

Mine Automation and Data Analytics

Prof. Radhakanta Koner

Department of Mining Engineering

IIT (ISM) Dhanbad

Week - 9

Lecture 43: t-test

Music Welcome back to my course, Mine Automation and Data Analytics. Today, we will discuss a test popularly known as the t-test. So, in today's lesson, we will introduce the t-test and this test we use when the population has an unknown variance. We will first discuss the motivation behind this t-test and what are the assumptions required to proceed with this exercise. We will also discuss some examples. For example, one sample t-test and then two sample t-tests will give you a better idea of what the t-test is all about.

There are many applications for t-test distribution in the engineering domain. More particularly, when dealing with large amounts of data, different machines are performing in the Mine automation process to compare the performance of the miner, the machinery, and the spare parts; this distribution will be beneficial. So, in this particular lecture, we want to define the mathematical representation of the t-test. So, assuming that x_1, x_2 up to x_n are the variables and that they follow the normal distribution,

The screenshot shows a video lecture slide with the following content:

- Overview of t test: Mathematically**
- $X_1, X_2, X_3, \dots, X_n$ - iid Normal (μ, σ^2)
- σ^2 is unknown
- $E(X) = \mu$; $\text{Var}(X) = \sigma^2$
- Testing for mean,
- Null hypothesis (H_0): $\mu = \mu_0$,
- Alternative hypothesis (H_A): $\mu > \mu_0$

The slide also features a small inset video of a man speaking in the bottom right corner and a video player interface at the bottom with a timestamp of 3:05 / 40:55 and YouTube controls.

As we discussed in our previous lesson, t-distribution, we assume that distribution is nearly normal when using it. So, the representation of $n \mu \sigma^2$ is identical here. So, the σ^2 is unknown here; the expectation is μ , and the variance is σ^2 . Here,

we test the mean between the two populations or from the sample to a standard value. We want to compare whether this mean value differs significantly from the standard or significantly from the other sample group.

So, the null hypothesis is that the mean we found in the sample is the same as the standard value, which is μ_0 , that we want to prove. Alternatively, μ is more significant than μ_0 or less than μ_0 . These are the alternative hypotheses. So, based on this tenet, we have to proceed with the t-test. Here, we want to mention some of the mathematical assumptions we follow regarding where to apply the t-test.

So, this t-test is used to determine if there are significant differences between the mean of two groups or between the mean of the sample or a known value. For example, several drill operators are working for the drilling operations, so if we want to measure the performance between them, drill operator one and drill operator 2 have significant differences in their average performance. Suppose we're going to prove it or analyze it. So, for this circumstance, t-distribution is a handy tool for analyzing the difference in performance, whether their performance is nearly equal or if there is a significant difference.

So, we assess the observed difference between the performance of these drill operators 1 and 2. Then, we compare using the t statistics whether this difference is significant or not. So if it is not substantial, then we have to conclude that their performance is at par, or the performance level of the drill operator that every drill operator should reach this particular minimum target or average target. We assess drill operator 1 to see whether it is the same as target 1. Based on that, we will give some ranking or something like that or consider their salary and pay package. So this kind of distribution is beneficial.

So, this t-distribution will be when we proceed. So, this is a mathematical distribution similar to the normal distribution with heavier tails on both sides. Here, we calculate the t-statistic. In the z test, we calculate the z value and then try to find the z value from the standard regular table. And whether the value lies outside or inside that significant level.

Based on that, we accept or reject the null hypothesis. Here, we also calculate the t-statistics, which measures the difference between the mean of the two groups. The more significant the t-statistics value, the more likely the difference between the group mean is not due to a random chance. When we performed the t-test, we tried to give you an example to help you understand the circumstances in which you can use this t-test.

So here we elaborate that a company, for example, X, is a company that provides telephone service and is a service center. They have two service centers in the same city and want to compare which service center performs better. So, a few hundred or a few thousand people

may visit these service centers in two corners of the city. However, we only want to get some population data for the customers visiting these service centers. We want 50 samples from service stations or centers A and B.

