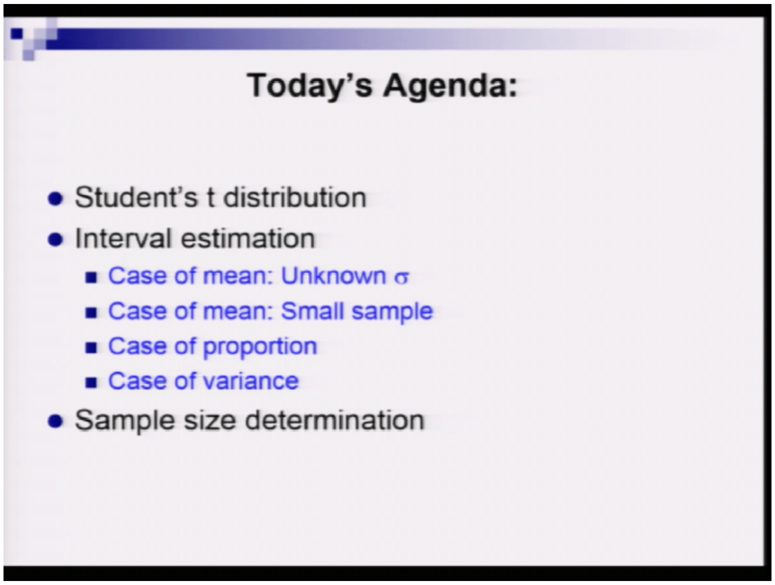
Applied Statistics and Econometrics
Professor Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology Kanpur
Lecture 10
Estimation (Part II)

Hello, friends. Welcome back to the lecture series on Applied Statistics and Econometrics. So let us have a look at today's agenda item first.

(Refer Slide Time: 00:26)



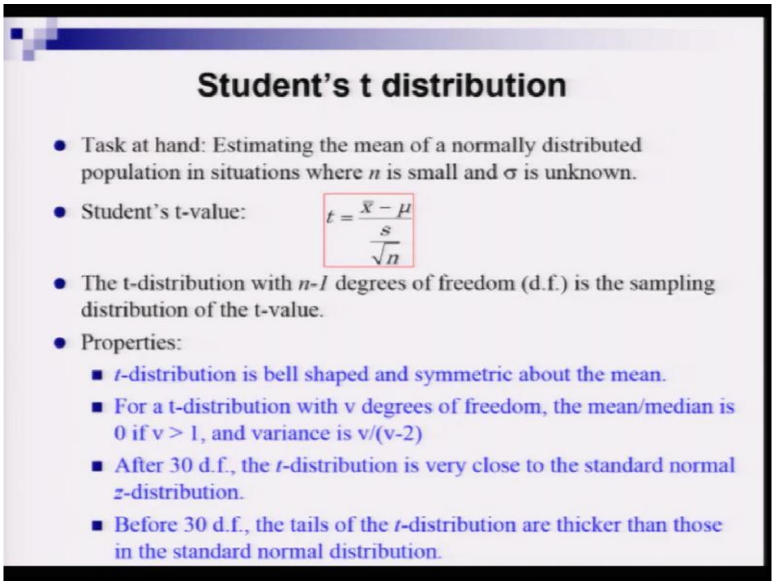
Today's Agenda:

- Student's t distribution
- Interval estimation
  - Case of mean: Unknown σ
  - Case of mean: Small sample
  - Case of proportion
  - Case of variance
- Sample size determination

So, we will start today's lecture with a brief discussion on a new probability distribution and that is called student's t distribution. And then, we will continue our discussion on interval estimation. We will discuss the case of mean for unknown sigma and then, case of mean for small sample. And then, we will also discuss the case of population proportion and case of variance. And finally, we are going to end today's lecture by have a discussion on sample size determination.

So, today's lecture will start with student's t distribution. Now, why we have to learn about one more distribution? It is because so far we have dealt with the case of large sample where the number of observations n is greater than 30, it is assumed. But if you have a small sample where n is less than 30, then the standard normal distribution that we have made use for confidence interval building will not work and I will show you, why. And, so, we need to figure out some solution and statisticians have provided us the solution. So, we are going to here work with the student's t distribution.

(Refer Slide Time: 01:43)



Student's t distribution

- Task at hand: Estimating the mean of a normally distributed population in situations where *n* is small and σ is unknown.
- Student's t-value: $\boxed{t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}}$
- The t-distribution with *n-1* degrees of freedom (d.f.) is the sampling distribution of the t-value.
- Properties:
  - *t*-distribution is bell shaped and symmetric about the mean.
  - For a t-distribution with v degrees of freedom, the mean/median is 0 if v > 1, and variance is v/(v-2)
  - After 30 d.f., the *t*-distribution is very close to the standard normal *z*-distribution.
  - Before 30 d.f., the tails of the *t*-distribution are thicker than those in the standard normal distribution.

So now, let us going to have a look at student's t distribution, its key features. So what is the task at hand? So our task is to estimate the mean of a normally distributed population in situations where n is small and sigma is unknown. And that is a quite common case. So to solve this problem, Gosset proposed a statistic called the student's t value, and that is defined as t and that is equal to the difference between sample statistic X bar and population feature or population parameter mu, the difference is divided by the root n.

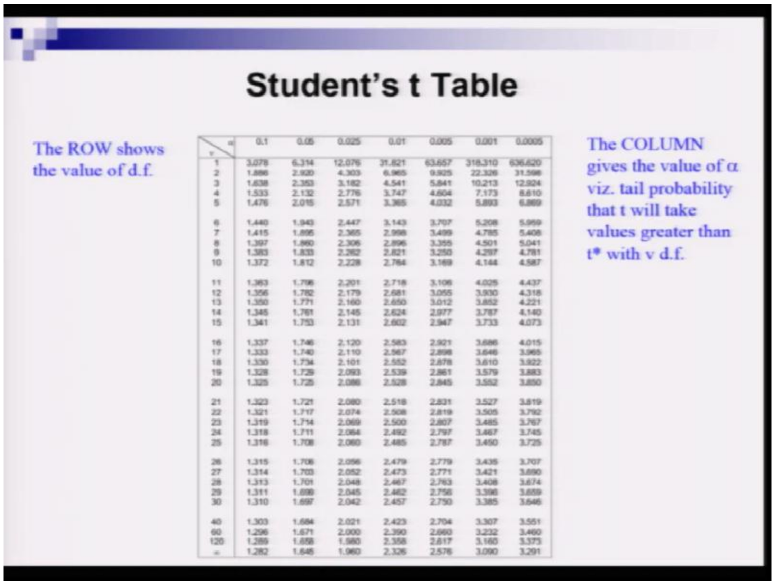
So this initially may look a bit complicated, but this t value that I am showing you, the formula here inside the red box, do not you find the similarity with some previous discussion that we had? If you remember, in the last class, when we were discussing the case of confidence interval, then actually, we had sigma in place of small s in the formula. So it is very similar to what we have been doing in the recent past.

