

X bar sample mean + 2.262 the this is a z value of the t value of the table into the standard error. So, even the formula remains the same if you can see the t or Z the formula remains the same. So, this is a finally becoming the upper confidence limit is equal to 11901 ton.

Upper confidence limit: $\bar{x}$ - 2.2626 $\hat{\sigma}_{\bar{x}}$

= 11400 tons + 2.262 (221.38 tons)

= 10899 tons

the lower limit is equal to 10899. So, with 95% confidence the manager will say that the mean weekly uses of coal lies between 10899 and 11901.

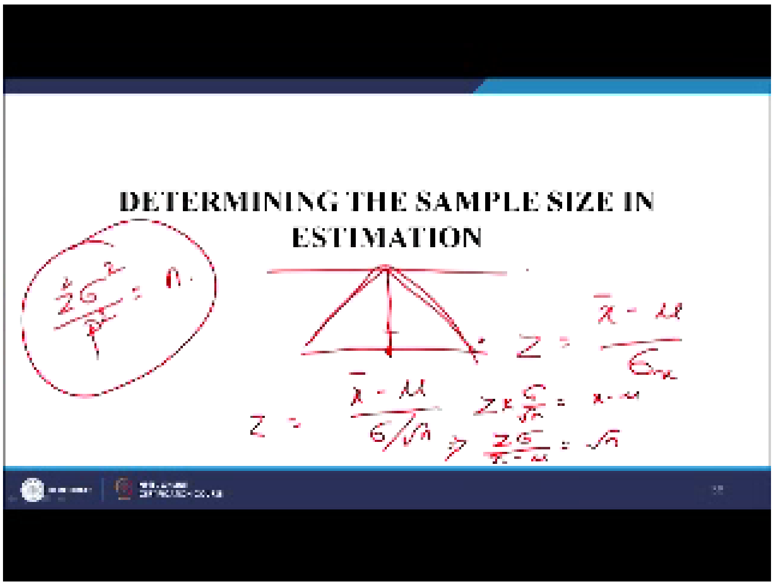
So, this is very important because from the management point of view they need to know they should lie between this value if you go less than this is a chance that you might have short of coal and if you go above it is unnecessarily wastage you are keeping too much of buffer unnecessary inventory which is also not good it is a waste of money. Had you changed this value let us say in this case from 95 to 99% what would happen?

Then one thing would have happened is as if you can see from a practical point of view if I am saying 99% confidence level instead 95 it means I want to be more confident when you have more confident when you have more resources correct. So, that means that Z value or t value from the table value this will not be 2.262 anymore it will be much is higher than for 99% is higher than the 95% value. So, when this is higher automatically the absolute value also will become larger than what this one is plus something right.

Plus Del X let say and plus also del X, so that it is what it says when you are increasing the 99 confidence level automatically the interval size range changes therefore the; now you see what has happened as you increase the confidence level your alpha the significance level was 1% has become 1% so as you reduce the alpha right. Now beta might also correspondingly change if you go back remember the last to last lecture.

So, we say that alpha and beta are related so the chance of having a beta is also changing with alpha this needs to be very clear ok.

(Refer Slide Time: 17:27)