Based on that, we assess which service center is taking too much time and which is taking less time. And whether the average time they take for each customer is substantially different or within the acceptable limit. This is an efficient problem, and we may find many similar applications. I have already given you an example for assessing the drill operator's efficiency. Similarly, the efficiency of the machines we can think about. Suppose several companies are providing machines.

So, two companies provide haul trucks that we use to handle the material of an autonomous haulage system. So we can also compare the performance of companies A and B. Is it significantly different, or is the performance level for both companies under the same conditions to the extent of acceptable limit? So, under these circumstances, we can use the t-test, which is an instrumental test. I hope you understand the application range or domain where we can apply the t-test to analyze the mean. The performance we measure is significantly different or not significantly different based on the t-statistics value.

$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

We have the t-statistics table similar to the jet table we showed you in the last lesson. So here, in this example, in service center A, it is found that the average time the service center A is taking is 22 minutes, and store B, another service station, takes 25 minutes. So, under these circumstances, we are taking only part of the data. There may be a few hundred people visiting these two service stations. However, we are only taking 50 random data out of it, and then we want to compare the mean value to see whether these differences are significant.

When Should We Perform a t-test?

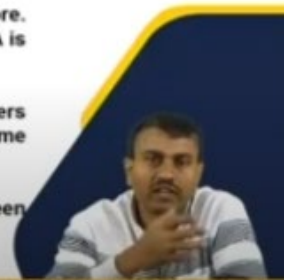
Let's first understand where a t-test can be used before we dive into its different types and their implementations. The best way to learn a concept is by visualizing it through an example. So, let's take a simple example to see where a t-test comes in handy.

Consider a telecom company that has two service centers in the city. The company wants to find out whether the average time required to service a customer is the same in both stores.

The company measures the average time taken by 50 random customers in each store. Store A takes 22 minutes, while Store B averages 25 minutes. Can we say that Store A is more efficient than Store B in terms of customer service?

It does seem that way, doesn't it? However, we have only looked at 50 random customers out of the many people who visit the stores. Simply looking at the average sample time might not be representative of all the customers who visit both stores.

This is where the t-test comes into play. It helps us understand if the difference between two sample means is actually real or simply due to chance.



If it is not very significant, then we can conclude that, yeah, the two service stations are performing at par okay, that is statistically acceptable, or if the difference is too much, then we have to think about why this performance difference is there some other reason or so. So what is the necessary task, or what actions must be taken? That is the subject matter of the telecom company. So, I hope you people can now understand where we can apply the t-test. Also, we have some basic assumptions to apply to this method or this statistical method under the circumstances. So, one assumption is that data should follow a continuous or ordinal scale, such as the IQ test score of the student we have already seen.

And data should be randomly selected. There should not be any bias, okay? There should not be any bias; we have to remember and follow. As we have already told, the data should be normally distributed, and the shape of the curve should be well separated. Extensive sample data or size data should be taken for data to approach a normal distribution.

And there are two kinds of things here that we want to test. Where the variance among the group is assumed to be equal under that case, we use the independent two-sample t-test. And where the variance between the groups is not equal, or we do not know that case, we go for the Welch t-test. We will discuss two types of t-tests in today's lesson. So first, the one-sample t-test.

This one sample t-test may be one good example when we have base-level data on what should be the performance of the drill operator. That is okay in terms of the average drilling time they are taking or the number of drill holes they do in a day or shift. That shift is better. So there is a standard mark, and based on that, the drill operator one is compared to that value; it is equal, similar, or there is a significant difference. So based on that, we can assess the performance.

So, one sample test is a typical example of that. So here, the mean value is our standard value, and the drill operator is the average value we found. We can estimate whether these are significantly different or not, and based on that, we can conclude. If there is no significant difference, then we can conclude that drill operator one performance level is at the expectation and, under these geo-mining conditions, okay. Otherwise, if it is not if the difference is too much, then we have to conclude that there might be some problem the drill operator is facing under these mining conditions.