So now, this t distribution, it is a new probability distribution with n minus 1 degrees of freedom. And this is the sampling distribution of the t value. Now, why degrees of freedom? Because, of course, you see the formula of t involves s the sample variance. And we all know that to calculate the sample variance, you have to first calculate the sample mean. So there are actually n minus 1 independent observations, we have had a discussion on degrees of freedom before. So hopefully, you remember that. And that is why t distribution has n minus 1 degrees of freedom.

Now, let us have a quick look at the most interesting properties of student’s t distribution. So, note that, the t distribution like standard normal distribution is also bell shaped and symmetric around mean. And we will derive, actually, we will not derive in the course, but what you can derive it easily that the mean and median for the t distribution is actually 0. So, now, we come to the second property and that is saying that for a t distribution with v degrees of freedom, the mean and median is 0, if v is greater than 1 and variance is v divided by v minus 2.

Now, the third property says that after 30 degrees of freedom, it implies when we are dealing with large sample, the t distribution is very close to the standard normal distribution. And if it is less than 30 degrees of freedom, it implies when we have small sample, then the tails of the t distribution are thicker than those in the standard normal distribution.

(Refer Slide Time: 04:41)



Student’s t Table

The ROW shows the value of d.f.

α v	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.076	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

The COLUMN gives the value of α viz. tail probability that t will take values greater than t* with v d.f.

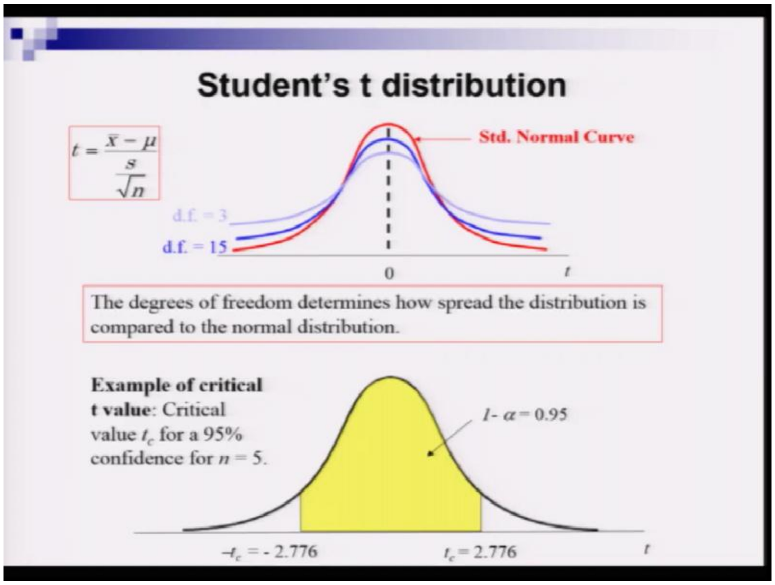
So now let us look at student’s t table and this student’s t table will help us to figure out the probabilities corresponding to different values of alpha and v. So as in the case of normal, we had taken the help of a standard normal table to compute various probabilities. In the case of t also, we are doing the same thing, because the PDF and CDF of t is also very formidable, look wise.

So it involves integrals and it will take a lot of time, if you want to derive everything on a piece of paper on hand or using a calculator. So, to help you, statisticians have made this table available. And let us see, how we can make use of this table. So, again, like the standard normal table, student's t table is also in a kind of matrix format. So, here we have rows and we have columns. So, row shows the values of the degrees of freedom, you see, it starts from 1, well, we will not get this kind of case in practice, because degrees of freedom 1 means that we are talking about n equal to 2, only 2 observations. But anyway, they have given us the values.

So, you see, this goes till 30, for all values of v and then after 30, there is a break and then they are reporting the cases of 40, 60 and 120. Actually, why, because after the v crosses the value 30, then actually you do not need a student's t table, you can actually make use of the standard normal table, as an approximation, it will be a very good approximation. So, that is why after 30 people do not report the different specific values of v, because it is going to be very lengthy otherwise.

So, now the column, so, the columns actually give us the value of alpha. And what is alpha? It is the tail probability that the random variable small t will take values greater than t star with v degrees of freedom. So, each and every cell here that you see for corresponding row and the choice of alpha, here these are not probabilities. You can very well see that these are not fractional numbers like bounded between 0 and 1. So, these are not the probabilities, these are basically the values of the t variable on the t axis. So, these are basically the cut points you can say on the t scale.

(Refer Slide Time: 07:36)

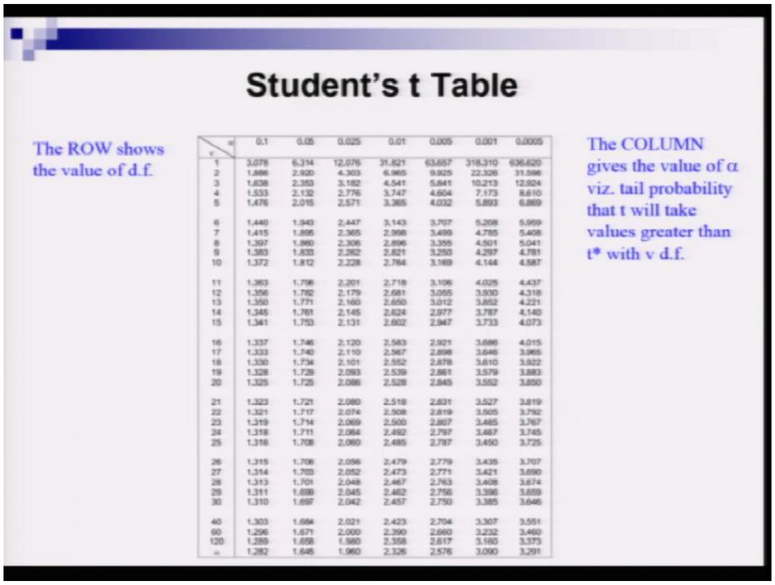


Student’s t distribution

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

The degrees of freedom determines how spread the distribution is compared to the normal distribution.

Example of critical t value: Critical value t_c for a 95% confidence for $n = 5$.



Student’s t Table

The ROW shows the value of d.f.

