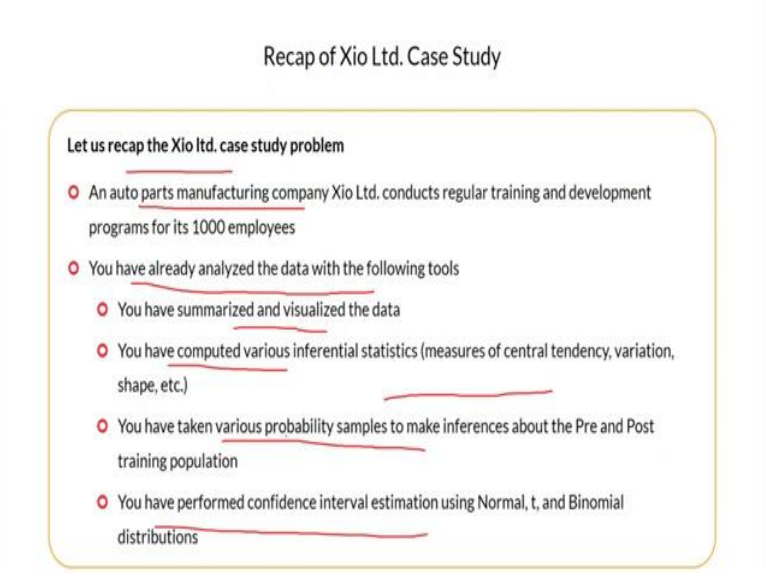
Artificial Intelligence (AI) for Investments
Prof. Abhinava Tripathi
Department of Industrial & Management Engineering
Indian Institute of Technology – Kanpur

Lecture – 23
Recap of Xio Limited Case Study

(Refer Slide Time: 00:00:13)



Dear participants, hello and a very warm welcome to all of you. Let us recap the Xio Limited case study. An auto parts manufacturing companies Xio Limited, conducts a regular training and development program, for it is 1000employees. It wants to check the current level of satisfaction of employees and that after the training program. You are the chief data scientist in the consultant firm engaged by Xio Limited.

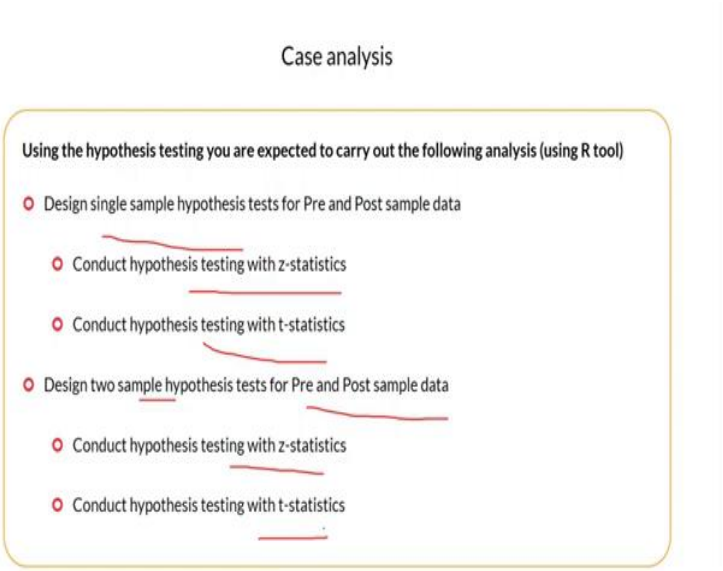
The firm has given you the primary surveyor data response feedback taken from these thousand employees, pre and post training program. And want you to examine the impact of training program on the employee satisfaction. You have already analyzed the data by summarizing and visualizing it. You have computed various influential statistics. For example, these included measures of central tendency.

For example, median, mode and mean, measures of variability. For example, range variance, standard deviation and mean absolute deviation. You extracted samples from the original data using various probability sampling techniques such as simple random sampling, systematic

sampling, stratified sampling and cluster sampling. You analyze these samples with all the analytical tools learned in the inferential statistics module.

You also performed interval estimation techniques using normal distribution and t distribution to generate confidence intervals. Lastly, you define detractors having satisfaction levels of less than 4.5 and assigned a score of 0. And promoters with satisfaction level of more than 4.5 and assigned a score of 1. You model the data using binomial distribution and made inferences about the population using samples from this detractor promoter data.

(Refer Slide Time: 00:02:11)



In this module on hypothesis testing, we will discuss the implementation of various aspects related to hypothesis testing. First, we will start by extracting a sample of 50 observations from pre and post variables. We will start with a simple case of single sample hypothesis test design. First, we will consider a case where population variance is known. In this case, we will conduct the hypothesis testing using z-statistics with the assumption of normal distribution.

We will make inferences about the population using our 50 observation sample, employing critical value analysis and p-value Analysis. We will design two-tailed tests to check whether the population mean is same as the desired satisfaction level of 4.5 or is it different. The null hypothesis here is that population mean is same as 4.5. Alternate hypothesis is that population mean is different from 4.5.