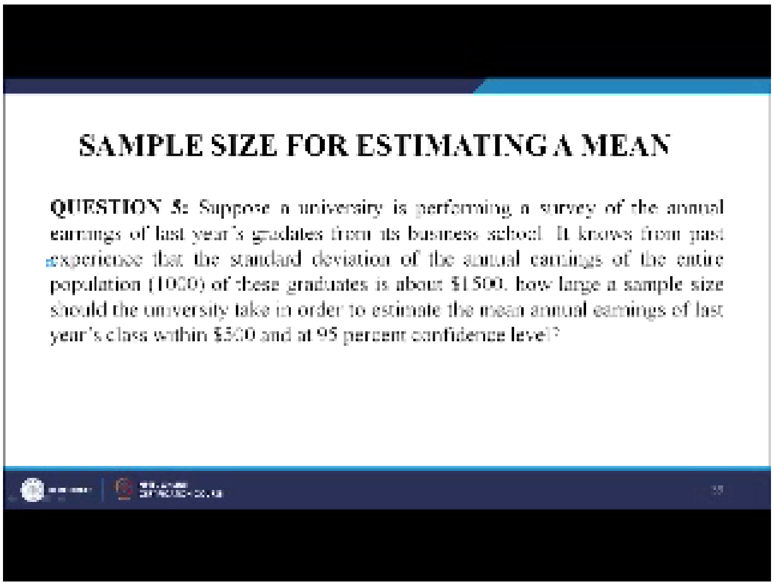


Determine the sample size in an estimation. Why this word estimation first let us needs to understand estimation is many a time you need to focus on a very close value and estimated value which is close to the mean very, very close or a focused point a point. We say sometimes it is a point estimate right. So, this is the mean let us say just say this is a mean when we want it to be as close it to the mean as possible right.

When you are talking how to determine the sample size right so earlier also you are seen the formula was Z or t whatever Z or t was equal to $(\overline{X} - \mu)/\sigma_x$
correct this was the formula so standard error was we said $Z = (\overline{X} - \mu)$ divided by $\sigma/\sqrt{n}$.
So, that means I can just simply calculate n the sample size from here also. If I know the Z value right so that means what $(Z*\sigma) / \sqrt{n} = (\overline{X} - \mu)$ right this implies root over of $(Z*\sigma)$ divided by $(\overline{X} - \mu)$ is equal to $\sqrt{n}$.

So if you just do it $(x - \mu)$ is nothing but the precision so $(Z\ \sigma)$ right Z Square σ square right divided by $(x - \mu)$ so μ is a population $\overline{X}$ is the sample so the difference is nothing but a precision and so precision can be is noted as p maybe. Let us say p or sometimes in books you will see it is a d so this is nothing but the difference between the population and the sample mean into square is equal to that is all ok.

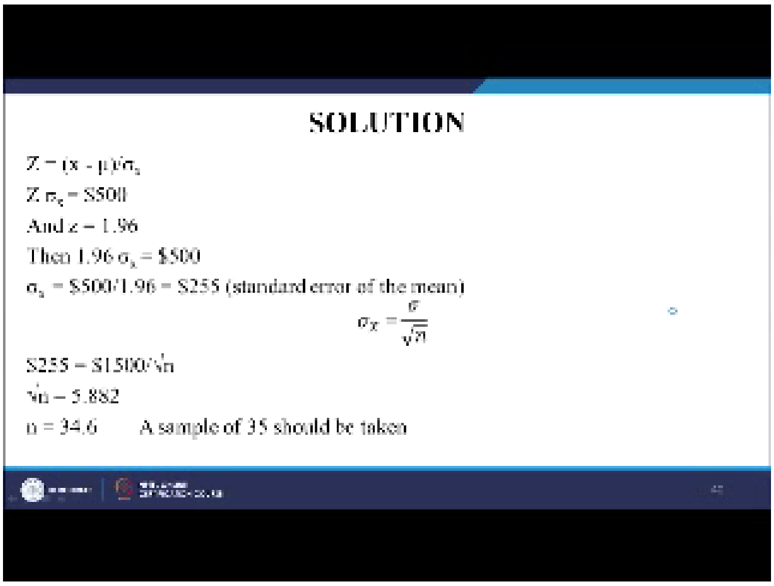
(Refer Slide Time: 19:23)



So, let us see this sample size for estimating a mean. The universityy is performing a survey of the annual earning of last year graduates sorry from its business school. It knows from past experience that the standard deviation of the earnings of the entire population of these graduates is about 1500. So, the population size is known it is a finite population is about 1500. How large is sample size should the university take in order to estimate the mean annual earnings of last year class within 500 at 95% confidence.

So, what it says let us go back standard deviation of the annual earnings of the entire population which is thousand people of this graduates is about 1500 Dollars. What should be the sample size the university should take to estimate the mean annual earnings of last year class within 500 of this years right at a 95% confidence level. So, how will you approach this problem? So, what you can do is you start thinking about a bit and then we will proceed.