The COLUMN gives the value of α viz. tail probability that t will take values greater than t* with v d.f.

ν \ α	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.076	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Now, we are going to look at the picture of the student’s t curve and the standard normal curve and we are also going to draw some conclusions based on these curves. So first, let us look at the diagram that is in the upper part of the slide. So here, I am showing you the famous standard normal curve using this red colored bell shaped curve. And as you all know, that for standard normal the mean is 0, so this is centered at 0.

And now, against the standard normal curve, I am going to show you two t curves and they are for different degrees of freedom. Because, of course, as I have shown you in the previous table that the cutoff values on the t scale, actually this degrees of freedom determines the spread of the distribution, when we are comparing t distribution to the normal distribution case.

So here, we are showing you 2 different cases, one is for a very small sample, where the sample size is 4 and degrees of freedom is 3 and the corresponding curve is given in lilac or purple color or light blue color, whatever you want to call it. And then, I am showing you another t curve for degrees of freedom 15 and that is in thick blue or dark blue color. So as you can see here, that the t curves are much flatter towards the tail compared to the standard normal curve. And the gap between the standard normal curve and the t curve actually decreases as we increase degrees of freedom.

So what does it mean? So it means that as the degrees of freedom increases, it implies the sample size increases, then the probability that the t random variable will take a value higher than value t star, that can be approximated by the standard normal curve as the degrees of freedom increases. Because the difference in the probability for the approximated number from the standard normal table and the actual probability coming from the t table is going to be small, small and smaller as the degrees of freedom increases.

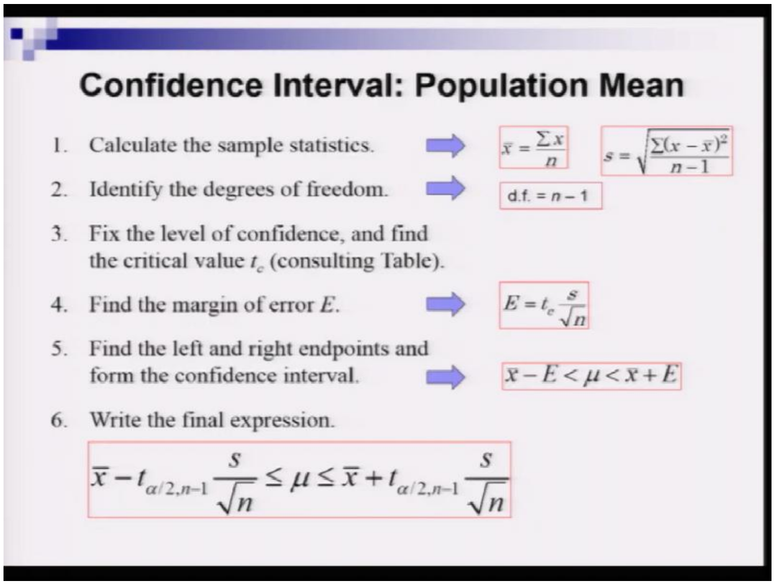
And finally, when we hit degrees of freedom 30 then actually in practical sense, there is no difference. So the bottom part of the slide now is going to show you an example how to find a critical t value and plot them in the t curve. So here, we are showing the case for n equal to 5 and if n is equal to 5 then we can assume or fix the level of confidence at 95 percent or 1 minus alpha that we have discussed before.

Now for these two values, let us see how we can figure out the probability from the student t table. So, remember that, we have 1 minus alpha equal to 0.95. So, that area actually is given by the yellow colored area and that is bounded by two different critical values minus tc and plus tc. And we have to figure out these threshold values minus tc and plus tc from the student’s t table.

Now how to actually find these values? So note that, from these 1 minus alpha equal to 0.95, that is an assumption, of course, from there, we can figure out that alpha actually takes value of 0.05, because alpha is basically a split into 2 parts here in this diagram. So, if they are equal, then each area will be alpha by 2 and that is going to be 0.025. So, now, we know that, we have to look at the table for the v value of 4 and alpha value of 0.025.

So, now, let me go back to the previous slide. And then, first, we go to the row and we spot 4 here. And then, we move rightwards across the columns and then we will look for the column title 0.025. And then, we stop and we see the t value reported there is 2.776. So, the example that we are showing you here in this diagram, for that example, the cutoff values or the critical values or the threshold values, whatever you want to call, it is going to be plus minus 2.776. And I have indicated these two values in the diagram.

(Refer Slide Time: 12:19)



Confidence Interval: Population Mean

1. Calculate the sample statistics. ⇨ $\bar{x} = \frac{\sum x}{n}$ $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$
2. Identify the degrees of freedom. ⇨ d.f. = $n-1$
3. Fix the level of confidence, and find the critical value t_c (consulting Table).
4. Find the margin of error E. ⇨ $E = t_c\frac{s}{\sqrt{n}}$
5. Find the left and right endpoints and form the confidence interval. ⇨ $\bar{x} - E < \mu < \bar{x} + E$
6. Write the final expression.

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

Now, we are going to discuss the case of building confidence interval for population mean from sample mean when the sample mean has come from a small sample. And it involves the use of a student’s t distribution and the student’s t table. So, let us have a look at different steps. So, the first step is to calculate the sample statistics and we know that what sample statistics we need to calculate. So, of course, the mean given by x bar and then the sample variance and then, from the sample variance, we have to calculate the sample standard deviation.

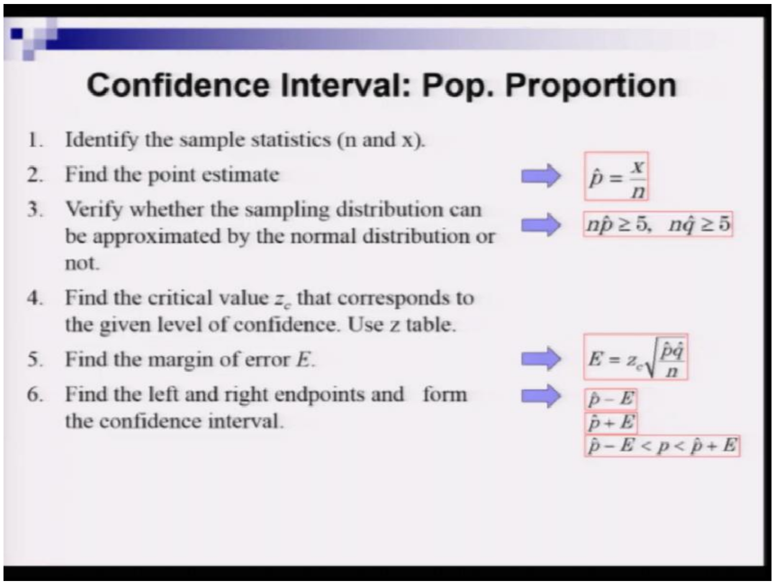
So, sample standard deviation is given by small s and that is basically the square root of sample variance and we have taken care of the degrees of freedom concerned, so that why we are dividing by n minus 1 here. And this is a good proxy when the sigma is not known. So, now, from step 1 we move to step 2, there we have to identify degrees of freedom. So, if we have n number of observations in the sample or in data points, then we have to subtract 1 to get the degrees of freedom.

