Essentials of Data Science with R Software - 2
Sampling Theory and Linear Regression Analysis
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Sampling Theory with R Software
Lecture - 22
Simple Random Sampling
Confidence Interval Estimation of Population Mean

Hello, welcome to the course Essentials of Data Science with R Software 2, where we are trying to attempt to learn the topics of Sampling Theory and Linear Regression Analysis. In this module we are going to discuss the topics of Simple Random Sampling. So, as I told you earlier in the last lecture, that in this lecture we are going to talk about the Confidence Interval Estimation of the Population Mean $\bar{Y}$.

And I am sure that you must have learned that what is the confidence interval estimation. But still I will just give you a very simple example to give you a feeling that why are we trying to do it here. Now, suppose I ask you a very simple question. How much time do you take in going from your home or say hostel to your college or say office?

Think about and let me know, your answer will be say 30 minutes. But, then my question is this exactly 30 minutes in 30 minutes 0 second. Now, you will rethink it is not always exactly 30 minutes, but sometime it might be 25 minutes, sometime it might be 28 minutes, sometime this might be 30 minutes sometime 31, sometime 32 maybe 35 minutes.

But, usually under normal conditions this time will not be a smaller than 25 minutes and more than 35 minutes. So, what are you trying to say now, earlier you said? You take 30 minutes of time. But, now would you like to rephrase your sentence that yes I take 25 to 35 minutes in reaching to my college or say office.

So, now if you try to see the difference between the two sentences, earlier you said 30 minutes, now you are saying 25 to 30 minutes. So, in the first case that was the value of the time at a point. So, that is called as point estimate. And in the second situation you are trying to estimate the time, but this is in the form of two points and they form an interval.

So, now you are trying to estimate the time in the form of an interval. So, when we are trying to estimate a parameter at a point that is called as point estimator or say point estimation, and when we are trying to estimate the parameter in an interval this is called as interval estimation. And, when we say about interval estimation then we try to construct the confidence interval estimation.

Why confidence interval? For example, if I say that my question will be what are the chances that you will take 29 minutes to 31 minutes in reaching to your college or office, and then I again ask you what is the probability that you take between 25 minutes and 35 minutes in reaching to your office.

So, definitely once I say 29 minutes to 31 minutes that is a pretty close interval. And you are not confident that whether you can always reach to your college or office in that time interval, but when I say that your time is between 25 and 35 minutes, then possibly you feel more confident. So, with every interval, there is associated a confidence coefficient.

And, when we are trying to state a confidence interval then there are two components, one we have to state the lower limit and upper limit of the interval. And secondly, we have to state the value of the confidence coefficient, confident coefficient will in simple language will indicate your confidence, or the confidence that the estimated value will lie between the lower and upper limits.

Now, next aspect up to now we have considered the estimation of Y bar $\overline{Y}$. And we have found the estimate of variance of sample mean, in both the cases we have not assumed any probability distribution. We have not assumed that your y_1, is y_2,, …, y_n, that is the sample values they are coming from a probability density function like, normal distribution or t distribution or binomial distribution or Poisson distribution or say anything else.

But, now when you are trying to consider the confidence interval estimation, then we need to make this assumption. We need to associate a probability distribution function or a probability density function with my population.

So, what we assume now that my population from where we are going to draw the sample, that is being characterized by the probability density function. Well you can take any probability density function, but in this lecture because we are at a beginners level. So, I will try to assume a normal distribution.