(Refer Slide Time: 20:33)



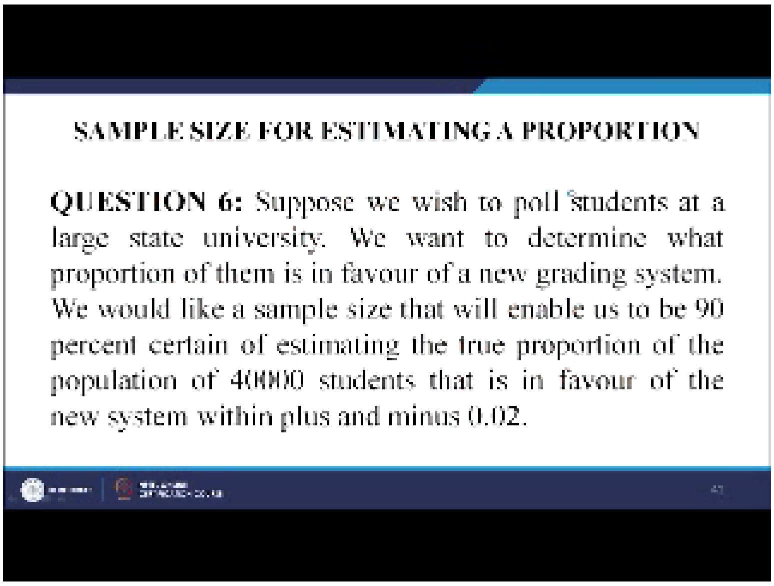
First let us calculate this is Z, we need to calculate this Z, now what is Z? $Z = (x - \mu)/\sigma_x$

$Z \sigma_x$ in this case is equal to 500 just look if you go back what it will say $Z * \sigma_x$, $(x - \mu)$that means the sample minus the population so this is the precision, the exact the difference it said that in the question it said it should not be more than 500 it should be close to; let just you see the annual earnings of last year class within 500.

So, we are saying this should be not more than 500 right so this is equal to $(x - \mu) = Z \sigma_x$ that means what $Z \sigma_x = 500$ ok this is what comes and Z is 1.96 because it is either 95% confidence level. So, what is my σ_x? $\sigma_x = 500 / 1.96$ so that is equal to 255 ok. So, this is 255 what is my standard deviation and standard deviation was 1500 and the population standard deviation. So, that means what we are saying $\sqrt{n}$ is equal to you see 255 sorry 1500 / 255 so n is equal to this much right $\sqrt{n}$.

So, that means is this is equal to whatever the n is equal to finally when n is equal to kindly be calculated 34.6 or 35 almost approximately 35. So, if you take 35 samples in this study then you will get a result which is almost 500 within 500 of the class limit of the earnings of the last year's graduates and this is at a 95% confidence level. So, I think this is clear this is very simple so you just need to understand one concept in this that this was you are talking about the closeness so $(x - \mu)$ so that is what you need to understand otherwise other things are simple right.

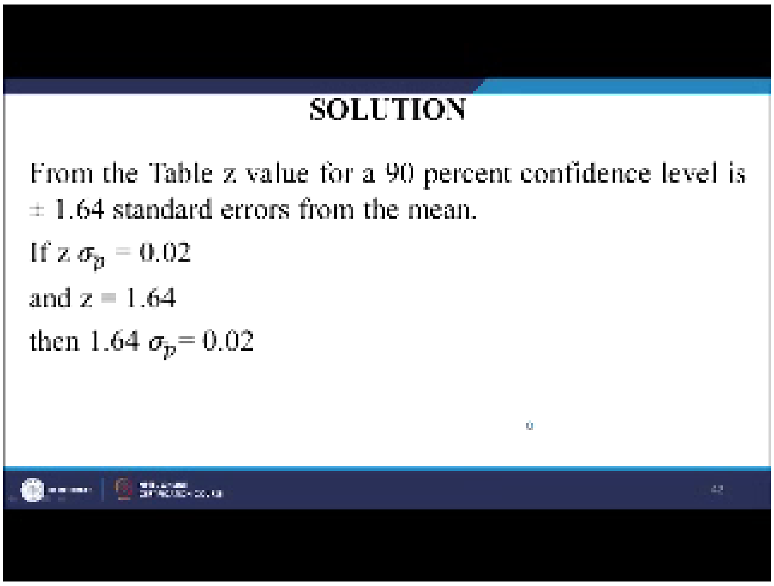
(Refer Slide Time: 22:50)



Now this is another condition the sample size for estimating a proportion so this was mean, we talked about now proportion. Let us take this case again suppose we wish to poll students at a large State University we want to determine the what proportion of them is in favour of a new grading system. So, that there is a old grading system and a new, students are always you know I thinking that the old grading systems are generally sometimes we think which is poor the old things are also poor not necessarily let us see this.