We have to investigate and go through how we can improve the performance of drill operator one. That is the task of the mine manager or the mine management. These statistical tools help us assess performance and suggest what needs to be done. This is the beauty of these tools.

So, we have a hypothesis. Hypothesis H0 shows no difference between the sample mean and the population mean. An alternative theory states that there is a significant difference in H1. So, the null hypothesis H0 is represented as μ , which equals μ_0 . The sample mean equals the population mean, and the alternative theory states that μ is not equal to μ_0 . The sample mean is not equal to the population mean.

This is the t-statistics calculation method or formula. t is equal to \bar{x} minus μ_0 . μ_0 is the null hypothesis assumption that we are getting of the population mean. The \bar{x} is the sample, meaning that we have the data and randomly selected it. s is the standard deviation of the sample, and n is the number of samples we have collected.

One sample t-test

Test Statistic:

The test statistic for the one-sample t-test is calculated as:

$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Where:

- \bar{X} is the sample mean.
- μ_0 is the population mean under the null hypothesis.
- s is the sample standard deviation.
- n is the sample size.

17:19 / 40:55

CC Settings YouTube 10

And you can remember that when we started introducing this distribution in t distribution, there was also a concept of degree of freedom. So we will come back to that. So, the assumption is that we randomly selected the sample from the population. Yes, we have chosen.

And data follow a normal distribution. Okay, normality is followed. Okay. And it is a crucial step: independence. The observations in the sample are independent of each other and not biased.

Okay. So, it is not influenced by one another. Okay. So that we have to ensure. So, we have to stick to this assumption, and we have to maintain it. Based on that, we will proceed to the next step.

Now, the decision rule, when we compare the t statistics value, this value, t value with the value found in the table with a degree of freedom of n minus 1. Okay. For a sample size of n. And if there is not much difference, we must accept it. Alternatively, we can use the p-value associated with the t statistics.

So, the t value is better because of that table we have. So, based on that, we have to decide on the null hypothesis and whether to accept it or not. And how do we conclude? So if the calculated t value is greater than the critical t value found from the table, the calculated t value, the t value equation we have seen already, this. So if this value is greater than the critical t value from the table that we will find and know, we will show you the table first and reject the null hypothesis. This indicates that these two means are statistically significant, and this difference is not by random chance.

The sample mean difference from the population mean is not due to random chance. There is some other reason or something else. Otherwise, we fail to reject the null hypothesis if the calculated t value is less than the critical value. In that case, we have to suggest that there is insufficient evidence to conclude that there is a difference between the sample mean and the population mean. So, if we reject the null hypothesis, the observed sample mean is unlikely to have occurred by random chance alone.

There is evidence to support the alternative hypothesis, indicating a difference between the sample mean and population mean. So, this is the t-test critical value. Here, you can find 95% confidence, 98%, 99, 99 and so on. So there is one tail; this value is first for 0.5, and for two tails, it is 0.1. For one tail, it is 0.025, and we mostly use this and two tails, which are 0.05. And here you can see the degree of freedom is increasing from 1 to 40, then 80, 100 and 1000. So based on that, if suppose we found that 55 and it is for t is equal to or for the two tail test 0.05, so 55, so 40 to 60, so this value we have to recalculate assuming that the difference is uniform.

We have to calculate and put the value at 55. So, let us see one example. This example shows that we have a sample of 20 students whose average score is 72. We want to calculate from the population means of that particular student batch that the total batch and population mean are supposed to be 70.

The sample standard deviation is 8, for example. And we have to test whether the sample mean is significantly different from the population mean at a 5% significant level. So this is our problem. So, let us start with the mathematical statement. So, we define the null hypothesis as a measure of mu equal to 70, similar to the population's mean. The second part is the alternative hypothesis, which is that there is a significant difference in H1 mu, which is not equal to 70.