In the next step, we will design one-tailed test to exactly identify whether population mean is more or less than 4.5. In the next set of tests, we will repeat all the previous analysis, assuming

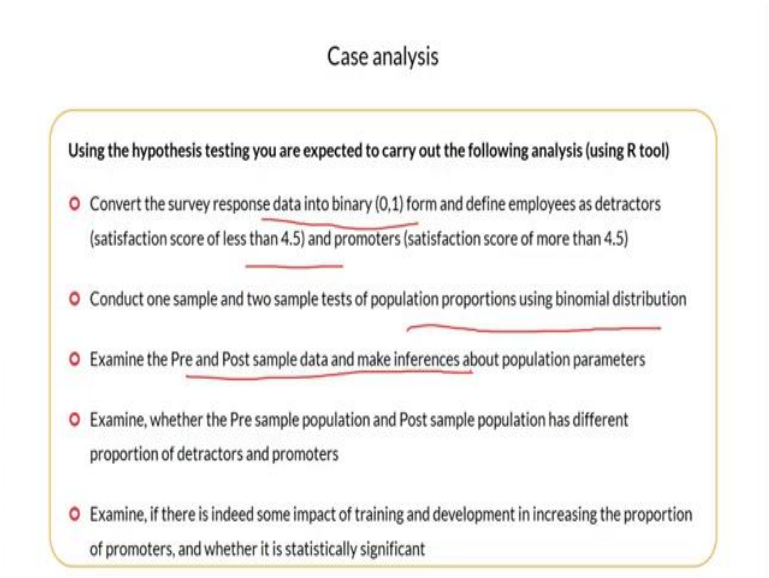
that population variance is not known. So, we will employ t-statistics and students t distribution to perform hypothesis testing. We will also examine the impact of sample size on hypothesis testing and analysis with z and t statistics.

And contrast the application of the two distributions, namely normal distribution and student's key distribution. We will repeat all this analysis for pre and post samples to make inferences about the impact of training and development program on employee satisfaction. Using the original data, comprising thousand observations will verify whether the inference is made using sample do apply on the actual population parameters.

After examining the pre and post samples separately, we will now examine them together through the application of two sample tests. We will directly compare the employee satisfaction levels between pre and post sample data and make inferences about the population parameters. We will conduct the hypothesis using z and t test statistics. We will examine whether the means of pre and post sample data are significantly different.

And can be used for making statistically significant inferences about the population using these results.

(Refer Slide Time: 00:04:39)



Finally, we will convert our survey response data into binary data that is in the form of 0 and 1 of promoters and detractors as we did in the first module. Those with a score of less than 4.5 are considered as detractors and assigned a value of 0. Those that are considered as promoters

are assigned a value of 1 who have a score of more than 4.5. We will model this data using binomial distribution.

And repeat all the one sample and two sample tests to make inferences about population parameters using the sample. That is we will try to examine whether the promoters and detractors have equal probability of appearing in pre and post samples. If not then we will examine whether there is a statistically significant impact of trading and development program on the proportion of the promoters and detractors on the pre and post data.

(Video Starts: 00:05:38) In this module, on hypothesis testing, we will start with downloading the data and generating random samples with 50 observations. In between some of us may have changed the working directory. However, if we followed our earlier best practice of setting the working directory correctly at the beginning itself, we can straight away run the setwd command.

And set our working directory wherever at our data and other materials stored. So, I can straight away on this setwd command and running this setwd command will set my working directory where all my data is stored. So, first we need to read the data. Data=read.csv('Data.csv'). Remember a data is stored at this location and I simply need to run this read.csv file without giving any location simply because our working directory is already appropriately set.

I can run this data.csv file and I can easily read it. Now, as a first step I need to generate sample with 50 observations. Pre=sample(Data$Pre, 50, replace=T)As we recollect, we have generated the samples with a very simple sample command, specifying the original population which is Data$Pre. And telling r that we need to generate a sample of 50 observations with replacement. So, I am putting the replacement argument as true which means that all the observations are sampled with replacement.

Similarly, Post=sample(Data$Pre, 50, replace=T),I can generate my post sample of 50 observations with the sample command. I can use Data $Pre 50 observations with replacement. So, observations are replaced back into the sample into population when the sampling is done. So, now that we have generated our pre and post samples of 50 length we will start with the hypothesis testing. And as a starting point, we will start our hypothesis testing implementation with single population.

As a first step in hypothesis testing, we state the null and alternate hypothesis. The framework of hypothesis testing requires us to specify two mutually exclusive hypothesis that is null hypothesis, often referred to as H0 and alternative hypothesis which is often referred to as H1. More specifically, we should choose H0 or null to be the case of no effect.

That is sample statistic is same as population statistic or there is no change or you can reject H0 and choose H1 to be the case when you want to show or prove something. That is the sample statistic is different from the population statistic. Now, the business case for this may be as follows. Xio Limited, as we discussed earlier is interested in individuals with the satisfaction of less than 4.5.

So, those individuals with the satisfaction score of less than 4.5 can be further targeted through training and development programs to improve their satisfaction level. This means that in this case, we want to test the hypothesis whether population mean is same as 4.5 or different from 4.5. That means your null would be that mean is equal to 4.5 and alternate hypothesis is, it is not equal to 4.5.

And therefore, it is a two tailed hypothesis. Please also remember, as a convention, we tend to say that we are rejecting the null rather than saying that we are accepting the alternate hypothesis. Here, we will also examine whether this mean is less than 4.5 or more than 4.5 by conducting single tail test. And we will start this hypothesis testing by installing the relevant package which is BSDA.install.packages("BSDA")

Once you install this package you also add this to your current working library with library(BSDA) command. This will add all the features and functionalities into your current working library and you can run this. Now, as a starting point we will run this z.test command to conduct the hypothesis, assuming that variance of the population is known. The syntax is quite easy. z.test(Pre, mu=4.5, alternative="two-sided", conf.level=0.95, sigma.x=sd(Data$Pre)) We provide the sample which is pre first.

Later on, we will also examine the post sample then we will provide the hypothesized mean which is 4.5. And then we also state the alternative hypothesis which is two-sided case. So, we are putting as two-sided. That is, it mean, should can be either less or more as well. We also

state the confidence level which is 0.95 we are putting. That means the synchrance level is 1– 0.95 which is 0.05.