Then we need to fix the level of confidence and then find the corresponding critical value consulting the student's t table, that I have shown you how to do. So, if you assume that 95 percent level of confidence, then you have to go back to the student's t table and then you have to figure out what would be the critical value corresponding to a particular degrees of freedom that you have calculated.

So, now, the fourth step talks about the determination or calculation of margin of error capital E. And that is basically given by the formula in the corresponding red box. So, that is basically the critical value that you observe from the t table shall be multiplied with s, the sample standard deviation and that then should be divided by root of n. And once the E is determined, it is time to find the left and right endpoints and then form the confidence interval.

So, once the E is calculated, we know what to do, we have to subtract E from the sample mean x bar and then that will be the lower limit of the confidence interval and then we have to add E to x bar as well to find the upper limit for the confidence interval. And if I want to write an expanded final expression, then it looks like kind of clumsy, but actually it is very simple because I have already explained you what to do. And you see that for mu the unknown population mean, I can write the final expression of the confidence interval in this small sample case.

(Refer Slide Time: 15:19)



Confidence Interval: Pop. Proportion

1. Identify the sample statistics (n and x).
2. Find the point estimate → $\hat{p} = \frac{x}{n}$
3. Verify whether the sampling distribution can be approximated by the normal distribution or not. → $n\hat{p} \geq 5, \quad n\hat{q} \geq 5$
4. Find the critical value z_c that corresponds to the given level of confidence. Use z table.
5. Find the margin of error E. → $E = z_c\sqrt{\frac{\hat{p}\hat{q}}{n}}$
6. Find the left and right endpoints and form the confidence interval. → $\hat{p} - E$
 $\hat{p} + E$
 $\hat{p} - E < p < \hat{p} + E$

So, now, let me move to the case of population proportion because population proportion is also a population feature that we may be interested to know about, but we have to make use of some small sample to infer about it. So, in that case how to proceed. So, again, we are going to look at a step by step, cookbook approach.

So, we are talking about 6 steps here. And the first step actually is to identify the sample statistics n and X. And then, from these two data, 1 has to figure out the point estimate and that is given by p hat equal to X divided by n. And X is of course, the number of events or some outcome and n is basically the total number of data points or observations that we have, out of which X number of outcomes have been realized.

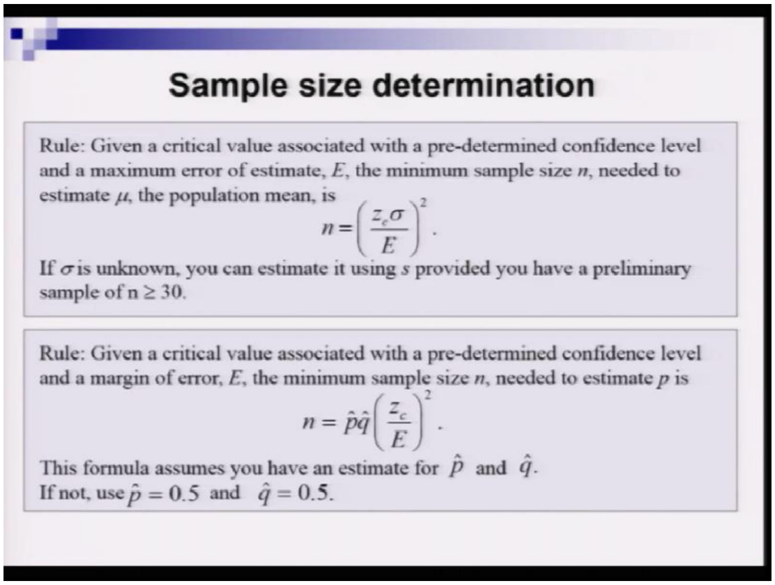
And then, in step 3, we have to find out whether the sampling distribution can be approximated by the normal distribution or not. So, when we realize the x’s and n’s, then they can be modeled actually in a binomial setup, but as we have studied earlier, and I told you that binomial can be approximated by the normal distribution in the large sample case and there are some statistical criterion to be made of course.

So first, we have to figure out whether we can actually approximate our sampling distribution for p hat by a normal distribution or not. And for that, what we have to do, we have to check whether n times p hat is greater than equal to 5 or not, and n times q hat, which is basically 1 minus p hat is greater than equal to 5 or not. And then, we have to find out the critical values zc, if we indeed get evidence for normal distribution assumption in this case. Then we have to fix the level of confidence and then, we have to consult the z table to get the critical values zc.

And then, finally, we can actually get the margin of error. And that is the formula given in the red box and that is capital E equal to the critical value zc times the square root of P hat, q hat divided by n. And once we get this E, next step is to find out the left and right endpoints of the confidence interval.

And that is basically done by subtracting E from p hat and that is basically will give me the left endpoint or the lower limit of the confidence interval and then if I add E to p hat then that will give me the right endpoint or the upper limit of the confidence interval. And finally, the confidence interval has to be written in one compact expression and that is basically the last box that is showing you.

(Refer Slide Time: 19:21)



Sample size determination

Rule: Given a critical value associated with a pre-determined confidence level and a maximum error of estimate, E, the minimum sample size n, needed to estimate μ, the population mean, is

$$n = \left(\frac{z_c \sigma}{E}\right)^2 .$$

If σ is unknown, you can estimate it using s provided you have a preliminary sample of $n \geq 30$.

Rule: Given a critical value associated with a pre-determined confidence level and a margin of error, E, the minimum sample size n, needed to estimate p is

$$n = \hat{p}\hat{q}\left(\frac{z_c}{E}\right)^2 .$$

This formula assumes you have an estimate for $\hat{p}$ and $\hat{q}$.
If not, use $\hat{p} = 0.5$ and $\hat{q} = 0.5$.

So, now, we are going to look at the sample size determination problem, again, because when we did it last time, probably it was done in a hasty manner or as you have not done confidence interval prior to it, probably you did not understand what is w? What do I mean by error margin and all? But now, we have had sufficient discussion on these concepts and probably, if I talk about it again, then probably it will be easier for you to appreciate and remember.

