Artificial Intelligence (AI) for Investments
Prof. Abhinava Tripathi
Department of Industrial and Management Engineering
Indian Institute of Technology – Kanpur

Lecture – 18
Finding the Mean of the Population

(Refer Slide Time: 00:13)

Finding the Mean of the Population

(Video Starts: 00:15) Recall that we found that the sample mean $\bar{X}$ was 2.2 ppm and we want to validate whether this mean of the sample will actually be the mean of population. Now, we understand one thing for sure. To find out the population mean exactly from the sample mean with zero error is near impossible. We will obviously make some error that is at the best we can say that the population mean should be 2.2 ± some error.

So, we might be able to conclude that the population mean will be 2.2 ppm ± 0.1 or 2.2 ppm ± 0.2 or something like that. But that is still useful to us, the aim of the problem at hand is to be able to tell whether the amount of lead in a noodle packet is greater than or less than 2.5 ppm. So, even if we can say that the mean of the population is 2.2 ppm with an error of ± 0.2.

We would know for sure that the mean lead content will be between 2 and 2.4 which is definitely less than 2.5. Now, in order to see if the sample is indeed a true representation of the population, let us do a simple experiment, let us go back to our noodles example where we already have a population whose parameters are known to us. Consider that you have the complete set of 30000 noodle packets that this company has manufactured, we are given that N = 30000.

Now, each data point in the CSV file is the lead content in ppm for that noodle packet. Now, if we check the mean of this data we get it as 2.199 and the standard deviation comes out to be 0.132. These are population mean and standard deviation values which are represented as μ and σ respectively. But for now, let us take a small sample size of 5 and see what it is parameters are?

So, we randomly choose five noodle packets out of 30000 and we find out that the mean of their lead content is 2.145 ppm. But this is a little far away from the actual population mean which was 2.199. It might have happened that we coincidentally choose noodle packets with lower lead content, let us take another example this time the mean comes out to be 2.27 ppm which is quite higher than the population mean.

Again, it is look like we coincidentally choose noodle packets having higher lead content. So, to normalize this effect of getting biased samples let us choose 100 samples instead of 2 this number could have been 70, 80 or 90. But for now, let us move with 100 samples, note that we are talking about number of samples here and not the sample size. Sample size here is still 5. If we plot this sample means on a graph, we find out that the distribution of the sample means looks like a normal distribution.

And if you see the centre point of this distribution it is around 2.2 which is in fact very close to our population mean. This certainly opens up many avenues, we saw that when we took a large number of samples and plotted the mean of each sample on the graph. We got a distribution that was quite close to a normal distribution as we keep on increasing the sample size, the sample mean will keep approaching the population mean.

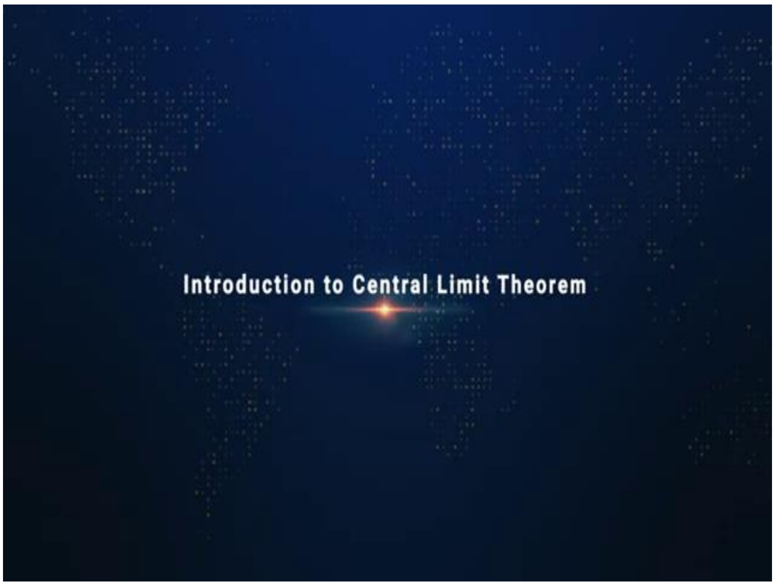
Now, when I say increasing the sample size, I mean the number of noodle packets in each sample and not the number of samples that we took. So, earlier our sample size was 5 if we increase it to 10 keeping the number of samples the same that is 100, the distribution becomes thinner and closer to a normal distribution. In fact, if you keep increasing the sample size, the distribution keeps getting narrower.

And the distribution seems to look more and more like a normal distribution. In fact, when you increase the sample size to 30, it becomes almost perfectly normally distributed with the means entered around the population mean. What does this mean for us? This experiment that we performed right now is the basis for central limit theorem. The central limit theorem states that when you take a large number of samples, the mean of the sampling distribution that is formed will be approximately equal to the population mean.