Interestingly, we also need to specify the population standard deviation. For population standard deviation, we will use the standard deviation of original data pre-sample variable which is pre-variable as sd. Using sd command we can compute the standard deviation of original pre-variable running this z.test will provide us with the hypothesis testing results. So, for example, if I run this my z-statistic is computed as – 5.5838.

Now, if I am going ahead with the critical value method, I need to compute the critical value. So, for that I will run this qnorm command which will provide me with the appropriate statistic using standard normal distribution. qnorm(p=0.05/2, lower.tail=T). And for that I need to provide the appropriate p-value which is 0.05 by 2. Why? I am using 0.05 by 2 because it is a two-tailed test, so, 2.5 percentage significance level on both the test.

Also, please notice that this z value is negative. Even though, we are testing two-sided tests but it indicates that the mean is much lower than 4.5. It is on the lower side, considerably lower side. And therefore, it would be more appropriate to put this lower.tail argument as true. That means we are generating the critical value on the negative side that is corresponding to the lower tail of 2.5 percent.

And if we run this command, the lower tail statistic is computed, as – 1.96 which is much lower than our z-statistic. Which clearly indicates that we can reject the null that our population mean is 4.5. This result we could have directly obtained by observing this p value which is very small, very, very small, as compared to 2.5 percent. That means, even just by looking at this p-value which is very significant.

We can easily reject the null that our population mean is 4.5. Now that we have understood the basic hypothesis testing using normal distribution that is z test. We will further discuss in the next video about hypothesis testing with t test as well. In the previous video, we discussed how to conduct hypothesis testing with critical value method and p-value method. We conducted a two-tailed hypothesis to check if the population mean was equal to 4.5.

Indeed, from our hypothesis testing method, we found that the mean was different from 4.5. And there was some indication that it was less than 4.5. In fact, we could have easily verified our results by running this mean command on our original pre-variable which is Data $Pre to see that the population mean is 4.00 which indicates that our inference from the hypothesis testing that mean is not equal to 4.5 was indeed efficient=mean(Data$Pre)

Also notice that we chose a significance level of 5 percent that means a type 1 error, probability of 5 percent. That is 5 percent chance that we can reject the null incorrectly. Also notice that z value was negative. If you remember the z value that we computed was –5.58, indicating that pre-sample mean may be much less than 4.5 which is also confirmed by the mean value of Data$Pre original data pre-variable here.

However, still we need to test in a more formal manner, using single-tailed hypothesis testing, whether the mean is more or less than 4.5. So, using a sample we will test that as well. And for that we can design our hypothesis testing procedure like this. So, we can use our single-tailed hypothesis design. And for this we will again use the similar command which is z.test, specify the pre, specify the mean as 4.5. z.test(Pre, mu=4.5, alternative=”greater”, conf.level=0.95, sigma.x=sd(Data$Pre))

But now, notice that I am giving the alternative argument as greater. That means I want to check whether my mean is greater than 4.5. Although, we already know that it is less much less in fact than 4.5 but still will check that. We will put a confidence level as 95 percent. That means a significance level of 1–0.95 or 5 percent. Again, since it is a normal distribution based hypothesis testing that statistic is used.

We need to specify our standard deviation from the population. And we are good to run this command. Notice that it gives a p value of 1 which indicates obviously that this hypothesis testing was largely spurious or not so, valid because the value is much less than 4.5. So, even without conducting this hypothesis test, we could have inferred that the mean is less than 4.5 which is also confirmed by the result of this hypothesis testing.

That we cannot reject the null that our population mean is greater than 4.5. So, we will move to the second step where we will test whether the population mean is less than 4.5. And for that again we will use the sample pre mu value of 4.5. But in this case, we will run our alternative

as less which means we are testing, whether our population mean is less than 4.5. And we will use our confidence level again as 0.95. z.test(Pre, mu=4.5, alternative="less", conf.level=0.95, sigma.x=sd(Data$Pre))

And we will specify the population variance also as standard deviation of original pre-data. Now, if I run this, please notice again my z value is – 5.5838. If I want to test my hypothesis using critical value, I can simply compute my critical value, as we did earlier with qnorm command. But in this case, I would specify my p value or significance level as 5 percent not 5 by 2. Because it is a single tail hypothesis, so, I will put p = 0.05.qnprm(p=0.05, lower.tail=T)

Again, we already know that we are testing for whether mean is less than 4.5. We will use lower.tail as true sort will get a negative z-statistic. And the statistic here is –1.64 which is much less as compared to this z value which clearly indicates that we can reject the null. And say the true mean is indeed less than 4.5. So, we will reject the null or in other words we can also say that we will accept the alternative.

And now that we have rejected the null, we can say that the mean is less than 4.5 with 95 percent confidence or 5 percent significance level. Second point, we could have easily inferred the same by just looking at the p value, p value is so small it is almost close to 0 indicating that with a lot of significance, we can reject the null. And state that mean is much less than and 4.5 and the result is test statistically very, very significant.

Again, you can see that Data$Pre we will again show it is 4.007.mean(Data$Pre). So, with 95 percent confidence we can say that our hypothesis is correct and the mean is less than 9.5. We can reject the null again. There is a 5 percent chance that we may incorrectly reject the null when it is actually, true. However, we can conclude our hypothesis testing here with normal distribution, using z-statistic.