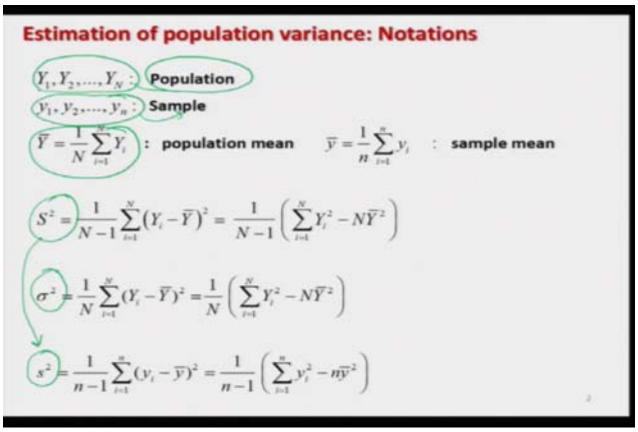
So, normal distribution have two parameters mean, which is indicated by μ and variance that is indicated by σ^2. So, mu and σ^2 are the values of population mean and population variance, but in the population. So, here our objective is that I would like to consider here the estimation of population mean.

So, I have two choices this σ^2 is known or unknown. So, I will try to construct here the confidence interval for the population mean under these two given conditions. Yes, what you are thinking that is also correct that what about σ^2; σ^2 is also unknown yes. If σ^2 is unknown, then you can estimate it and you can estimate it as a point estimate as well as confidence interval estimation.

But, surely here I am not going to do it, but that is available in any confidence interval estimation book, where the topic of confidence interval estimation under normal distribution is considered. So, I would recommend you that you please try to look into those books.

So, now let me start our this lecture, first I will try to give you the quick revision of the result whatever we have found earlier. Because, we are going to use them here and then I will try to give you a result, which is going to be used in the confidence interval estimation although I believe that you know it, but just for the sake of say completeness I will try to state it ok.

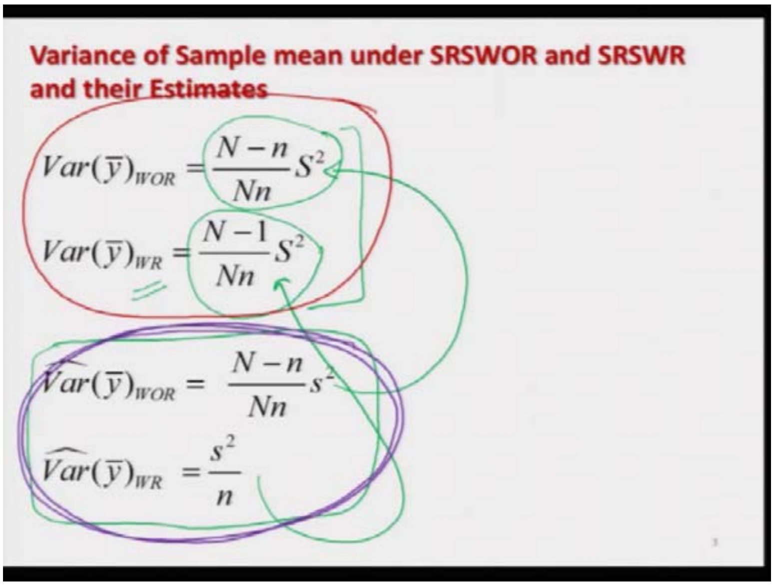
(Refer Slide Time: 08:32)



So, now first a quick revision about the symbols so, that I already have done in the earlier lectures also. So, this $Y_1, Y_2, \ldots, Y_N$ they are going to indicate the population of size N y_2,,…, y_n, they are going to indicate the sample values that is the sample of size n, $\overline{Y}$ is going to indicate the population value.

And similarly S^2 and σ^2 they are the quantities which are reflecting the population variance with the divisor N - 1 and N. And a s^2 that is the sample counter part of this S^2 and we had seen in the last lecture that, this quantity also helps us in estimating the variance of the sample mean under SRSWR and SRSWOR ok.