The second part of the theorem states that the standard deviation of the sampling distribution will be equal to σ. Which is population standard deviation divided by the square root of n, where n is the sample size. Just remember that the sampling distribution standard deviation would be σ by square root of n. Finally, the central limit theorem states that if the sample size that you take is greater than 30.

The sampling distribution will become normally distributed these three findings in short constitute the central limit theorem. **(Video Ends: 04:35)**

(Refer Slide Time: 04:36)



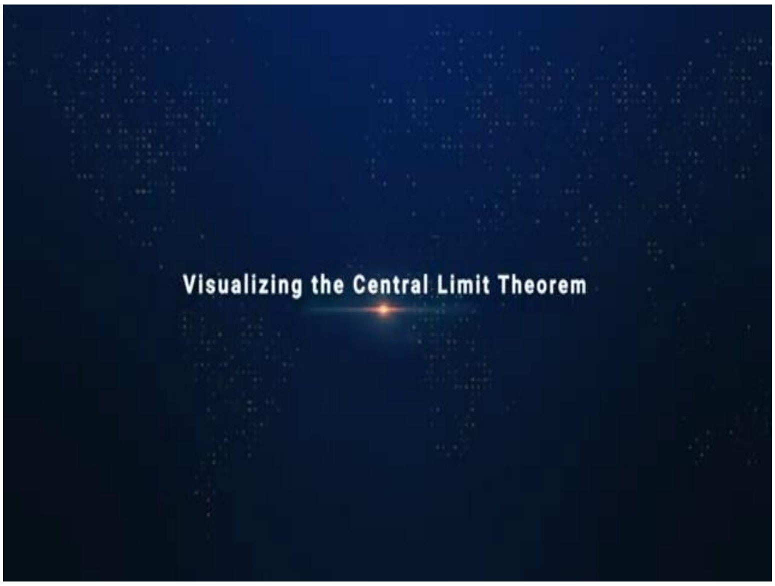
Introduction to Central Limit Theorem

(Video Starts: 04:38) Let us first look at a few terms and notations, the whole bunch of packets that is 30000 packets is called the population and the small collection of packets that we select from it is called a sample. If you look at this table you can find all the notations and formulae for both the population and the sample. The population size is denoted by N, its mean by μ and its standard deviation by σ.

The variance in turn is σ squared, the formula of the sample have a small twist here. The sample size is denoted by lowercase n and the mean by $\bar{X}$ which is nothing but the sum of all the observations in the sample divide by the total number of observations which is the standard way of computing the mean. So, there is nothing different about calculating the mean for the sample.

This $\bar{X}$ after measuring the lead content was found to be 2.2 ppm parts per million. But now, if you come to the sample variance S squared, you can see in the denominator you have n – 1 instead of n. **(Video Ends: 05:38)**

(Refer Slide Time: 05:40)



Visualizing the Central Limit Theorem

To concretize your understanding of the central limit theorem, let us try and visualize the central limit theorem. **(Video Starts: 05:45)** We plot means sampled from a non-normal population with 100 samples for different sample sizes. The four plots corresponding to our four sample sizes are 1, 5, 10 and 30. Examine the four density plots for the sample that are coloured in blue and compare with the normal density distribution in red here.

We start with a sample size of 1, the resulting curve deviates a lot from the red curve as we keep on increasing the sample size from 1 to 5, 10 and 30 the blue curve comes closer and closer to the red curve. This graph is now starting to look like a normal distribution and the reason for this is the central limit theorem. And what this states, is that if you plot the sampling distribution. Then this distribution approach is a normal distribution regardless of what your parent population is.

And this is especially true when the sample size is above 30. So, here you can see that even with a sample size of 30 the distribution almost approximates a normal distribution. If you look different means from different sample sizes they appear to be similar, you see that this is also the mean of the population. And then you come to the third graph and here you see that this is the mean of the sampling distribution.

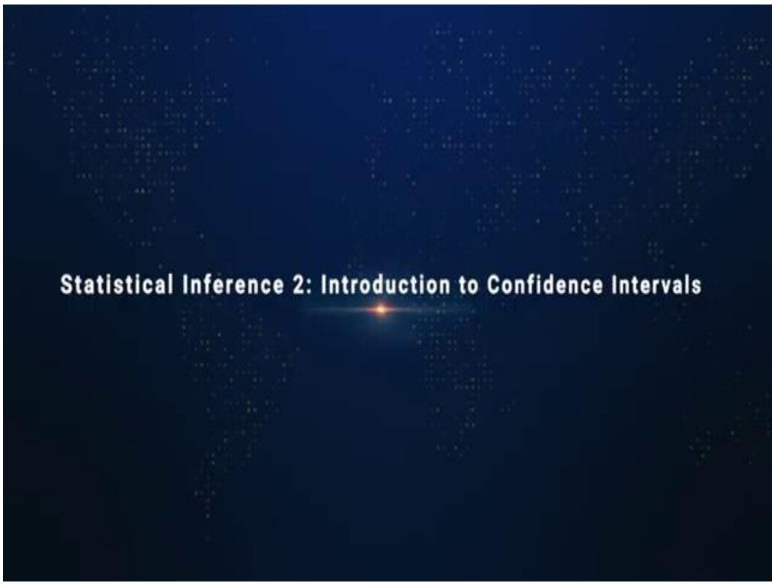
And you can see that both these means are almost equal, the standard deviation of sampling distribution is the standard deviation of population divided by square root of sample size. As you keep on increasing the sample size, the curve will start to look like normal distribution. Now, as you increase the size of N, you are bringing the N value closer to the population size value.