In subsequent videos, we will implement the hypothesis testing using students t distribution and t-statistic. In the previous video, we have understood how to conduct hypothesis testing using z-statistic? When the assumption of normal distribution was valid. And we also knew the population standard deviation. In this video, we will conduct the hypothesis testing with t-statistic that is, we will conduct the hypothesis testing with t-statistic using student t density distribution.

And we will assume that population variance is not known to us. Also, we will see that t-statistic is more effective when the sample sizes are small. And for t tests degrees of freedom is an important property or an important feature that needs to be computed to perform t test. So, to start with, first, we will start with two-tailed test and for that we will design our t test like this. We will run the t test command.

t.test(Pre, mu=4.5, alternative="two-sided", conf.level=0.95)

We will provide the argument, pre-sample variable pre, they specify the mean as 4.5. Because we are assuming or we are trying to test whether population mean is same as 4.5 or different than that, will provide the alternative which is two-sided. That means we are conducting a two-tailed test, will provide the confidence level as 0.95. That means significance level of 5 percent.

Also, please notice we have not provided any argument regarding the standard deviation of population. Because assumption here is that standard deviation of population is not known to us. Now, we will try to run this command and once we run this command notice, the t-statistic which is – 5.8733. This is the t-statistic for this test and if we are doing the critical value analysis, we need to generate the critical value of t-statistics.

For that we would require the degrees of freedom which can be easily computed as length of sample size –1.DOF=length(Pre)-1. So, this is my degrees of freedom. Once I have my degrees of freedom, I can generate the t-statistic, the critical value corresponding to 5 percent level of significance using this qt command. I will write qt and probabily t value which is significance level is given as 0.05 divided by 2.qt(p=0.05/2, df=DOF, lower.tail=T).

When using 2.5 percent or 0.05 divided by 2 is simply because it is a two-tailed test. So, we are conducting the hypothesis test that 4.5 is different from the population mean. That means the population mean, can be either more as well as less. So, it is a two-tailed test and that is why we are using 0.05 divided by 2 as our significance level. We also provide the degrees of freedom to our argument.

And then we specify lower.tail as true, why we are doing this? As we discussed earlier, the t value appears to be negative. And therefore, it seems it is more appropriate if we examine the t values on the lower side that is, we examine the critical values on the lower side. So, we

generate this critical value which is –2.00 as compared to our t value from this test which is –5.8733. It appears to be quite small.

And therefore, we can safely reject the hypothesis that population mean is 4.5. Now, a more simpler way would have been to simply look at the p value of this test which is very, very small, close to 0. And therefore, with a lot of significance p can easily reject the hypothesis directly looking at this p value. And say that our population mean is not same as 4.5. So, with this we have implemented the t test in this video.

And in the next video we will try to see whether we can find, if the mean is more than 4.5 or less than 4.5 by conducting single-tail test. Will also examine what are the critical differences, while we are conducting t-statistic and z statistic, related hypothesis testing. In the previous video, we conducted the two-tailed hypothesis testing using t test statistics. Now, let us compare our results across z test statistic and t test statistic.

And for that we will start with the sample size of 100. Let us start with the sample size of 100. We extract that from our original data of thousand observations. Now that we have extracted our sample, we will start with the hypothesis testing using z test statistics. Let us look at the result. A very low p value which is almost equal to 0 is obtained. Let us conduct the same test using p test statistic two-tailed test.

Again, a very low p value which is almost close to 0 so, not much difference. Although, still it is higher but still we can consider them to be almost similar level of significance. So now, we will decrease our number of observations. As of now it appears that both of these tests are giving almost similar level of significance which is very close to 0. But now let us take 10 observations, a sample comprising 10 observations.

We will extract that sample from our original data again. Let us first consider the two-tailed test with z test statistic, it is 0.006. Now, again we will perform the same test with the test statistic. And the value is much larger notice the p value it is much larger. It is 0.0195 as compared to a p value of 0.006 with z-statistic. What do we infer from here? Please notice when the sample sizes are low, the p value is higher with t-statistic and it indicates a lower significance.

That means t test statistic, in stills a more sense of reality when you have low number of observations. And also, you do not know the population variance then it is more proper and more appropriate to have more evidence in favour of rejecting the null. That means, if you want to reject the null, you would probably want to have more evidence when your sample sizes are much lower. Add population variance is also unknown.

And that is what is facilitated by t-statistic hypothesis testing. Given the nature of students, t distribution, it is much flatter around the tails and the peak is lower at closer to mean. It indicates that this is more closer to reality. That means when observations are low, number of observations are smaller. And also, we do not know the variance in standard deviation of our population.

Then it is better to have slightly more evidence relatively higher evidence to reject the null. And that is what is precisely achieved with the t test statistics. In case, when we have large number of observations then as we have noticed already. The inferences and the significance of tests is quite similar between whether you use t-statistics or z-statistics hypothesis testing.
So now, we will move to single-tailed test and we will try to establish whether the population is has a mean of more than 4.5 or less than 4.5.

So, in this we will conduct the t test to see whether population mean is more than 4.5 or less. So, first, let us see if it is more, although we already have some intuition looking at the previous z values and t values that it is lower than 4.5 but still, we will design our test. So, we will give the sample pre mu value of 4.5. Alternative argument as greater to see, if the mean is greater than 4.5 to establish the confidence level as 0.95 that means a significance level of 0.05. t.test(Pre, mu=4.5, alternative=”greater”, conf.level=0.95)

Let us run this notice, a p value of 0.99 which again indicates that our test is giving slightly spurious results why, so? Because already we know that our population mean or the sample mean also is much lower than 4.5. So that means this test is rather spurious and will focus on the lower tail test. That means negative side of the statistic or tail. So, we will conduct the t test use our pre sample as we did earlier. t.test(Pre, mu=4.5, alternative=”less”, conf.level=0.95)

We will put mu = 4.5 alternative as less and confidence level as again 0.95. Please notice here we are not giving the standard deviation or variance argument why? Because assumption is that variance of the population is not known to us. So, from here we can compute the relevant t-statistic which is coming as –2.8363. Now, if you want to conduct the critical value analysis, you can simply generate the critical t value here using the qt function, you can use qt and specify the p value of 0.05 that is 5 percent. qt(p=0.05, df=DOF, lower.tail=T)Please note we are not using 2.5 percent. Why? Because it is a single-tailed test. So, we are looking at only one tail, the lower tail.