You like to sample size that will enable us to be 90% certain of estimating the true proportion of the population of 4000 student's sorry 40,000 students in favour of the new system within plus and minus 0.02. So, now here also let us see what to be done.

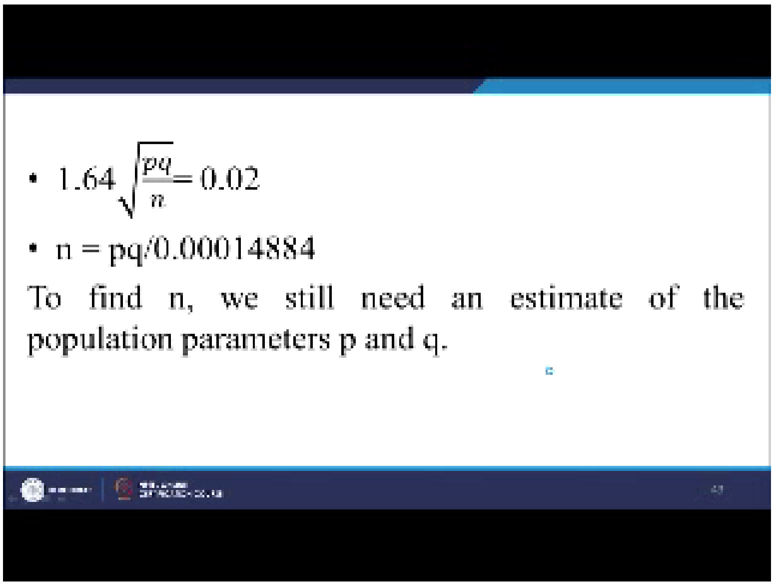
(Refer Slide Time: 23:44)



In this case the table Z value for 95 is 1.64, standard error from the mean right. Now Z $\sigma_{\bar{p}}$ is equal to same thing if you go back so we said p as for the for example the mean what it was $(x - \mu)$ right. So, in case of proportions what did we say $p - \bar{p}$, so $\bar{p} - p$ is what we required? So, it says $\bar{p} - p = 0.02$ correct and I thing that is what it is saying let us go back to the questions. We would you like a sample size that will enable us to be 90% certain of estimating the true proportion of the population of 40000 students that is in favour of the new system within plus and minus 0.02.

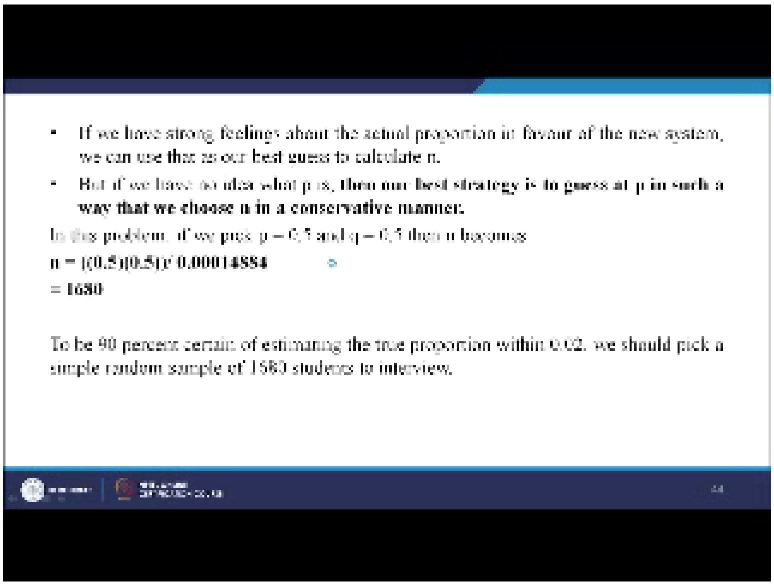
So, this difference between these two $\bar{p}$ and p right the success of ratio is need to be only 0.02 so that means what it is saying if this is then that means if you calculate you find out the your Z $\sigma_{\bar{p}}$ is equal to 1.64 that was the Z value/ 0.02.