You will start seeing that the sampling distribution of sample means will start looking more like a normal distribution. And you see that the mean is becoming more equal to the population mean. What about the standard deviation of the sampling distribution? Compare the standard deviation value of the population which is 28.80 with the standard deviation of the sampling distribution.

The standard deviation of the sampling distribution approaches closer to the standard deviation of population divided by the root of n. For example, the standard deviation of sampling distribution for n = 30 is 5.3 which is closer to 28.8 upon root 30 which is equal to 5.26. Also see that the normal distribution became thinner as the sample size increases this is because the standard deviation decreases remember here the standard deviation is σ by root n.

So, obviously σ by root 5 is going to be higher as compared to σ by root 30 which is why this graph seen earlier becomes more packed. And now what do you think will happen, if I further increase the sample size say from 9, n = 30, to n = 50? The normal distribution will become even narrower. So, this is your centre limit theorem. **(Video Ends: 08:22)**

(Refer Slide Time: 08:23)


Statistical Inference 2: Introduction to Confidence Intervals

Statistical inference 2 introduction to confidence intervals. **(Video Starts: 08:27)** Let us go back to the noodles example and see if we are able to derive conclusions about the population using the sample. If you recall we took a sample of 100 packets and find out that its sample mean was 2.2 ppm and standard deviation was 0.7 ppm making use of what you learned about sampling distributions.

We can directly assume that the sample mean of 2.2 ppm that we have calculated belongs to one of the infinite possible sample means that is present in the probability distribution. And this probability distribution is nothing but our sampling distribution. Remember that the sampling distribution is nothing but the distribution of all the possible sample means that can be generated from this population.

And this means that our sample mean which was 2.2 ppm is definitely one of the sample means present in this distribution. As per the central limit theorem if the sample size is greater than 30 then we can assume that the sampling distribution is normally distributed with mean equal to the population mean which is unknown in this case. And standard deviation equal to σ by square root of n, where σ is unknown and n is the sample size.

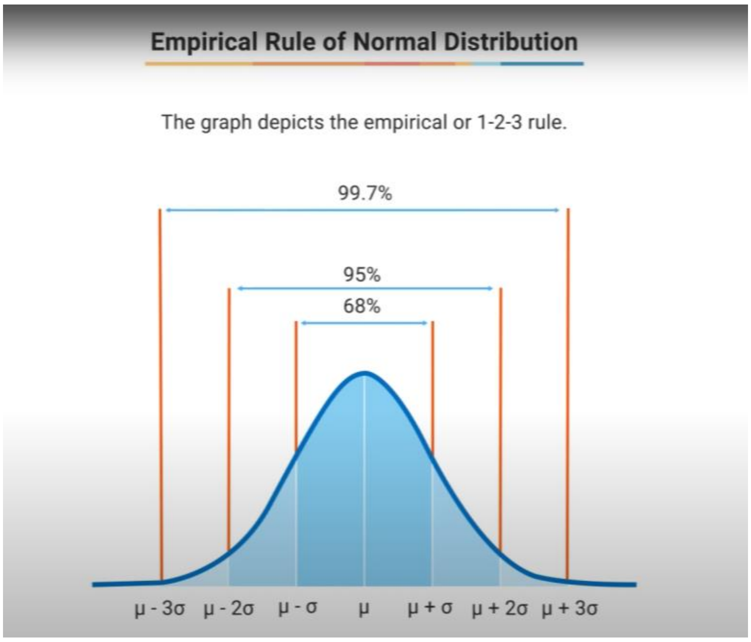
So, with this we can conclude that our sample mean which was 2.2 ppm belongs to a normal distribution with mean equal to the population mean. And standard deviation equal to σ by root n, again this distribution is called the sampling distribution. And the reason we know that the sampling distribution will be a normal distribution is because of the central limit theorem.