And therefore, we are using the argument p is 5 percent. We also need to supply degrees of freedom but before that we will put the sample size as 50. So, we will put the sample size as 50 and compute our degrees of freedom as length of sample size –1. And then we will run this will keep the degrees of freedom as DOF. And also, we will put the argument lower tail as T because we are looking at the lower side of the tail, the negative side of the t-statistics.

So, we can easily compute our relevant t-statistics as –1.67. Notice that the magnitude of this critical value is much lower as compared to the test statistic, t value that we computed earlier which is –2.8363. That means we can reject the null with a reasonable amount of confidence. And also, even without doing this critical value analysis, we could have directly looked at p value as well which suggests that the significance level of this test is very high.

So, with a lot of confidence, we can reject the null and say that mean is less than 4.5. That is pre sample mean. Before the training and development mean observations, the sample mean and the inference that we are making about the population. That population mean is less than 4.5. That is what we can infer from here. Now, again, we would like to again check whether changing the sample affects the level of significance or not.

So, let us do this testing with 10 observations only. So, let us do this testing with 10 observations. Pre=sample(Data$Pre, 10, replace=T)
Post=sample(Data$Pre, 10, replace=T)
And this time we will only run the single tail test. That is on the lower side which is this so, we will run this lower tail test. First, z test notice, the p value 0.001. Now, let us run the same test with our lower tail for the t test statistic. And notice, the p value is much higher indicating a lower level of significance.

So, this again confirms the argument that we already made. That in case of t test statistic when you have lower number of observations. And also you have made the assumption that you do not know the population variance, the level of significance of the test is lower. That means you need a larger amount of evidence to reject the null. And be sure about your influence. So, with this we have understood the t test statistics and z test statistics, the critical difference between them.

And we have already seen, how to infer the population mean statistic using the sample data from t test statistics and z test statistics using hypothesis testing? In the previous video, we have examined the pre sample data that is before training a development program. In this video, we will examine the post sample data. We will use similar z and t test statistics to conduct hypothesis tests.

And we will examine the impact of training and development. We will use p values to make inferences and first we will start with the z test statistics, hypothesis testing and then we will move on to t test statistics. So, first we will start with the post sample data with z-statistic. So, similar argument we will use z.test will supply our post data which is the post sample data extracted from our previous 1000 observations.

So, we will just extract the data 50, 50 observations we will extract. z.test(Pre, mu=4.5, alternative="two-sided", conf.level=0.95, sigma.x=sd(Data$Post)) So, we will extract the 50 observations. And then we will put a mean of 4.5 alternative argument we can set as two-sided. So, first we will start with two-sided. That means whether the poor sample data has a mean of 4.5 or it is different than that. We will again set our confidence level as 0.95. It means significance level of 0.05.

Please notice, since we are conducting z test statistic hypothesis, we need to supply the variance of population with this sigma.x which is same as standard deviation of original population which is data dollar post. Here, we are assuming that original thousand observations that we had is like a population data from which we have extracted the sample. So, this is our z test statistic. Now, let us compute the z-statistic, it is 5.0742.

In case, we want to do a critical value analysis. We can simply provide the argument, qnorm, and supply the p-value which is 0.05 divided by 2, because it is a two-tailed test. So, we need to supply 2.5 percent as discussed earlier. Then we will supply the lower.tail as F. [qnorm(p=0.5/2, lower,tail= F)]. Please notice that the z value is positive here. In fact, it is particularly large. And therefore, it would be better if we look at the upper tail, which is the positive side of the tail.

The critical value corresponds to the upper tail. So, we will generate that the value is 1.96. Now, let us compare this 1.96 critical value with the test z value, which is 5.07, which is considerably large and indicates that we can reject the null with a lot of confidence. And that means that the population mean is different from 4.5. It is not equal to 4.5. And without going to the critical value, even we can infer that from the p-value as well.

Where p-value is very small, indicating a very high amount of significance and a lot of confidence at which we can reject the null. Now, let us conduct a single-side test which is the z-test on the positive side. That mean is more than 4.5, and for that, again, we will supply the sample post. We will put the mu value equal to 4.5,and the alternative this time is greater. That means we are checking whether the mean is more than 4.5 population mean is more than 4.5 or not.

Again, we will supply the confidence level as 0.95. And we will supply the sigma.x argument which is the standard deviation of our population which is data dollar post. Now, when I run this command notice, the z value is again 5.0742, and it clearly rejects the null. That is the true population mean is greater than 4.5 a very strong result that suggests that the mean is greater than 4.5. Now, we can very well check that as well.

We are not doing the lower side of the test because it is quite obvious here, given the z-statistic and our previous hypothesis test that it is obviously more than 4.5. So, we are not doing the lower tail test. Let us look at the mean of the population, so that is Data$Post. And it appears that, indeed, the mean of the population, in this case, the population would be the original thousand observations, is much larger than the 4.5 value.