So, I will not spend a lot of time maybe a minute or two to discuss the sample size calculation formulas again. First, I will start with the case of the population mean, and of course, we want to get a proxy value of population mean by making use of the sample mean and if that is the case, then how we can get the sufficient number of in n or sample size. So, there is a rule.

So given a critical value associated with a predetermined confidence level, so note that you have to first fix the confidence level and then, of course, there will be a critical value zc associated with that fixed alpha. And then, maximum error of estimate capital E, we can actually figure out the minimum sample size requirement n to estimate mu from the population mean.

And that formula n is given here, it is not new to you, you have seen it before. So it is the critical values zc times sigma divided by E and the whole square. So if sigma is unknown, then okay, no problem. If you are dealing with a large sample, then you can use a proxy for it by calculating s the sample standard deviation that works in practice, when you have a large sample.

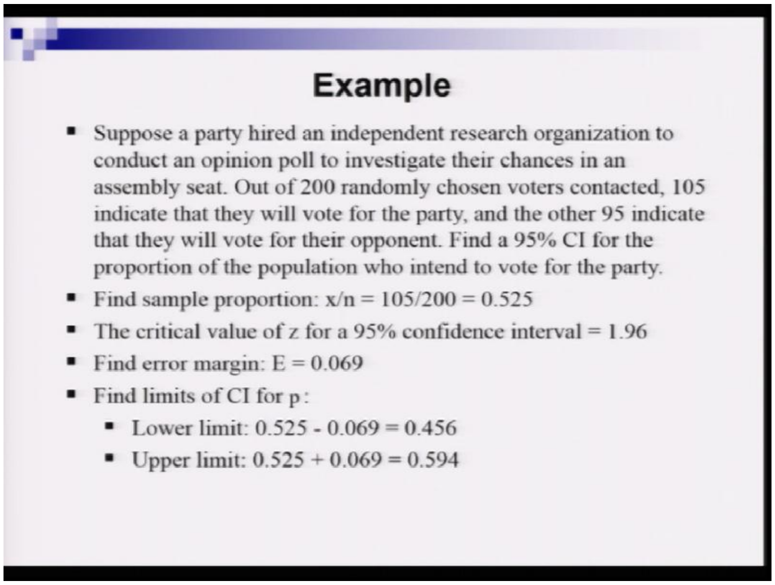
Next, we move on to another rule, and this is for the population proportion. So again, we have to assume that the critical value we know and that is for a predetermined confidence levels and a margin of error E, so you have to fix E and the alpha values. And then, the minimum sample size required to estimate p can be done using this formula n equal to p hat times q hat. And then zc, the critical value divided by e square.

Now this formula assumes that you have an estimate for p hat and q hat. But what if, if you do not have the estimates for p hat and q hat. So I told you earlier also, so if you do not have a priori information, you can also make use of 0.5 as the proxy value for p hat and q hat. So now we are going to look at one example where I am going to show you how to find confidence interval for unknown population proportion.

Now, where is this exactly useful? So you see, just right before the election in any country, many independent research organizations conduct surveys or public opinion polls, and they try to give some kind of an idea about the likeliness of a political party or a candidate to win certain election. So not only in India, in U. S. also and in many countries also, these public opinion polls are conducted.

Now, of course, sometimes their forecast matches, sometimes their forecast does not match the reality. But well, how do they address this problem? So for that, they, they actually make use of this statistical estimation theory. And I am going to show you one particular hypothetical example.

(Refer Slide Time: 22:42)



Example

- Suppose a party hired an independent research organization to conduct an opinion poll to investigate their chances in an assembly seat. Out of 200 randomly chosen voters contacted, 105 indicate that they will vote for the party, and the other 95 indicate that they will vote for their opponent. Find a 95% CI for the proportion of the population who intend to vote for the party.
- Find sample proportion: x/n = 105/200 = 0.525
- The critical value of z for a 95% confidence interval = 1.96
- Find error margin: E = 0.069
- Find limits of CI for p:
  - Lower limit: 0.525 - 0.069 = 0.456
  - Upper limit: 0.525 + 0.069 = 0.594

So let us assume that there is a political party and that party is going to contest 1 assembly seat and they want to know about their chances of winning that particular seat. So they have hired an independent research organization. So that research organization will conduct an opinion poll to investigate the party’s chances in that particular seat.

So to start that organization has collected 200 randomly chosen voters and they have surveyed these 200 people and out of these 105 indicated that they are going to vote for that party and other 95 indicate that they are going to vote for their opponent party. So in this case, as a statistician, how do you find a 95 percent confidence interval for the proportion of the population who actually intend to vote for that particular political party?

So, I have already shown you the steps in the previous slide. So if you follow the steps, the first step would be to find the sample proportion p hat, and that is given by x divided by n. So how many said yes, that is basically captured by the variable X and a particular value of that variable X is small x and that should be divided by the number of sample size or number of voters in this case. So you can get the sample proportion value of 0.525 or you can say 52.5 percent of the respondents have said that they are going to vote for the party.

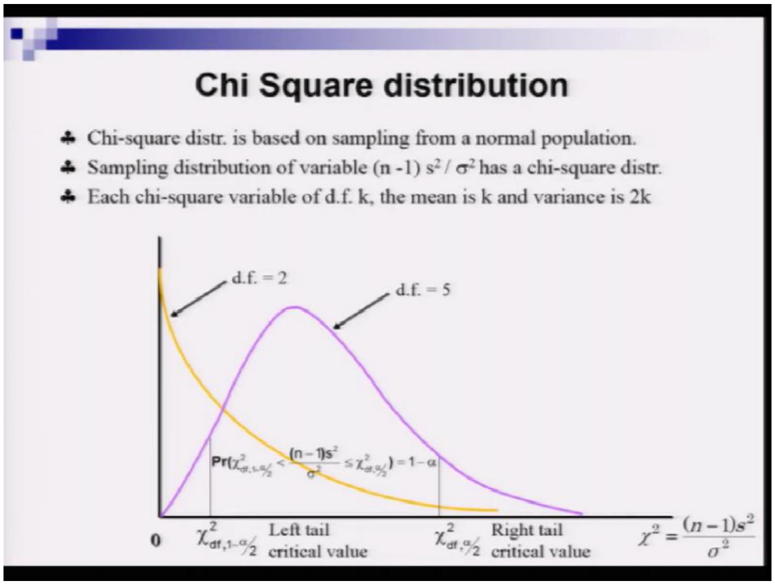
Now the next step would be to find the critical value of z. Now, why z? Because look, we have 200 that is a large sample and then, if you multiply that number 200 with the p hat, and also the q hat, which is 1 minus p hat, then you are going to see that it satisfies the criteria that we have set for normal approximation to discrete data in the case of population proportion. So in that case, we can actually safely apply the standard normal table to find the critical value for 95 percent confidence interval.