So, the sampling distribution is like this and our sample might lie somewhere in this distribution so, let us start with the estimation. Right now, we know that the mean of this normal distribution is equal to μ which is the population mean. But we do not quite know what this mean is? So, let us come back to that later. This distribution standard deviation is σ. Now, remember that σ is the population standard deviation but wait we do not know what the population mean is?

How are we supposed to know the population standard deviation? Now, there may be some cases where prior to sampling, you were able to obtain a good estimate about the population standard deviation. In that case you can use the value of σ as it is but what if that is not the case. In fact, in most practical scenarios such as our noodle example σ is usually unknown.

In such cases, we will need to substitute the sample standard deviation in place of the unknown population standard deviation σ. The sample standard deviation is represented by S and which we found out to be 0.7. Hence replacing S in the formula and replacing n with 100 which is the sample size, we get the standard deviation of the normal distribution as 0.07.

Note that here we are using the sample standard deviation as a substitute for the population standard deviation. This is something we do since in most cases we will not know the population standard deviation. Now, come back to our example we know that the distribution is normal. And thus, we can make use of the different properties that we have learned about the normal distribution.



Recall that one of them was the empirical rule also known as 1-2-3 rule so, let us quickly revisit that rule. The probability of X lying between my $\mu - \sigma$ and $\mu + \sigma$ is around 68 percent one. Two, the probability of X lying between $\mu - 2\ \sigma$ and $\mu + 2\ \sigma$ is around 95 percent and three the probability of X line between $\mu - 3\ \sigma$ and $\mu + 3\ \sigma$ is 99.7 percent.

Using this rule, we can say that the probability that the sample mean which is 2.2 will lie between μ – 2 standard deviation to μ + 2 standard deviation will be equal to 95 percent. Note that μ is the population mean and it is the same as the mean of the sampling distribution which we are making use of. Standard deviation is something we just found out using σ by root n which we calculated as 0.07.

So, going back to our problem we said that 1-2-3 rule states that 95 percent of the values in a normal distribution lie between mean ± 2 standard deviations. We do know that the standard deviation of this normal distribution is 0.07 which we just found out. Hence this equation after replacing 0.07 becomes probability of $\mu - 0.14$ is less than 2.2 which is less than $\mu + 0.14 = 95$ percent that is probability of 2.2 lying between $\mu - 0.14$ and $\mu + 0.14$ is 95 percent.

Now, using this equation can you tell what will be the value of mean, μ lying between 2.2 – 0.14 to 2.2 + 0.14? Think about it. It is the same thing you simply just rearrange the terms of the expression probability of $\mu - 0.14$ less than 2.2 less than μ + 1 0.14. So, you got the expression probability of 2.2 – 0.14 less than μ which is less than 2.2 + 1 0.14. What we did was, we simply carried over μ to this side and 2.2 to either sides.

So, since both of these are the same equation, you can say that probability of 2.2 – 0.14 is less than μ and μ is less than 2.2 + 0.14 = 95 percent or in other words probability of μ being lying between 2.06 to 2.34 is 95 percent. Hence, you can say that with 95 percent probability that the mean will lie between 2.06 to 2.34 and you found out this value just by using the 1-2-3 without even knowing the value of μ.

So, before moving further let us get some terms state, the probability associated with this claim is called the confidence interval. In this case since we are concluding about the population mean with 95 percent probability, we can say that the confidence level is 95 percent. In many cases instead of confidence level, you may also be given the level of significance which is denoted by alpha.

The significance level is given by 1 – confidence level so, if the confidence level is 95 percent or 0.95 then the level of significance will be 1 – 0.95 which is 5 percent or 0.05. Similarly, if you are given the level of significance or told that alpha is equal to some value say 0.05 then you can say that the confidence level which we want to conclude is going to be 95 percent. Next you have the margin of error which is the maximum error made in the sample mean.

In this case, margin of error is 0.07 into 2 which is 0.14 and finally, the final interval of value that you get is called the confidence interval. In this case, our confidence interval is 2.06 to 2.34. The upper bound of the confidence interval that is 2.34. We can conclude that the noodles that are being manufactured on average do not contain higher than the accepted threshold of lead which is 2.5 ppm.

And we did all of this using just a sample a few noodle packets which is amazing. **(Video Ends: 15:16)**

(Refer Slide Time: 15:17)

Statistical Inference 3: Confidence Interval Construction

Welcome to the statistical inference 3: Confidence interval construction. **(Video Starts: 15:21)** If you look back at everything you learned till now, it was all about solving a simple problem, calculating an interval for our population mean. And in order to do this you learned something about sampling, sampling distributions, central limit theorem and many other things. So, in this discussion let us actually generalize this approach.

So that if we give us any large population and ask us to give an estimation about the population mean. We know, how to solve this problem. And come up with an unbiased confidence interval. First you need to decide whether it is practical to actually obtain all the values in the population? So that you can calculate the mean accurately because in such a case your problem is already solved.