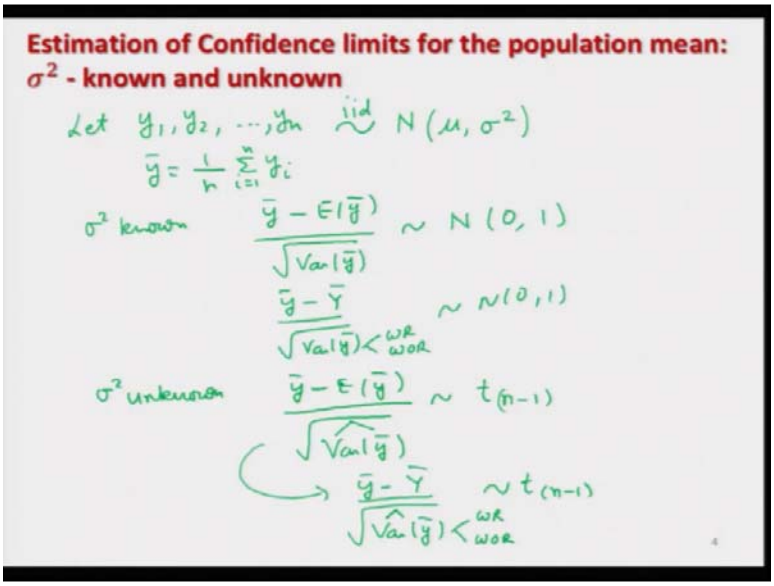
(Refer Slide Time: 09:26)



And, you had seen that we had estimated the variance of $\overline{y}$ that the sample mean under WOR and WR. And those expressions are here like this $\frac{N-n}{Nn}S^2$. And for the variance of $\overline{y}$ under WR, this is $\frac{N-1}{Nn}S^2$.

And, in case if you want to estimate it on the basis of given sample, then these two quantities they are indicating the unbiased estimator of population variance of $\overline{y}$. So, this estimator is going to estimate the variance of $\overline{y}$ unbiasedly and this quantity is going to estimate the variance under $\overline{y}$ WR unbiasedly. So, this is for your quick review.

(Refer Slide Time: 10:21)



Now, let me try to state here a simple result. So, if we assume say let y_1, y_2,…, y_n this is a random sample; that means, they are i i d identically and independently distributed. And this observations are coming from a normal distribution whose mean is indicated by mu and variance is indicated by σ^2, right.

Now, in case if you try to define here a quantity $\bar{y}$, which is actually your sample mean which you have considered $\frac{1}{n}\sum_{i=1}^{n} y_i$, then I have here two cases one is σ^2 is known. Under this thing ($\bar{y}$ - E($\bar{y}$))/ $\sqrt{Var(\bar{y})}$. , this will follow a normal distribution with mean 0 and variance 1, right.

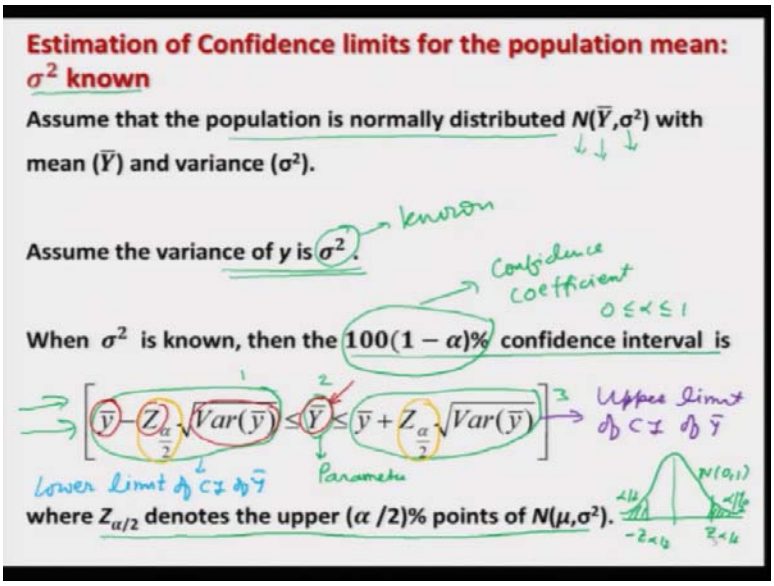
So, in your case you can translate it to $\bar{y}$ and E($\bar{y}$) in case of SRSWOR WOR, it was $\bar{Y}$ and the variance of $\bar{y}$ that you already have obtained under the two situation WR and WOR right, which is here you can identify I will just indicate it by another color red. So, this is the expression ok.