Now 95 percent, why 95 percent? Because that is the standard norm, you can make it 90 percent confidence interval as well. But for that, you do not have to go back to the standard normal table, again. Because remember, I have spoken about one magic figure and that is the z critical value for 95 percent confidence interval and that is 1.96. So this magic figure we are going to use here.

And not only here, later also, in the case of hypothesis testing, you will see that we will make use of this number again and again, whenever need arises. So, the next step would be to find the error margin, capital E, and I have again shown you the formula how to compute error margin, and if you plug values in that formula, you will get 0.069.

Now, the last step is to find the limits of the confidence interval for population parameter p. And we all know what to do. So to find the lower limit, we have to subtract that error margin from the sample proportion. So if you do that, then you get the value 0.456. And to obtain the upper limit of the confidence interval, you have to add that same error margin to the sample proportion. And if you do so, you get a value of 0.594. So this is the way you can actually construct the confidence interval for an unknown population parameter p in this particular example.

(Refer Slide Time: 27:23)



Chi Square distribution

- Chi-square distr. is based on sampling from a normal population.
- Sampling distribution of variable (n -1) s^2 / σ^2 has a chi-square distr.
- Each chi-square variable of d.f. k, the mean is k and variance is 2k

Now, before we move on to the confidence interval creation for unknown population variance, let us look at a digression. It is not a very big digression in real sense, because you will see that when we will be discussing the case of confidence interval building for population variance, we are going to make use of chi-square distribution. We have already introduced chi-square distribution, but that is not in this lecture. So just to make this lecture self sufficient, it is not a bad idea to have 1 minute or 2 minutes discussion on chi-square distribution first. And then look at the problem of confidence interval in the context of unknown population variance.

So just to remind you few basic things about chi-square distribution from the previous lecture, that chi-square distribution is based on sampling from a normal population. So that is a mandatory thing. So you can say that it is a strict assumption that chi-square distribution, if you want to deal then the sampling should be from a normal population. So the original data should at least look like a normal distribution, it should come from an approximately normal population.

Then the second point is that the sampling distribution of newly created random variable n minus 1 times the sample variance s square divided by population variance sigma square as a chi-square distribution. So here, note that, n is my sample size and s square is basically my sample variance. So that is the unbiased estimator of the population variance and sigma square is unknown population variance.

So the last point that I want to mention here is about the mean and the variance of the chi-square variable. So if you remember the previous lecture, chi-square random variables and the probability distributions, they always have associated degrees of freedom. And so if we assume that we have degrees of freedom k here, so the mean of the chi-square random variable will be k and variance will be 2k.

Now, let us look at a diagram of chi-square distribution. And I am going to show you various things through this one single diagram. So of course, in the x axis, I am going to plot various values of the continuous random variable chi-square, which is defined as n minus 1 a square whole divided by sigma square. And of course, along the y axis, I am going to plot the probability value so basically, it actually measures PDF.

Now, I am going to show you 2 different chi-square probability density function curves, one is in orange and that you see it is continuously falling. So that is basically for the degrees of freedom 2 and then, you see another PDF drawn and this is in the pink color, and that is basically constructed for degrees of freedom 5. Now, you see that as we have increased the degrees of freedom, there is a change in the shape of the PDF.

And these will continue to surprise you, as you keep moving the degrees of freedom values. If you increase the degrees freedom values say from 5 to 10, 10 to 15 and then 15 to 30, then you will see gradually, the chi-square distribution is going to be flatter and flatter. So the maximum distance of the curve from the x axis that is going to decrease and decrease and we know it is going to be more flat towards the tails. And note one thing that for smaller degrees of freedom chi-square distribution, actually, it is a positively skewed distribution. And that is on the shape part.

Now, we are going to look at interesting critical values and the corresponding probabilities associated with the chi-square distribution. So here, we are going to look at the left-tail and the right-tail critical value. Note that, as, so chi-square is a very special distribution, it is not like the normal or not like the standard normal, or even t, which are all symmetric. So actually, if you know one critical value, there is a mirror image of the other critical value and you can write in a plus minus zc or plus minus tc. Remember, we actually used the trick for z and t.

But this trick is unfortunately, not going to work with chi-square distribution, because chi-square is an asymmetric distribution. So you have to separately figured out the left-tail critical value and the right-tail critical value. So let us assume that our confidence limit is 95 percent. And that is basically 1 minus alpha, if you remember the previous set of lectures, we have already introduced 1 minus alpha and 95 percent is basically the most popular choice for 1 minus alpha. That is why I am talking about that.

But I am going to show you in this diagram a very general case, so, it pertains to any other choice for 1 minus alpha, if it is 99 percent, then also you can explain through this diagram, if it is 90 percent, then also you can through, then also you can explain through this particular diagram only. So, in general sense, we can write the probability of the chi-square random variable taking a value between the left-tail critical value and the right-tail critical value is 1 minus alpha.

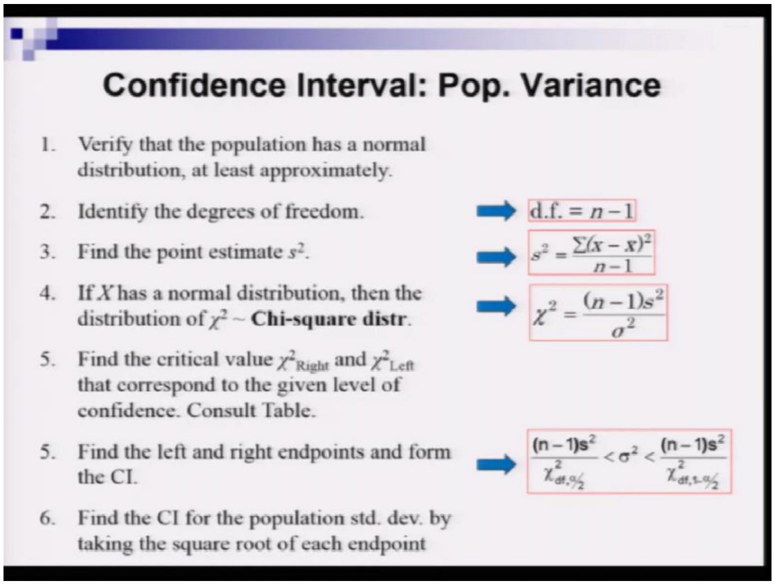
Now, how to find this left-tail critical value and the right-tail critical value? So say we have 1 minus alpha equal to 95 percent. So, in that case, alpha will take value 0.05. So, if you now subtract, alpha divided by 2, which is going to be 0.025 from 1, then actually you have a probability value of 0.975.