We have a significance level alpha that is equal to 0.05, and we know the X bar value 72; we know the standard deviation of sample 8, and we also know the sample size, which is the number of samples; it is 20, and n is equal to 20. So, the degree of freedom is 19. So we calculate the T, 72 was the sample mean, 72 minus 70 divided by eight divided by root over 20. So, this value is calculated as 1.58 and is now based on the degree of freedom, 20 minus 1, n minus 1, 19. So, if we return to this table, there is a 19. And it is for, I mean, a two-tailed test.

4. Calculate the Test Statistic:

$$t = \frac{72-70}{\frac{8}{\sqrt{20}}} \approx 1.58$$

5. Determine Degrees of Freedom:

1. $df = 20 - 1 = 19$

6. Find Critical Value or P-value:

1. At $\alpha/2=0.025$ and $df=19$, $t(\alpha/2, df) = 2.093$. (With the help of the **t distribution table shown earlier**)

7. Make a Decision:

1. Since $|1.58| < 2.093$ and the p-value is greater than 0.05, it fails to reject the null hypothesis.

8. Interpret the Results:

1. There is not enough evidence to suggest that the average score of the sample is significantly different from the population mean at the 5% significance level.

The diagram on the right shows a normal distribution curve with two shaded areas in the tails labeled 'Reject Region'. Above the curve, the null hypothesis is stated as $H_0: \mu = \mu_0$ and the alternative hypothesis as $H_1: \mu \neq \mu_0$. The title of the diagram is 'Two-Tailed Test'.

At the bottom of the video frame, there is a video feed of a presenter, a progress bar showing 23:31 / 40:55, and YouTube interface icons including CC, settings, and a play button.

So it is for df, which is equal to 19. So 19 is this value. For T of 0.05 two-tailed test, this is 19, two-tailed test 0.05, so 2.093. From that, we found that this 2.93 is greater than the value of 1.58. So 1.58 is less than the 2.09. So, it indicates that we fail to reject the null hypothesis. So, there needs to be more evidence to suggest that the average score of the sample is significantly different from the population mean at a 5% significant level. Now go to the next

stage. We have given you the example for one sample t-test. There are two sample t-tests when there are two, three, and multiple groups.

And we want to compare their performance. So, under those circumstances, we have to use this t-test case. So, there are two assumptions. Assumption number 1 is that the variance for the two populations is equal if it is yes. We have to follow these nomenclature. The t-calculation is from this value on this equation. And s-pool means the standard deviation of the pooled population we have to find from this formula. The degree of freedom for this case is n1 plus n2 minus 2, n1 minus one plus n2 minus one, and n1 plus n2 minus 2. There might be a complex scenario when the population variance is not equal. So, under this, we have to follow this nomenclature.

Two sample t-test cases - Overview

Assumption: Is the variance for two populations equal?

Yes

$$t_{cal} = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$df = n_1 + n_2 - 2$$

No

$$t_{cal} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

$$df = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{(s_1^2/n_1)^2}{(n_1 - 1)} + \frac{(s_2^2/n_2)^2}{(n_2 - 1)}}$$

It is complex, and this follows the test. We will show you both the example cases using the example. So, the two sample t-tests equal variance case. So, we will follow the first nomenclature here. Here, we must find the s-pool because the s-pool square is there.

So we have to root it over square root. Then, we have to calculate the t-calculation. For this t-calculation value, we have to compare and get the value under this confidence level of alpha and the significance level from the t table. And then we have to compare. Similarly, we have followed the one-sample t-test.

Here the df is n1 plus n2 minus 2. So here, the fundamental assumption is that there is no significant difference between the mean of the two groups. And their mean is found to be mean1 and mean2. And we are assuming in the null hypothesis that mean1 equals mu2. Alternatively, mu1 is not equal to mu2, mu1 is greater than mu2, or mu1 is less than mu2.