(Refer Slide Time: 25:00)



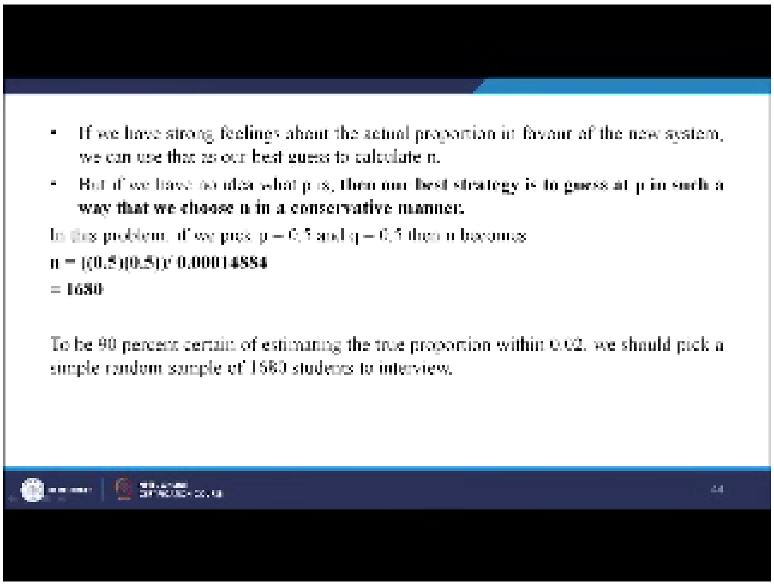
Now that is equal to now if I go $\sigma_{\bar{p}}$ is nothing but pq/n we have already done so n= p*q to find n you still need estimate of the population parameters p and q since it was not given to us we could not, suppose it would have given us then we would have said p is suppose is let say .4 in one case we are done so, p was .4 q was .6 so if this becomes .4 * .6 divided by this value.

(Refer Slide Time: 25:35)



So, if we have strong populations feelings about the actual proportion in favour of the new system we can use that are best ways to calculate n. But if you are no idea what p is suppose you have the idea of what is the rate of success or your accepted value of proposal accepted ratio and accepted proportion then it is ok.

(Refer Slide Time: 25:56)



But then if you do not have any idea the best strategy is to guess that p in such a way that we choose it in a conservative manner now what is conservative manner. So, in this condition if you do not know you think take p = 0.5 and q = 0.5 so when you do not know anything the probability of success let us take it 50% and the product failure, we take it 50% right. So, then n becomes in the last case had it been given as I said 4 and 6 are taken from the last one example which I had talked about proportions.

Had not been given then I would have taken .5 * .5 and by this much so this gives me the sample size of this. So, that means to be 90% confidence certain of estimating the true population true proportion within 0.02 right we should pick a simple random sample of 16080 students to review. So, if you take less than that then you will not be able to find right. So, this is what it is saying so I think what I will do it instead of going to the next lecture now what I will do it I will discuss what we are done in this lecture.

So, we understood the importance of P value also right and then we understood the relationship between P, alpha and other things and then we help, it helps to calculate the interval in the class the range the interval is the range for the study for a condition where the mean testing for means or maybe testing for proportions but and you are also help it also helping you in calculating not only the interval but also sample size by using the same formula that we have been using for Z or t which is $Z = (x - \mu)/\sigma_x$.

So, standard error will definitely become slightly different when you are calculated for proportions right. So, when you are calculating for propositions as we said the standard deviation is nothing but root of pq so ultimately the formula changes. So, now I think you are sure that what is this mean what is this proportion and how the formula for the class interval calculation changes and the sample size calculation changes in terms of a mean in terms of a profession.

What happens when there is a finite population and when there is an infinite population that also you have learnt? What happens when the probability of the p and q is not given to you what do you do? What do you do when the population standard deviation is not given to you? And what do you use? You use the sample standard deviation in that case. So, this things need to be very clear for a researcher because in real life in practical nobody is going to give you in a paper or something.

So, this data you need to sometimes calculate and therefore some kind of deviation some little variance is acceptable but then you should have a clear idea how do I measure? How do I calculate? These things we will later on also do it on the software and I will show you how this is calculated but for another one or two class we will get into this calculation on the paper or let us say on the board right.

I hope this class was clear to you the most important thing that I would say in this class which I understand the probability of the p value and then how to calculate the interval right thank you so much have a nice day.