So, this quantity will follow N(0, 1), normal distribution with mean 0 and variance 1. And similarly when σ^2 is unknown, then in that case the quantity $\bar{y}$ - E($\bar{y}$)/ $\sqrt{Var(\bar{y})}$ but, this variance of $\bar{y}$ cannot be estimated because σ^2 is unknown or equivalently S^2 is unknown. So, this can be replaced by estimator.

What is this estimator? You can see here, now I will try to indicate it with another color, you can see here this is the estimator, right. So, this quantity will follow a t probability distribution with n - 1 degrees of freedom, right. So, in your case this translate to $\bar{y} - \bar{Y}$, because E($\bar{y}$) is $\bar{Y}$ in both the cases SRSWR and WOR.

But, in the denominator this will be variance of $\bar{y}$. and this can be used under the two condition, whether the sample is drawn by WR or by WOR and this will follow a t distribution with n - 1 degrees of freedom. So, these are the two results which I am going to use further in the construction of confidence interval estimation, ok.

(Refer Slide Time: 14:04)



So, now first I try to consider the case, where σ^2 is known. So, we assume that the sample is coming from a population which is normally distributed. So, it has got a normal distribution with mean $\bar{Y}$ which is the population mean and variance σ^2 . So, we are assuming here that the variance of every y that is every observation is the same that is σ^2 .

Now, I have here two option that this σ^2 is known or σ^2 is unknown. So, in this case I am assuming that this σ^2 is known. So, under this condition that is when σ^2 is known the 100(1 - α) % confidence interval for $\bar{Y}$ is given by this quantity by this interval. So, if you try to see there are several quantities which are written in this expression.

So, let me try to explain you one by one this quantity 100 (1 - α) % this is called the confidence coefficient ok. And α is a value which is lying between 0 and 1 and α is also related to the test of hypothesis, where we try to fix the type 1 error as α. So, you might have heard that 95 % confidence interval or 5 % level of significance. So, α is the similar value same value.

Now, here there are three quantities say here quantity number 1, quantity number 2 and quantity number 3. So, what are these things let me try to explain you one by one. So, first I come to quantity number 2 which is your $\overline{Y}$. So, this is the parameter which you want to estimate for which you want to find out the confidence interval. The confidence interval means the interval which is comprising of lower limit and upper limit.

So, this quantity here 1, this is $\overline{y} - Z_{\frac{\alpha}{2}}\sqrt{Var(\overline{y})}$, this is called as the lower limit of confidence interval of $\overline{y}$. And, similarly if you go to this quantity number 3 here, this is called as upper limit of confidence interval of $\overline{y}$.

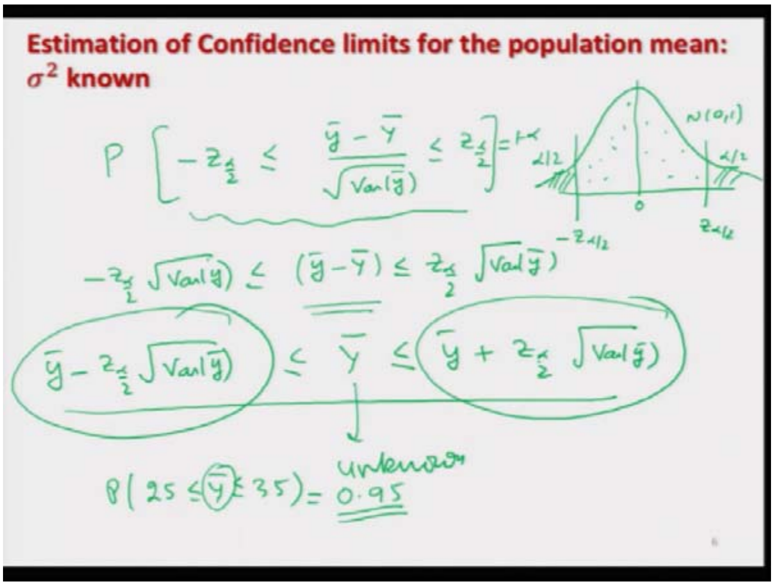
So, now you can see here that when you are trying to make a statement that you would like to find out the confidence interval of a particular parameter this amounts to state that you want to find out two quantities, one is the lower limit and other is the upper limit of the parameter.