So, for the corresponding degrees of freedom, you have to find out the percentage point or the critical value chi-square the particular degrees of freedom with respect to the probability value of 0.975. So that will give you the left-tail critical value. And for the right-tail critical value, it is simple. So it is going to correspond to the probability value of 0.025 after dividing alpha by 2 and if you look at the corresponding degrees of freedom value, then you can figure out the percentage point from the corresponding PDF.

But note that, this is very complicated, because with the change in degrees of freedom, there is a change in the shape of the PDF of chi-square random variable. So, how can we find out the percentage points or the threshold values are the critical values whatever you want to call them, and I am not showing you the form of PDF of chi-square distribution, because that is very messy and complicated.

So, it is very difficult indeed to get these critical values yourself by calculating on a piece of paper. So, the statisticians have again come up with a statistical table and they have actually provided us these different percentage points or the critical values for combinations of degrees of freedom and alpha value.

(Refer Slide Time: 34:49)



Confidence Interval: Pop. Variance

1. Verify that the population has a normal distribution, at least approximately.
2. Identify the degrees of freedom. → $\boxed{\text{d.f.} = n-1}$
3. Find the point estimate s^2. → $\boxed{s^2 = \frac{\Sigma(x-x)^2}{n-1}}$
4. If X has a normal distribution, then the distribution of χ^2 ~ **Chi-square distr**. → $\boxed{\chi^2 = \frac{(n-1)s^2}{\sigma^2}}$
5. Find the critical value χ^2_{Right} and χ^2_{Left} that correspond to the given level of confidence. Consult Table.
5. Find the left and right endpoints and form the CI. → $\boxed{\frac{(n-1)s^2}{\chi^2_{df,\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{df,1-\alpha/2}}}$
6. Find the CI for the population std. dev. by taking the square root of each endpoint

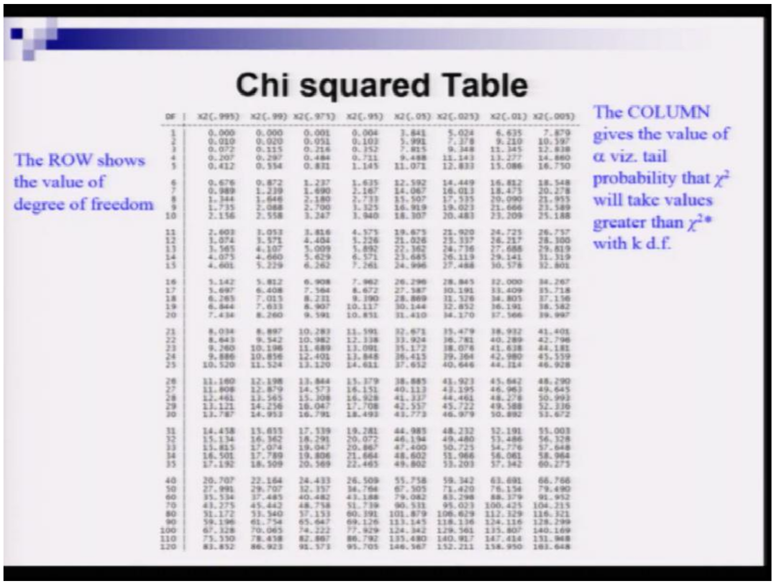
So, we this brief discussion on the chi-square random variable and chi-square distribution, let us now proceed with the discussion on confidence interval of the population variance. So, as I told you, that we are going to make use of chi-square distribution here, so it is indeed very important to check whether the population has a normal distribution or at least approximately it looks like it is tending towards normal curve and all, the relative frequency curve is tending to the, tending towards the normal curve or not.

And once that is verified, you actually have to identify the degrees of freedom that is simple. So you have to subtract 1 from the sample size and that is your df. And then, you have to find out the point estimate sample variance and we all know how to do it, it is not new to us. And then, if the random variable x has a normal distribution, then the derived random variable chi-square follows a chi-square distribution, we have spoken about this in the previous slide only. So it is also not new to you, this is basically not a step, actually this is just something in between that you need to remember.

And that actually tells you that you have to consult the chi-square table and you have to find the critical values, chi-square right and chi-square left that correspond to a given level of confidence and the degrees of freedom. And once this left and right critical values are obtained, you can actually find the confidence interval by following this particular formula that I am showing you in the very last box in this slide. And you see that this has 2 endpoints.

Now, the lower endpoint actually is basically ratio and its ratio of n minus 1 times the sample variance and the chi-square right critical value. And the upper limit of the confidence interval is again ratio and this time n minus 1 times sample variances is divided by the chi-square left critical value. So, once this confidence interval is built for the population variance, it is very easy to find out the confidence interval for the population standard deviation as well, because you just have to take the positive square root of each endpoint and then you will find the confidence interval four sigma, the population standard deviation.

(Refer Slide Time: 37:29)



Chi squared Table

The ROW shows the value of degree of freedom

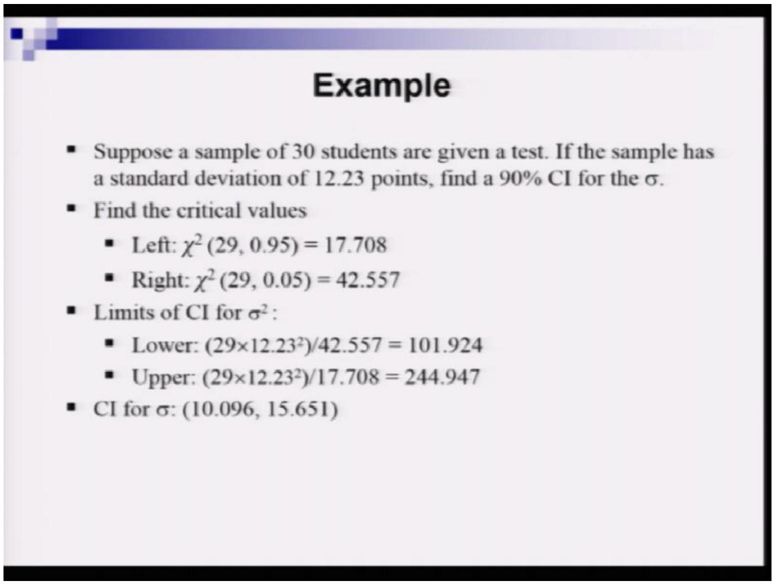
DF	x2(.995)	x2(.99)	x2(.975)	x2(.95)	x2(.05)	x2(.025)	x2(.01)	x2(.005)
1	0.000	0.000	0.001	0.004	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672
31	14.458	15.655	17.539	19.281	44.985	48.232	52.191	55.003
32	15.134	16.362	18.291	20.072	46.194	49.480	53.486	56.328
33	15.815	17.074	19.047	20.867	47.400	50.725	54.776	57.648
34	16.501	17.789	19.806	21.664	48.602	51.966	56.061	58.964
35	17.192	18.509	20.569	22.465	49.802	53.203	57.342	60.275
40	20.707	22.164	24.433	26.509	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	124.342	129.561	135.807	140.169
110	75.550	78.458	82.867	86.792	135.480	140.917	147.414	151.948
120	83.852	86.923	91.573	95.705	146.567	152.211	158.950	163.648