So for this, two-tailed, and this is for one-tailed. This is the definition of the hypothesis for these two sample t-tests. Now, we have to collect the data for the two independent samples, each with its own set of observations. And we have to verify this assumption that both samples are independent, yes. And both populations follow normal distribution, yes.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

In inhomogeneity of variance, the variance of the two populations is equal. That is one assumption we have already shown you. And based on that, we are proceeding. So here are the t-statistics to be calculated based on this formula. \bar{x}_1 minus \bar{x}_2 divided by s_{pooled} into root over one by n_1 plus one by n_2 . n_1 and n_2 is the sample size for group 1 and group 2. This is the mean of group 1; this is the mean of group 2. So, s_{pooled} is to be calculated using this formula. n_1 minus one into the standard deviation square of group 1. n_2 minus one into the standard deviation square of group 2.

$$s_{pooled} = \sqrt{\frac{(n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2}{n_1 + n_2 - 2}}$$

And n_1 and n_2 is the sample size. And df is calculated based on this n_1 plus n_2 minus 2. So, let us proceed to the example. So, here we have taken an example of two classes, class A and class B.

And this is the data, okay? A mean score is found. We are assuming the population variance for these two classes is the same. So, class A has a sample size of n_1 equal to 30. The mean \bar{x} -bar score is 75, and the standard deviation is 8. For the second, that is class B, the sample size is 25, the mean score is 72, and the standard deviation is 7. So, the null hypothesis states, we are stating the null hypothesis that there is no significant difference between the mean score of class A and B represented by $H_0 \mu_1 = \mu_2$.

An alternative hypothesis is that there is a significant difference between the mean score of class A and class B; that is, $H_a \mu_1 \neq \mu_2$. So, we will use and test under the significance level of α , equal to 0.05. So these are the calculation steps. We calculated the S_{pooled} for group 1 and group 2 from this equation. And it is found that 7.4. We now calculate the t statistic, and it is found that it is 1.51. After putting the S_{pooled} value and n_1 , n_2 value, it is found to be 1.51. From this table, do we have to find the sample degree of freedom? So, the degree of freedom is 30 plus 25, 53 minus 2, that is 53. So, this is the value from the 53 of this, 0.0212.

Example: Two sample t-test (1/2)

Calculate the Test Statistic:

$$s_{pooled} = \sqrt{\frac{(n_1-1) \times s_1^2 + (n_2-1) \times s_2^2}{n_1+n_2-2}}$$

$$s_{pooled} = \sqrt{\frac{(30-1) \times 8^2 + (25-1) \times 7^2}{30+25-2}}$$

$$s_{pooled} \approx 7.40$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$t = \frac{75 - 72}{7.40 \times \sqrt{\frac{1}{30} + \frac{1}{25}}}$$

$$t \approx \frac{3}{1.99} \approx 1.51$$

And based on that, we have to calculate the 53 value of the degree of freedom at 0.05. So, this is found to be alpha for alpha is equal to 0.05 and 53, 2.004. This 1.51 calculated in the last slide is less than the value we saw from the table.

So, we need to reject the null hypothesis. Example 2. Suppose this is a handy example. Many experiments are going on with teaching methods, whether teaching method A is better or teaching method B is better. So, many debates are going on. Similarly, in the mining scenario, we can think of the training of the operators.

Nowadays, virtual training and related training are also available. So, the field-level training and the virtual-related type of training are more contributory for the operator. So, we have to get more information about the system. So we can compare based on the data. Here, the t-test is a valuable tool for assessing this performance, and we must compare it.

So, we have taken one example. So, by teaching method A, we found that the student's average score is 85. And for that sample, s_1 is ten, and the sample size is 35. For the student taught using method B, okay, the sample size is 40, the mean score is 80, and the standard deviation is 8. Now, for the null hypothesis, we have to state that μ_1 is equal to μ_2 , and for the alternative hypothesis, we state that μ_1 is not equal to μ_2 . So, here we found the S-pool. We had to estimate first, and we had to calculate the t based on that. Our sample size was 35 plus 40, 75 minus 2, and 73.