Now, here also you can see there is one quantity here $Z_{\alpha/2}$, this $Z_{\alpha/2}$ this is indicating the $\alpha/2$ % points on the distribution of $N(\mu, \sigma^2)$. If you know the probability density function of $N(\mu, \sigma^2)$ looks like here this. So, if you try to so suppose this is you are here, suppose N(0, 1). And suppose if you try to say that this is α % point or this is $\alpha/2$ % point on this side α /2 % on this side. Then, this is suppose $Z_{\alpha/2}$ and this is - $Z_{\alpha/2}$.

So, this value of $Z_{\alpha/2}$ can be obtained from the tables. The tables are available which are computing the value of Z for different values of α. Now, if you try to look at the interpretation of this confidence interval, this confidence interval is saying now my this $\overline{Y}$, $\overline{Y}$ is unknown to us.

Now, we draw a sample on the basis of sample we compute the values of $\bar{y}$, variance of $\bar{y}$ and we obtain the value of Z from that table. So, I can compute the lower limit as well as upper limit of this interval. So, now I am saying that the probability that $\bar{Y}$, which is unknown to us will be lying between this lower limit and upper limit is 1 - α right.

(Refer Slide Time: 19:34)



How this interval is obtained let me try to first give you a quick review. So, as I explained you earlier that this ($\bar{y}$ - $\bar{Y}$)/ $\sqrt{Var(\bar{y})}$. this will follow a normal distribution. So, I can say that suppose this is lying between - $Z_{\alpha/2}$ and $+Z_{\alpha/2}$ so, in a probability distribution function of N(0, 1).

So, this is your here quantity $Z_{\alpha/2}$ and on the left hand side the quantity is $-Z_{\alpha/2}$. And this area on the right hand side and left hand side, which is the shaded area this is actually α/2 and here is the mean, ok. So, this value will lie between - $Z_{\alpha/2}$ and + $Z_{\alpha/2}$; that means, in this interval right. And I am saying that the probability of this event is 1 - α.

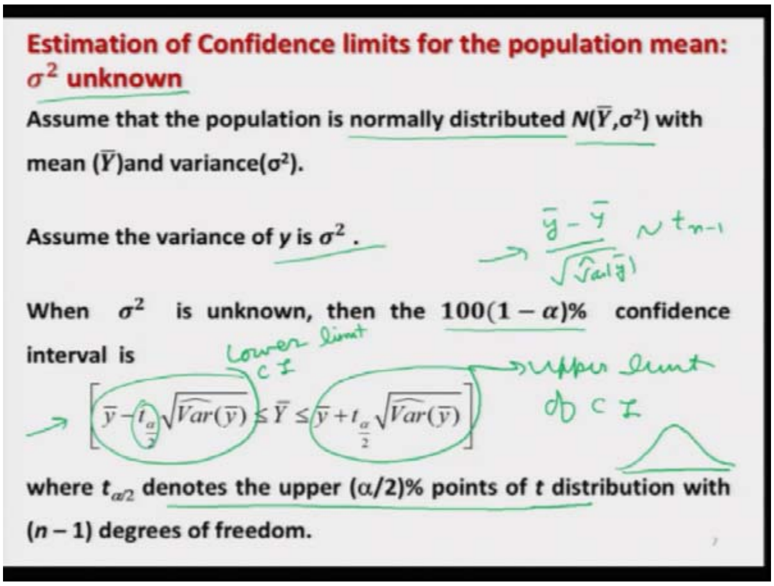
So, now what I have to do I simply have to simplify this quantity in the bracket for example, I can write down here $\bar{y}$ - $\bar{Y}$ will be lying between - $Z_{\alpha/2}$ $\sqrt{Var(\bar{y})}$ and + $Z_{\alpha/2}$ $\sqrt{Var(\bar{y})}$.