This is also an early indication that, indeed there is some impact of training and development on the population, what have we inferred from the sample statistic. So, with this, we have

understood the hypothesis testing on post-data using z-test statistics. And now, we will conduct the t test statistic hypothesis testing on our data. Now that we have examined the pre and post-sample separately, using simple single-sample hypothesis testing.

In this video, we will analyze and compare the pre and post-samples together using two sample tests. So, we will analyze pre and post-samples together using two sample tests. First, let us start with simple z test statistics. So, we will start with z test hypothesis testing. As we have already seen, we will use the z.test command. And we supply pre and post-arguments alternative as two-sided.

We will set the confidence level at 95 percent. Now, in this case, we need to supply the population standard deviation for both the pre and post-samples. Because we are conducting z test assuming that population variances for both the samples, are known, we will supply them sigma.x. So, first we will supply the population variance of the pre-sample. Then we will supply sigma.y, and we will supply the population variance of the post sample.

So, now that we have supplied the relevant arguments, let us try to see what we can infer about these z-statistics. We will focus on the p-value and z-statistic here. Notice this p-value clearly indicates that there is a significant difference pre and post. Notice the mean of pre is 3.99, and the mean of the post is 4.90. And their difference in the confidence interval is –1.14 to –0.69 0, does not lie in this.

So, there is no 0 in this interval, and given the p-value, the significance level of this p-value, it clearly indicates that the difference in their mean is not 0. And obviously, the post-value of the mean is higher as compared to the pre-value. So, this is an early indication that clearly, we can reject the null that their means are equal. And early results suggest that the post is higher than pre.

Now, we will do some more analysis on this. And first we will start with the one-tailed test. So, we will apply our z test for pre and post. We will supply the alternative argument as greater. So, it would be a one tailed test again the confidence level is 0.95. And we will set our standard deviations argument two as we did earlier. We can supply the population standard deviations. z.test(Pre, Post, alternative= “greater”, conf.level=0.95, sigma.x=sd(Data$Pre), sigma.y= sd(Data$Post)). Let us run this test.

Notice the difference is negative, that is $pre - post$ is negative. And therefore, the p-value is 1, which means this test is spurious. That means clearly always the post values are more than p. And therefore, this test becomes spurious. So, we will focus on the other side of the tail, the positive side. So, let us conduct that test using the z-test statistic. This test was simply spurious because the pre-values were clearly lower than the post.

So, the hypothesis is that post p is greater than post was clearly not rejected. So, we will conduct this pre, and post-alternative; we will supply the argument less, which means pre is less than post. And now, we will design our hypothesis with a confidence level of 0.95. And we will supply the population variances. If I run this, we are testing one-tailed test and checking whether pre is less than post.

So, the null hypothesis is, it is not less than post. And the alternate hypothesis is that p is less than post and clearly, it is rejected. It appears that the hypothesis that this pre is not less than the post is rejected with a lot of confidence. The p-value is much lower, and a very high level of significance is associated with this. That means we can very easily say that post is greater than pre. We can also confirm this by running the population difference.

So, we can directly check the differences between the population from pre and post. And it should not surprise us to find that, indeed, post is more than pre. That means the difference is negative. So, a pre is lesser than a post, by a difference of –0.98. So, which also confirms that our inference drawn from the sample is efficient. Now that we have conducted the z-test of differences between population means of pre and post-sample data.

We will move toward t-statistic test. That means we design our hypothesis using t-statistics. In this video, we will conduct the hypothesis testing using t test statistics. We will conduct two sample t test hypothesis testing, and compare the means of pre and post sample and make inferences about the data. First, we will conduct the two-tailed test and then we will design single-tailed test to check if there is indeed an increase in employee satisfaction due to training and development programs. So, we will conduct that t test and t tests are done when population variances are not known and sample sizes are small. So, we will start our simple t.test command. First, we will supply the pre sample then post then we will set the alternative

argument as two-sided. So, here we are simply comparing the pre and post means we will set the confidence level as 0.95.

Please note we are not providing here the standard, deviation, or variance of the population, as we did earlier with the z-statistic because the assumption here is that those variances are unknown. So, we will run this test. And notice the t value of – 8.7441. And also, please notice the p value, which clearly indicates that with a lot of confidence, we can reject the null. And say that the mean of pre and post-sample are not the same.

So, we can make that inference about the population. Also, there is some evidence to suggest that the mean of pre-sample which is around 3.99, and mean of post sample, which is around 4.90 are quite different. However, we will have more to say about that in a subsequent test. So, let us start with the single-tailed test. In the single-tailed test, let us first compare the greater alternative argument is greater that means we are saying pre is greater than post.

Although, if we look at our t-statistic that we computed earlier, it would suggest that pre is indeed lower than post. And therefore, this hypothesis is slightly spurious but it still will try and test that we will provide the confidence level of 95 percent and let us run. Look at the p value. It is equal to 1 which again substantiate that this hypothesis was sort of spurious. Because we already knew that pre is much lower than post.

But to check the statistical significance of that and prepare our hypothesis design around it will rather use the upper tail. That means we will use the lesser argument. We will put our alternative as less that means post is greater than pre. So, post minus pre is on the positive side, and pre minus post is on the negative side. So, we will set the confidence level as 95 percent and let us run this.

Notice the t value is negative and very significant, which again conforms to our original hypothesis. That pre is much lower than post, or we can say that significantly lower than post. So, this result is statistically significant, which is also confirmed by a very low p-value which indicates a very high level of significance. That means we can reject the null with a lot of confidence. And we can make the inference that pre is significantly lower than post.