The COLUMN gives the value of α viz. tail probability that χ^2 will take values greater than χ^{2*} with k d.f.

So, now, I am going to give you this picture of chi-square table which is coming from a textbook. Now, here again, you see like other tables I have shown you, this is also in matrix form and row actually, a particular row shows the values of the degrees of freedom and then a particular column gives me the value of alpha, which is basically the tail probability that the chi-square random variable will take values greater than a particular threshold value chi-square say chi-square star with k degrees of freedom.

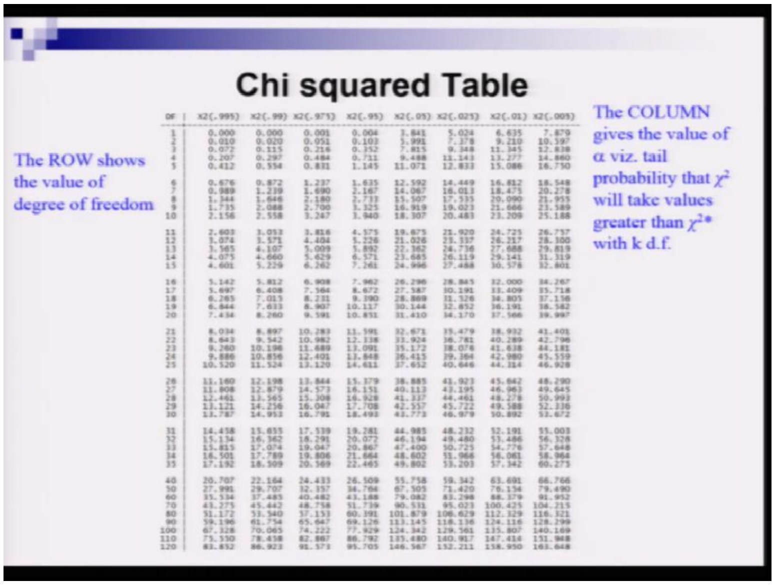
So, now, with this table, let me look at an example, and through that example, I will also show you how to consult this particular table and get the critical values or the threshold values from chi-square distribution.

(Refer Slide Time: 38:25)



Example

- Suppose a sample of 30 students are given a test. If the sample has a standard deviation of 12.23 points, find a 90% CI for the σ.
- Find the critical values
  - Left: χ^2 (29, 0.95) = 17.708
  - Right: χ^2 (29, 0.05) = 42.557
- Limits of CI for σ^2 :
  - Lower: $(29\times12.23^2)/42.557 = 101.924$
  - Upper: $(29\times12.23^2)/17.708 = 244.947$
- CI for σ: (10.096, 15.651)



Chi squared Table

The ROW shows the value of degree of freedom

DF	x2(.995)	x2(.99)	x2(.975)	x2(.95)	x2(.05)	x2(.025)	x2(.01)	x2(.005)
1	0.000	0.000	0.001	0.004	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672
31	14.458	15.655	17.539	19.281	44.985	48.232	52.191	55.003
32	15.134	16.362	18.291	20.072	46.194	49.480	53.486	56.328
33	15.815	17.074	19.047	20.867	47.400	50.725	54.776	57.648
34	16.501	17.789	19.806	21.664	48.602	51.966	56.061	58.964
35	17.192	18.509	20.569	22.465	49.802	53.203	57.342	60.275
40	20.707	22.164	24.433	26.509	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	124.342	129.561	135.807	140.169
110	75.550	78.458	82.867	86.792	135.480	140.917	147.414	151.948
120	83.852	86.923	91.573	95.705	146.567	152.211	158.950	163.648

The COLUMN gives the value of α viz. tail probability that χ^2 will take values greater than χ^{2*} with k d.f.

So, here is our example. So suppose you have a sample of 30 students who were subjected to a particular test or exam and the sample actually results in standard deviation of 12.23 points. So, in that case, the task is to find out a 90 percent confidence interval for the sigma, the population standard deviation. So, now, we are going to first look at the critical values from the table because sample standard deviation is already calculated for us. So, there is no need to find out the sample variance.

Of course, 30 is a large number. So, again we can assume that well approximately it is a normal population. It may not be but let us assume because it is a simple example. So, now, we have to find the critical value. So, we have to find the left critical value and the right critical value from the chi-square distribution and for that we need to consult the table. So, basically for the left critical value, the value is from the chi-square table where degrees of freedom is 29 and alpha is 0.95.

Why it is 0.95? Because you remember we were discussing about 1 minus alpha divided by 2. Here, we are looking for the 90 percent confidence interval that is why. So residual is 10 percent. So now you divide this 10 percent into 2 equal parts, so it will become 5 percent towards the tail? So you have to look at the table for the alpha value of 0.95 at the left-tail. And then, for the right-tail, you have to consider alpha value 0.05.

So, now let us go back to the table. So, now you come down the rows and then you start where you figure out df takes value 29. And then you move towards your right across the columns. So my first task is to find out the chi-square value for alpha of 0.95. And from this table, you can see that the value is 17.708. Now my next task is to figure out the value for alpha 0.05. And you see, it is basically the very next value in that same row and that is 42.557.

So, from this table, we have got the values and now, the formula I have already shown you, after plugging these values of left and right critical values in the formula for confidence interval for sigma square, you get the lower limit value as 101.924 and the upper limit value as 244.947. So if you take square root of these 2 numbers finally, you land up with the confidence interval for the population standard deviation and of course, that is a lower number 10.0. So, now, we are done with our discussion on estimation theory. From the next lecture, we are going to look at the other part of statistical inference, which is called hypothesis testing. Thank you.