And, now in case if you try to just solve this quantity, you try to solve that what will be the limits for this $\bar{Y}$ you will see here this is $\bar{y}$ - $Z_{\alpha/2}\sqrt{Var(\bar{y})}$. And $\bar{y}$ +s $Z_{\alpha/2}\sqrt{Var(\bar{y})}$. And this is here the same result which is given here you can see here right ok. So, I have given you one proof, now I will not be giving you any proof in the further slides, but they will follow exactly on the same line.

So, now you can see here that you are trying to say that I do not know $\bar{Y}$ this is unknown to us. And, if I try to find out this limit and this limit which can be computed on the basis of given sample of data. Then, the $\bar{Y}$ will be lying between these two limits for example, if I say the probability that $\bar{Y}$ lies between 25 and 35; that means, you take 25 to 35 minutes in reaching from your home to your office or college suppose this is 0.95.

So that means, there are 95 % chances that the estimated value of $\bar{Y}$ will be lying between 25 and 35. And yeah means you can choose the different values of α and you can create, different confidence interval with different confidence coefficients, right, ok.

(Refer Slide Time: 22:51)



So, now I come to the second case where we assume that σ^2 is unknown to us. So, again we make the same assumption here that the sample is coming from a population which is following a normal distribution. Normal with mean $\bar{Y}$ and variance σ^2 and every unit has got the same variance σ^2.

And, when I am making this assumption here that that the variance of every unit is σ^2 that also reflect the same discussion, which we had earlier where I indicated that how are you going to choose that whether you have to use simple random sampling or say another type of sampling.

So, you have seen that while we are trying to estimate the variance or this confidence interval estimation, we are assuming that the variance of y is actually σ^2; that means, every unit has got the same variance. So, if your population has got a very high variability that different sections of the population have got different variances which are varying a lot.

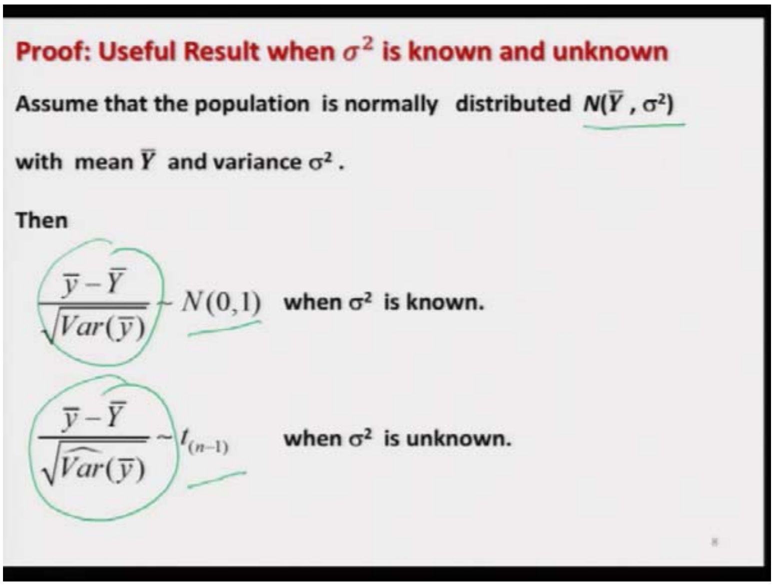
That means, the simple random sampling may not give you a good outcome, because this assumption is being violated and that is my precise reason to give you this algebra. So, that when you are trying to do it for some bigger data sets like as in the case of data sciences, then you should know whether your results what you are achieving are good or bad.