Overall from these results, we can easily infer that there is a substantial impact of training and development activity. That means employee satisfaction before training and development, which we sampled in our pre-data. And then, post-training and development, there is a considerable increase which is statistically significant as well. That is what we infer about the population. Now, let us check what is happening in the population itself?

So, we will take the differences between post and pre and let us see if it is indeed large. We are directly working with population now. So, we can see that the difference is indeed quite large. It is 0.98 which suggests that the inferences that we made about the population are indeed valid and they hold with the population. Let us recollect the business case with Xio Limited.

The firm considered those with a survey response of less than 4.5 as detectors and assigned a value of 0. Those with a score of more than 4.5 were considered promoters and were assigned a value of 1. The firm wants to focus on those detectors with the 0 scores, and it wants to design customized programs suitable to these detectors. Now, in order to perform this analysis, we will model the data and design hypothesis tests using the binomial distribution.

So, first, we will start by examining pre-sample data with a single sample test. First, we will create our Pre_Bino variable, as we did earlier. We will run this if else code; we will assign a value of 0 to those with a score of less than 4.5 and others 1. Now, when we can create a proportional test, prop.test function, you can use that. We will use the summation of Pre_Bino which means all those with a score of 1.

And we will provide the sample size, which is 50, which will provide the confidence level, which is 0.95. That means 5 percent of significance. And will provide the alternative argument as two-sided. If I run this function essentially, I am testing a hypothesis that whether it is a score of 0 or an individual score of 1, that means there are two binary scores, 0 or 1. This equals the probability of any individual having this course.

First, we can perform critical value analysis using the critical chi-square value. It is quite easy. We can provide a confidence level that is 95 percent and a score of 1. qchisq(0.95,1). Since we are looking at the upper side of the tail higher side of the tail, we are providing the confidence interval. If you are looking at the lower side of the tail, you would have provided a significance level that is 0.05.

So, if I run this, I find a critical value of 3.84 which is much lower than 19.22 or in another way we can also look at the p-value and infer the same. The p-value is quite low, in fact very close to 0. That means with a very high significance level we can reject the null and say that any individual has a score of 0 and 1 if we pick any individual randomly having a score of 0 and 1, does not seem to have a 50 percent probability.

So that means there is a bias; the probability that individual has a score of 0 or 1 or he is a detector or promoter is not the same. Now, let us examine this in detail. We need to check whether there is a higher probability for promoter or detector. So, to conduct that kind of test, we need to design one-tail test. And therefore, first, we will put the argument as greater. Prop.test(sum(Pre_Bino),50, conf.level=0.95, alternative= "greater"))That means we are checking whether those with a score of 1 have a higher probability than those with a score of 0.

And let us run this command and notice the p-value is 1, which indicates that this hypothesis was indeed spurious; why? Because in our previous test itself, we could see the sample estimate of 0.18. And sample interval is 0.09 to 0.319, which is lower than 50 percent, which already indicated to us itself that those with a score of 1 have a lower probability than 50 percent.

So, in this hypothesis, where we are hypothesizing that null is that their probability is 50 percent. And the alternative is that probability is greater than 50 percent is spurious. So, we will modify our hypothesis, and we will look at the other tail. That means other arguments remain the same. That is, we will supply the Pre_Bino sum of the Pre_Bino variable 50 sample size, confidence level of 95 percent.

But the alternative argument would be less. Prop.test(sum(Pre_Bino),50, conf.level=0.95, alternative= "less")). That means the probability that an individual score of 1 is less than 50 percent. That is my alternate. My null is that the probability of an individual having a score of 1 is 50 percent, and the alternative it is less than 50 percent. So, this is my alternative hypothesis. If I run this command, it is indeed the case. That means my chi-square value clearly indicates and my p-value indicates that with a lot of significance, we can say that getting a score of 1.

That means, if I pick any individual randomly, the chances that his score is 1 is less than 50 percent. Now, I can also apply the qchi square test here. How? So, I can simply put chi square; since it is a single tail test, I will put the argument as 0.90,1 qchisq(0.90,1). And I will have it is 2.70, which is again much less than 19.22. So, with a lot of confidence, I can say that it is less than 1.

Again, if it was a two-tailed test, I would have provided 95. Since it is a single-tail test, I am considering 0.90. Now, we will examine the same for our Post_Bino variable. So, till now, we have examined the Pre_Bino variable. And we have reached the conclusion that in the population, the probability that if an individual has a score of 1 or he's a promoter, it is much less than 50 percent. We can also check this as well.

How do we check this? So, we can use the population data, which is Data$Post, For the population. We can create a new variable, Pre_Bino population. We can assign a score of 0 if it is less than 4.5. We can assign a score of 1 if it is more than 4.5. And we have this population with a score of 0 and 1. Now, what can we do? We can divide this Pre_Bino population with the length of the sample, which is 1000, and let us look at that.

So, we need to add the sum here. Once we assign the sum and divide it by 1000 sum(Pre_Bino_Population)/1000, it is 25.5 percent. So, in the population itself, the probability to observe an individual with a score of 1 is 25.5 percent which is much less than 1. And this is what precisely we estimated from our sample list hypothesis testing. With this, we found that in our pre-sample data, the probability of observing an individual with a score of 1, that is, promoter ,is much less than 50 percent.

So now, we will examine our post-sample data and make inferences using the binomial distribution. In this video, we will analyze the post-sample data using the binomial distribution. We will conduct the hypothesis test for proportions and make inferences similar to that we did for pre-sample data. We will also examine the probability of an individual being detected. That is a score of 0 or being promoter that is a score of 1 if it is the same or different.