But suppose getting this population data is not actually feasible at all which is usually the case this is where your inferential statistics knowledge comes into play. There is no way to know the actual population mean and standard deviation. So, for this population that we have with us beside its mean is μ and standard deviation is σ. And the objective of this problem is to get an estimate of the population mean that is μ, also most important of all this population can follow any distribution.

It need not be a normal distribution, it can be a uniform distribution or any random distribution that comes to your mind. So, how do you start this problem? You start by collecting a sample, you have even learned multiple sampling techniques such as simple random sampling and stratified sampling. So, it is up to you to choose an appropriate sampling technique.

Such as that whatever sample you have can be considered a proper representation of your population. So, let us say you have selected your sample with the size of small n. Now, for the sample calculating various parameters such as mean and standard deviation are possible so, let us calculate those. Let us call the sample mean as $\bar{\mathrm{X}}$, sample standard deviation as S and sample size is n.

Now, remember that we know the values of $\bar{\mathrm{X}}$, S and n and using these values, we need to somehow find an interval for our population mean μ. To solve this problem, you learned that we can make use of central limit theorem, let us recall this theorem. The central limit theorem is based on a probability distribution known as sampling distribution. So, firstly what is the sampling distribution?

Let us say, we have our population and we start collecting different sample of size n. And let us say you collect your first sample calculate, the sample mean call it $\bar{X}_1$

1 then calculate your second sample calculate the sample mean call it $\bar{X}_2$and so on. Now, you can continue this process an infinite number of times and you will be left with an infinite number of sample means that are $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$and so on all the way up to X̄ infinity.

Now, suppose you consider your random variable X such that it is the sample mean then the possible outcomes of X will be $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$ and so on up till X̄ infinity. Then the probability distribution that this random variable X follows is called sampling distribution. You will also hear some people saying it sampling distribution of sample means which is the same thing so, we have the sampling distribution.

Now, our central limit theorem says that this sampling distribution that you have generated is going to be more and more like a normal distribution as you increase the sample size is n. In fact, we assume any sampling distribution has a normal distribution if the sample size is greater than 30 which has a mean of μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. We have selected our sample whose sample mean is X̄ and sample standard deviation is S and the sample size is n.

Now, using these values you need to come up with an interval for our population. Now, if you consider a confidence level of Y percent and if we apply the central limit theorem. We can derive that our population mean lies in the range $\bar{x} - z^* \frac{s}{\sqrt{n}}$ to $\bar{x} + z^* \frac{s}{\sqrt{n}}$, where Z is the critical value associated with Y percent confidence interval. For example, if the confidence level at which you are looking to estimate the population parameters is 90 percent.

Then the value of Z will be 1.65 if the confidence level is 95 percent then the value of Z will be 1.96 and for 99 percent it will be 2.58. These are the standards Z values for confidence levels traditionally taken in the industry. You can choose the confidence level according to your problem for example in our noodles problem. We want to be highly confident in our results, since the entire company's business depends on this.

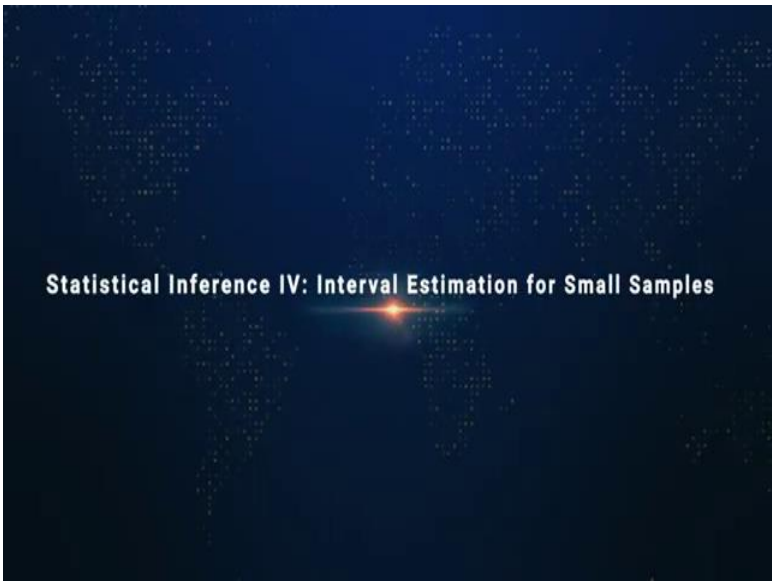
And hence in such cases choosing a confidence level of 95 percent or 99 makes more sense. On the other hand, if you are even 90 percent confident that making a change in your home page will result in generating more leads, you can probably go ahead with it and make the change. So, based on different scenarios, you choose different confidence levels. To make it easy for us to remember there are five step approach that you can use for estimating the mean of a population.