So, now I am going to use here the result which we discussed earlier that is $\bar{y}$ - $\bar{Y}$ upon $\sqrt{Var(\bar{y})}$ and its estimator this follows a t distribution with n - 1 degrees of freedom.

So, when σ^2 is unknown, then this results hold through and under this condition the $100(1-\alpha)$ % confidence interval will come out to be like this. Now, you can see here this is the lower limit, lower confidence interval is a lower limit of confidence interval and this is here the upper limit of confidence interval.

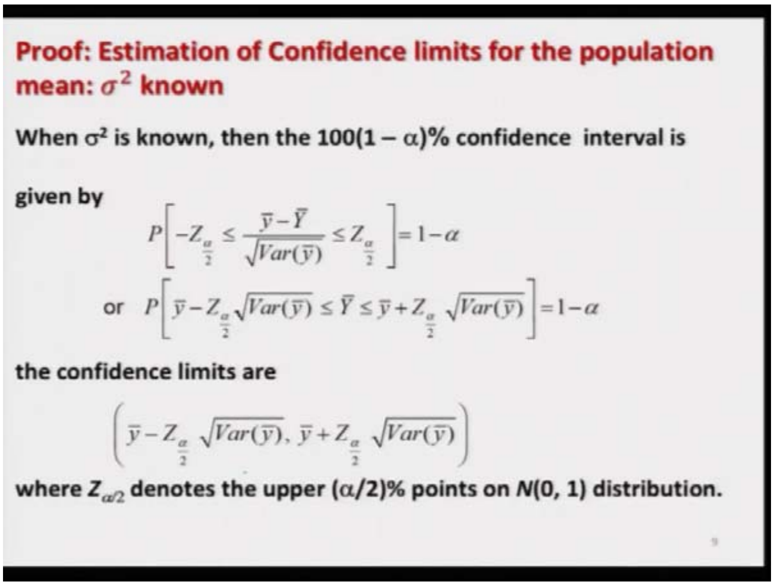
And, this value here t this $t_{\alpha/2}$ indicates the upper $\alpha/2$ % points on the t distribution with n - 1 degrees of freedom. The distribution of t and normal they are more or less similar and when these degrees of freedom are more than 30, then the t distribution and normal distribution they becomes almost the same right.

(Refer Slide Time: 25:41)



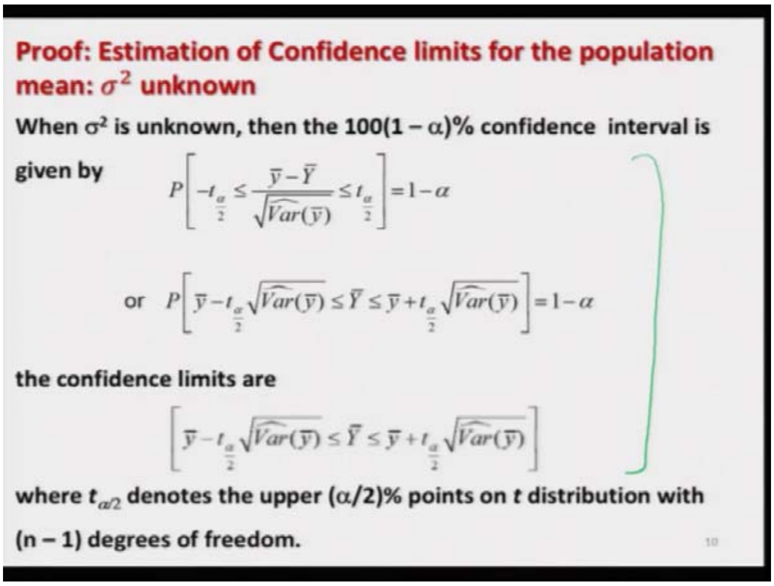
So, now you can see here we also have computed the confidence limit when the σ^2 is unknown to us right. So, that is the same result what we have used that I have written here more clearly for your records ok, that if the population is N($\bar{Y}$, σ^2). Then, this quantity follows a normal distribution and this quantity follows a t distribution this is just for your records right.