So, for the 73 again, 73 is 60 and 80 between. So, 60 and 80 between 2 and 1.990 in between this value, okay. So, based on this value, we found from the table that it is 1.994. So, 1.994 is

less than what we saw in 2.17. So, it is, we have to reject the null hypothesis. So, we follow methods A and B significantly differ in the student's scores. Now, in the complex case, we have seen that population variance was known, and we have assumed that population variance was equal. The population variance for the case we have just discussed may not be the same.

So, under that condition, we follow the t-test. So, here, the other things remain the same: null hypothesis. The mean group average is equal or significant difference not equal for 2-tail and 1-tail μ_1 greater than μ_2 or μ_1 less than μ_2 , okay. We have to get the independent sample data from its observation, and we have to follow, obey the assumptions, and be satisfied. The samples are independent, and both populations follow normal distribution. The population does not need to have equal variance, okay?

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

So, under that, t is to be calculated using this formula. But the pdf has complex formulas, as you have already seen in the previous slides. So, df to be calculated based on this formula, s_1 whole square divided by n_1 plus s_2 whole square divided by n_2 whole square divided by s_1 whole square divided by n_1 whole square divided by $n_1 - 1$ and similarly s_2 whole square divided by n_2 whole square divided by $n_2 - 1$. So, we will get an approximate value of df here. Again, we will return to the t-table we have found and compare whether this value is too much or less okay.

$$df \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Based on that, we can say whether this difference is significant. Other assumptions and basic tenets are the same. We have to find the critical t-value on the t-distribution and the table and compare them. So, if the calculated t-value is greater than the critical value, we reject it, and if the computed t-value is less than the critical value, we fail to reject the null hypothesis. An interpretation is that if the null hypothesis is rejected, it indicates a significant difference between the means of the two groups. If the null hypothesis is not rejected, it suggests insufficient evidence to conclude a substantial difference between the mean and the two groups.

So, we will show you the procedure. This is because we have a wide range of applications for geotechnical purposes and for assessing machines' performance. These different kinds of machines operate in an automation scenario in a mine. When multiple parties are involved in the process of operations, assessing their performance may require this kind of test, which is an instrumental test. So, let us concentrate on this data. So, we have already seen the example of the teaching method.

So, in teaching method one, we assume that population variance is not equal. Sample n1 is 25, the mean is 78, the standard deviation is 7, and for method 2, n2 is 30, the sample mean is 82, and the sample standard deviation is 9. And test whether there is a significant difference in the average score between the two teaching methods at a 5% significance level. So, the same H0 assumption, mu1 is equal to mu2, H1 mu1 is not equal to mu2, and here alpha is 0.05, and these are the data for group 1 and group 2. So, we calculated the t value to be minus 2.06, and from the DF calculation, we found that df was nearly 51.

Example

4. Calculate the Test Statistic:

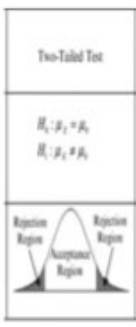
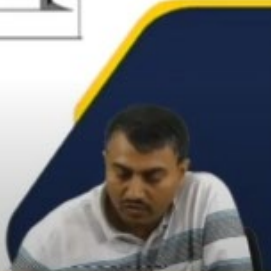
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

On Substituting values: t value : -2.06

5. Determine Degrees of Freedom:
 Degrees of freedom (df) are calculated using the formula:

$$\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

df = 51.69 (rounded down to the nearest whole number, df = 51)

39:23 / 40:55 CC HD YouTube 34

And from the 51 again, we have to return to this particular one between 40 and 60. So, it would be greater than 2.0 and less than 2.021. So, this value alpha, alpha by 2.025 for the degree of freedom 51, is 2.009. So, minus 2.09 is less than minus 2.009. So, we reject the null hypothesis. So, there is enough evidence to suggest that this difference is significant. So, these are the references. Let me conclude in a few sentences. So, we discussed the t-test for the population unknown variance and their motivations. We have clarified the required assumption and shown the three examples for one sample t-test, two for the two sample t-tests, and the other for the variance equal to the population. For the population, variance is not known or is not equal. Also, we have enumerated using one example. Thank you.