First you collect a sample of size n greater than 30 from a population. In step 2 then you compute the sample mean $\bar{X}$ and sample standard deviation. In step 3, now you can assume that your sampling distribution is a normal distribution with mean equal to population mean μ which is unknown. And the standard deviation σ by root n which can be approximated by $\frac{s}{\sqrt{n}}$.

Step 4, you select the confidence level at which you want to estimate the population mean for example, 95 percent, 99 percent or maybe 90 percent it can be lower or higher than these numbers based on your requirement. Finally, compute the confidence interval given by $\bar{x} - z^* \frac{s}{\sqrt{n}}$ to $\bar{x} + z^* \frac{s}{\sqrt{n}}$. Now that you know how to determine the confidence interval for a population.

We will take a look at some other scenarios that you might come across and learn how you can calculate the confidence interval for these scenarios. **(Video Ends: 21:05)**

(Refer Slide Time: 21:07)



Statistical Inference IV: Interval Estimation for Small Samples

Statistical inference IV: Interval estimation for small samples. **(Video Starts: 21:10)** So, far you had the luxury of collecting lots of data points which meant your sample size was at least 30 which allowed you to apply the central limit theorem. Hence, you need a minimum sample size of 30 to be able to use the Z-distribution for calculating the confidence interval. However, in real life you will often come across situations where you have to work with small samples.

For example, assume you are working for a pharma company, you need to compute an interval for the effect of a medicine on patients and you only have 15 volunteers. In such cases it is better to use the T-distribution when the population standard deviation is not known. Instead of using the Z-distribution, where we estimate the population standard deviation to be equal to the sample standard deviation.

So, in this discussion, you will see how to calculate the confidence interval using a T-distribution when your sample size is less than 30? So, first let us understand what this T-distribution is? The T-distribution is centred at 0 but it is standard deviation is proportionally larger compared to the Z-distribution. The only difference is that the T-distribution has a shorter peak and wider tails.

Consider the figure shown here the exact shape of T-distribution depends on the size of sample for smaller sample sizes 30 distributions are flatter and for large sample sizes as you can see here how the sample size relates to the T-distribution. As sample size is increase degrees of freedom increase and as they approach 30, the T-distribution approaches to normal distribution which is in black.

So, let us take an example where you can understand how you can estimate the interval for smaller samples. Imagine you work from a pharma company and are testing the effects of medicine on 15 volunteers. Hence, your sample size is 15. Now you find that medicine increases the presence of a particular hormone XYZ in patient's blood by 10.038 micro units. So, this value becomes our sample mean.

Since the population standard deviation is not known to us, we estimate this using the sample standard deviation. Let us say that our sample standard deviation comes out to be 0.072 so, how do we solve this problem? Recall, the general approach you learned for estimating the interval using Z-distribution. We will use this exact same approach except for a few changes. So, let us go over each step that you have learnt.

The first step was to collect the sample size n, here n = 15, since the sample size is less than 30, we will go ahead with the T-distribution. Let us move to step two, now step two was to compute the sample mean and standard deviation. Since the population standard deviation is known to us, we are estimating this with the sample standard deviation. Again, the values of sample mean $\bar{X}$ and the sample standard deviation S is 10.038 and 0.072.

Coming to step three, earlier you learned that step three was where we assumed our sampling distribution was a normal distribution. However, since we are dealing with a smaller sample, hence in our case the sampling distribution will follow the T-distribution. Now, like the normal distribution there is an entire family of different T-distributions each T-distribution is distinguished by what statisticians call degrees of freedom which are related to the sample size of the data set.

For a sample size of n, the degrees of freedom for the corresponding to distribution would be n – 1. For example, for a sample size of 100 T-distribution would have 100 – 1 that is 99 degrees of freedom denoted DFT 99. This is why for smaller sample sizes the T-distributions are flatter than for large sample sizes. In fact, as the sample size increases the degrees of freedom also increase.

This makes the T-distribution look more like a standard normal distribution or the Z-distribution. The point where they become very similar to each other is about the point where the sample size is 30. Hence for any sample size that is greater than 30, the T-distribution can be approximated to a normal distribution. So, in our case the sample size is 15 this means that the degrees of freedom should be 15 – 1 which is 14.

Hence the distribution that we will use to calculate the confidence interval will be a T-distribution with degrees of freedom equal to 14. Now that we know our sampling distribution for step three. Let us move to step four if this step four was to select the confidence level at which you want to estimate the population mean, let us assume that the confidence level we want to estimate is 95 percent.

Let us move to step five, this is the step where we calculated the final confidence interval using the formula $\bar{x} - z^* \frac{s}{\sqrt{n}}$. The only difference in the formula will be that in place of Z we use t value. The confidence interval for our population mean will thus be given by $\bar{x} - t^* \frac{s}{\sqrt{n}}$ to $\bar{x} + t^* \frac{s}{\sqrt{n}}$.