(Refer Slide Time: 26:04)



And here I am just trying to give you the algebra what I shown you manually, once again this is just for your records that when you are trying to look into the slides you should have a clear cut expression without any mistake right.

(Refer Slide Time: 26:30)



And, similarly this is for the second case where I have computed the confidence interval when σ^2 is unknown. So, my advice is that I have given you these expressions on this slide so, that you can try to do this algebra at least once with your own hand right. So, now I would like to stop here in this lecture, but now we have completed the theory part, we had a population whose population mean was unknown to us.

And usually in practice we are very much interested in measuring the central tendency of the data, we would usually interested in the average values. So, this is the place this is the theory which is going to help you. And these are the tools which are going to give you a clear cut answer. So, this population mean was unknown to us now what we have what we have concluded is that take a simple random sample.

But, before using the simple random sampling you have to make sure that the variation in the population with respect to the characteristic under study, that is not very high means if you try to go to the different parts of the population if you try to take up the samples, you must check. Well, you can do it by looking in the values from your experience or you can also conduct the test of hypothesis also.

But, you have to first make sure that this is happening only then you must use this simple random sampling. And if the variation is high if the variation is changing, then another option is this you can sub divide this population into group such that those groups have got the same variability. But that we are not discussing here we are going to talk about it, when we consider the stratified random sampling.

So, now you had unknown population you have found the arithmetic mean of the sample values, you also have found the involved variability. Whatever estimate you have given you have estimated the standard errors of the sample mean, they will give you an idea that how much is the variability involved in the values which you are trying to say; that means, if you say that my sample mean is 20.

So, you are saying that there is some variation also. So, the values are closely concentrated around 20 that is what we mean, when we try to connect the concept of mean and variance together. Now, when we are trying to connect the concept of mean and variance together, then confidence interval also comes into picture. Now, you can see that when we have found the lower limit and upper limit of the unknown parameter, then they involve two components one is sample mean and another is standard deviation or standard error.

So, confidence interval takes care of the value as well as variation and that is why the interpretation of point estimate and confidence interval differs point estimate says ok, this value is concentrated only at this particular point. And the confidence interval says that the value is concentrated in this interval. So, one is saying the value is concentrated at a point and say another is saying that the value is concentrated in the interval.

This usually you will see the midpoint of the confidence interval is the same as the point estimate, when you are assuming a symmetric distribution like as normal one. But, if you take a different distribution then this may not hold for example, if you try to construct the confidence interval for σ^2, then we have to assume that this follows a chi square distribution which is not symmetric. So, in that case this will not hold.

So, what you are trying to say in the present case under the normality, that this is your point estimate and all the values are lying in the interval around this point estimator. So that means, when I say around this is taking care of the standard deviation or standard error or in simple words the variability of $\bar{y}$. And, and what do you mean by variability of $\bar{y}$? You have seen that once you have a population of size N, then you are trying to take a sample of size n, different samples will give you, different values of sample mean.

So, now with these tools you can study that how much is the variation and what do you expect with what confidence or with what probability that your estimated value will lie between these two intervals. So, that is all about the theory, now in the next turn I will try to take up the issue that how to compute these things on the R software that is pretty simple straight forward.

So, you take care of yourself try to practice try to revise the concept. So, that you are 100 % prepared with this concept to employ on the R software. Because we need to write a small functions and we need to take interpretations. And I will see you in the next lecture till then. Good bye.