So, what is the t critical value? To calculate this value, we will use either the T-table or you can also compute t-values on T-table easily using R software. So, we are going to search for 0.05 significance value and two-tail test. For now, we do not have to worry about what is one-tail and two-tail, we select the critical t-values from the table corresponding to the degrees of freedom.

For example, in this case degrees of freedom is 14. So, we will find the critical value as 2.145. The lower bound of the interval will be $\bar{x} - t^* \frac{s}{\sqrt{n}}$ and the upper bound of the interval will be $\bar{x} + t^* \frac{s}{\sqrt{n}}$. So, let us substitute the values we have and then we get 10.038 – 2.145 which is our t critical times S which is 0.072 and then divide by square root of 15 this comes out to be 9.998 i.e. $10.038 - 2.145^* \frac{0.072}{\sqrt{15}}$ = 9.998.

Similarly, the upper bound will be about $10.038 + 2.145^* \frac{0.072}{\sqrt{15}}$ and this value comes out to approximately 10.077. So, we can say that our confidence interval for the presence of that hormone lies the range of 9.998 and 10.077 with a 95 percent confidence interval. So, let me quickly recap what you have learnt in this video, you must use the T-distribution then the number of data points that you that is a sample of size of less than 30.

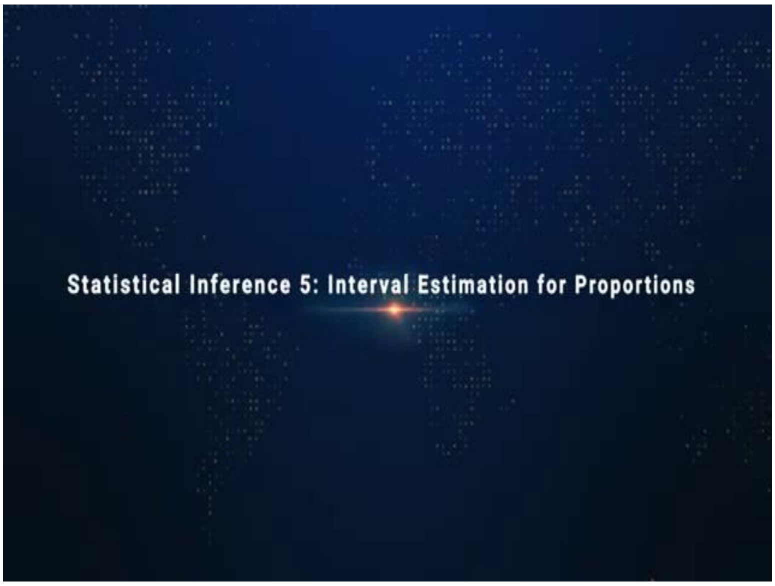
And the T-distribution is also preferred when the population standard deviation is unknown. We also learned that the T-distribution depends on additional parameter called degrees of freedom that is df. The degrees of freedom is calculated as a sample size –1, as the sample size increases it means that the degree of freedom is also increasing and as a result the T-distribution tend to become narrower and narrower.

At sample size greater than or equal to 30 the T-distribution is essentially indistinguishable from a normal distribution. Which is why if you tried solving the noodles problem using the T-distribution, you will get the same result as you got with the Z-distribution. So, going forward we can see that the flowchart that you can use for deciding between Z and T-distribution.

If the population standard deviation is unknown and the sample size is greater than or equal to 30 then that distribution is perfect over T-distribution. This is so because T-distribution approximates the variance using the sample size as you saw when we increase the degrees of freedom but this is not the case with Z-distribution. However, if the sample size is less than 30 then even if the population standard deviation is known it is best to use the t-test as it is idly suited to dealing with small sample.

Now, if we use the T-distribution then there is the formula for calculating the confidence interval. The lower bound of the interval is given by $\bar{x} - t^* \frac{s}{\sqrt{n}}$ and the upper bound is given by $\bar{x} + t^* \frac{s}{\sqrt{n}}$. Now, using Z-distribution and T-distribution you can estimate the confidence interval for any sample whether the sample size is greater than 30 or less than 30. **(Video Ends: 28:43)**

(Refer Slide Time: 28:44)



Statistical Inference 5: Interval Estimation for Proportions

Statistical inference part 5: Interval estimation for proportions. **(Video Starts: 28:47)** Here, we have assumed that all the values in our sample are numerical values but what if the values are categorical in nature. A very common example of this is the exit polls that are conducted during the election. The idea of these polls is to estimate the number of people who voted for a certain candidate or party.

Suppose you sample a few voters and calculate the proportion of voters that voted for a certain party. Now how will extrapolate this value to the entire population? In this discussion, we will just show how to solve such problems and as an example, we will take up this exact problem. In this discussion, we will show how to solve set problems when the data is categorical in nature and in such cases, we calculate the proportion instead of mean.

Let us see this example if you are working as a part of political science company that specializes in water pools and design surveys to keep political office seekers informed of their position in a race. This is done through telephone service where they ask registered voters who they would vote for if the election were held at that day. Now, suppose there is an election campaign going on and through these interviews you found that 220 registered voters out of 500 contacted favour a particular candidate.

Now, your company wants you to develop a 95 percent confidence interval estimate for the proportion of the population of registered voters that favoured that candidate. As you can see, the data in this problem is categorical in nature either the voter voted for that candidate or did not vote for that candidate. For this reason, we will work with the proportion of voters that voted for the candidate.

Hence the proportion of voters voted for the candidate is 220/500=0.44. Now, the objective here is to compute a confidence interval for the sample proportion value which is 0.44. The idea is to compute an interval around 0.44 for example 0.43 to 0.45. And let us call this sample proportion as $\bar{p}$= 0.44 as we mentioned earlier the approach for solving any interval estimation problem remains always the same.

Let us go back to our five step approach that we discussed earlier and understand how we can solve this problem? Step 1 here is to collect a sample of size and as you can do that in our case sample size is 500. Step 2 is to calculate the sample mean and standard deviation, however since our data is categorical in nature, we only have data on sample proportion. We found the sample proportion to be 220/500 which is 0.44.

Step 3, was to assume our sampling distribution as normal distribution. Now, as you have learnt for any interval estimation problem, we derive the interval from the sampling distribution. When we use the normal distribution, we first generated the sampling distribution of sample means which was $\bar{X}$. And we found that the sampling distribution was a normal curve with parameters μ and $\frac{\sigma}{\sqrt{n}}$.

In our example since we want to find the interval estimate for the population proportion if you recall the condition of approximating the sampling distribution of sampling means, it was that the sample size should be greater than 30. Similarly, we have a condition for being able to apply the sampling distribution of sampling proportion, the condition is that first the sample size which is n times the population proportion p should be greater than 5.

And second the sample size n*(1–p) should be greater than 5. So, let us see if these conditions are satisfied now since the best estimate of the population proportion that we have is the sample proportion, hence we will substitute the sample proportion in the equation. Since we know n = 500 and p is approximately 0.44 we already know that np = 220.

Similarly, if you calculate n*(1 – p) it will come out as 280, since both these values are significantly greater than 5. We can say that our sampling distribution follows a normal distribution and we can go ahead and use the formula for the confidence interval of the population proportion as we see shortly. So, let us move to our step 4 for now. Step 4 was to select our confidence level, it has been given to us that the company wants us to develop a 95 percent confidence level for the proportion.

Finally, we come to step 5 which is where we compute the final conference interval or calculating the confidence interval. We will use the formula given $\bar{p} - z * \sqrt{\frac{\bar{p}*(1-\bar{p})}{n}}$. So, this is the lower bound of our interval and the upper bound of the interval will be $\bar{p} + z * \sqrt{\frac{\bar{p}*(1-\bar{p})}{n}}$. Just by looking at the formula, you can say that the sampling distribution that we have used is a normal distribution with mean as the sample proportion. And the standard deviation as $\sqrt{\frac{\bar{p}*(1-\bar{p})}{n}}$. You already know that $\bar{p}$ = 0.44 and the sample size n is 500. We also have to calculate Z but you have already learned that for 95 percent confidence interval Z will be 1.96.

So, let us put these values so, if I just look at the margin of error Z or $z * \sqrt{\frac{\bar{p}*(1-\bar{p})}{n}}$. And putting in the values we will get $1.96 * \sqrt{\frac{0.44*(1-0.44)}{500}}$ = 0.3965. Overall, the value will give us 0.44 – 0.0435 = 0.3965 which is the lower limit and 0.44 + 0.0435 = 0.4835. So, our confidence interval is 0.3965 to 0.4835.

Thus, we can conclude that there are 95 percent chances that the proportion of all voters that favour the candidate is within range of 0.3965 to 0.4835. So, now let us quickly go over the approach for solving the interval estimation for the population proportion. Wherever the data is categorical in nature we calculate the proportion instead of mean. So, the aim of the problem is to be able to estimate an interval around the sample proportion p bar.

To start off these are n*$\bar{p}$ should be greater than 5 and n* 1 –$\bar{p}$ should be greater than 5. If both these conditions are met then we can use the formula for inference. The formula is given by quite simply $\bar{p} - z * \sqrt{\frac{\bar{p}*(1-\bar{p})}{n}}$. Here Z is basically your critical Z value that will depend upon the confidence level that is provided to us. **(Video Ends: 34:54)**