And also examines if there is any impact of training and development on an employee being a detector or promoter. So, we will start with post-sample data. We will model it using the binomial distribution. So, first, we need to create our Post_Bino variable, as we did for the pre-

sample Pre_Bino variable. So, we will create this Post_Bino variable using the same, if else argument. And we will assign a score of 0 if it is less than 4.5 and 1 otherwise.

Now, will conduct a prop.test, and we will supply the argument sum of Post_Bino, which is those individuals with a score of 1 summation will be the sum of all those 1s and sample size 50. We will provide the confidence level as 95 percent it remains the same as earlier. And will provide our alternative hypothesis, which is two-sided. That means the probability if it is 0 or 1 is equal to 50 percent, it is null, and whether it is more or less, that will be our alternate hypothesis. So, if I run this argument, I get a chi-square value of 16.82 and a sample statistic of 0.8. Sample estimate suggests that the probability of having a score of 1 may be higher than 0. Will again look at the critical value; we can examine the critical chi-square value by supplying 95 percent confidence and degrees of freedom as well.

We can run this and we can see the chi square is again 3.84 which is much **less as** compared to the test statistic chi square which is 16.82. We can also infer the same from our p value that is 4.11e to the power –5 which suggests that with a lot of significance, we can reject the null. That means finding the probability that an individual scale score of 0 or 1 is not 50 percent it is significantly different.

And some indication of some intuition is also provided that it is more than 0.5, which we can got from the interval. The confidence interval starts from 0.658 on the lower side and 0.894 at the higher side, and the sample estimate is 0.8. So, there is some evidence suggesting that the population mean value may be higher than 0.5. So, with this, we move towards a more precise test, a one-tailed test to establish whether this probability is more than 50 percent or less than 50 percent of finding an individual with a 0 or 1 score. So, we will again conduct our prop.test, and some Post_Bino variable summation 50 is the sample size, and the confidence level is 95 percent. Let us give the alternate hypothesis as greater , Prop.test(sum(Pre_Bino),50, conf.level=0.95, alternative= “greater”)). Since we already have that intuition that the probability of finding the individual as 1 may be higher.

So, we will examine that hypothesis, and it is indeed the case it seems that the chi-square value is 16.82 and the p-value is much lower. That means a very high level of significance. So, when we are rejecting the null that observing an individual with a score of 0 or 1 has equal probability. We are rejecting it with a lot of confidence and a lot of significance.

That means the probability of identifying an individual with the value of 0, 1 in the population is indeed not the same. And the probability of having one is much greater, with the sample estimate of 0.8. So, we can also conduct the opposite hypothesis, although it will be spurious because we have already seen that the probability of having an individual with a score of 1 is much higher.

But still will conduct the opposite hypothesis. The same arguments will use a sample size as 50 confidence interval of 95 percent. And the alternative is less, which means that we are saying that identifying an individual with a score of 1 has less probability. That means less than a 50 percent probability than identifying and finding an individual with a score of 0. Prop.test(sum(Post_Bino),50, conf.level=0.95, alternative= "less")). So, if I run this hypothesis, please notice the p-value is 1, which clearly indicates that this was a spurious hypothesis.

We already knew that somebody with a score of 1 would have a higher probability than somebody with a score of 0. This we can also establish by directly looking at the population. How do we do that? So, first, we will construct the Post_Bino variable for our population. So, we assign a score of 0 to those observations that are less than 4.5 and otherwise 1. Post_Bino_Population=ifelse(Data$Post <4.5,0,1)
So now, I can see in the population itself. What is the proportion of those individuals having a score of one?
Sum(Post_Bino_Population)/1000. Let us see that, and if I run this, it is 75 percent, which clearly indicates that individuals having a score of 1 are almost 75 percent of the entire population. This also is a very clear signal that there is indeed some impact of training and development. And therefore, in the pre-sample analysis of proportions using the binomial distribution, we found that individuals with a score of 1 had a much lower probability, much less than 50 percent pre-sample.
And after training and development, that probability has increased considerably. So, there is a significantly higher probability than more than 50 percent. That an individual will have a score of 1, means employee satisfaction has improved after training and development programs. So, this is what we can infer from these single sample tests taking pre and post-variables separately.

In the next set of videos, we will examine this using two-sample analyses, two sample hypothesis testing by using pre and post-samples together. In this video, we will conduct two sample tests of population proportion using the binomial distribution. We will model the data and we will compare the proportion of detractors and promoters in the pre and post-sample. And then, we will make inferences about the population proportion, pre and post-training and development programs.

And we will try to find out whether promoters have increased after the training and development program. So, we will conduct two sample tests of population proportion using normal distribution, and to start with, we will again use that prop.test command. We will now supply not one sample but two samples very easily. So, first, we will provide Post_Bino, and then we will provide Pre_Bino. Next is the sample size, which is 50 for both samples.

Then we have a confidence interval as 95 percent. And will provide an alternative hypothesis as two sided. So, we are providing alternate hypothesis as two sided. prop.test(x=c(sum(Pre_Bino), sum(Post_Bino), n=c(50,50), conf.level=0.95, alternative= "two-sided")). So, essentially we are testing here that the proportion of detractors and promoters is the same across both populations using the data from a sample. So, if I run this command, please notice that the p value is very, very small.

That means this level of significance is very high when we reject this null. So, with a lot of confidence and significance level, we can say that the proportion of promoters and detectors is not the same across both samples now, whether promoters have increased after the training and development program that will try to identify using the one-tailed test. So, we can conduct them very easily using the same function again.

Again, we will use a similar set of commands, sum Post_Bino, and we can provide the sum for Pre_Bino. Then we can also supply the sample size as 50, 50 for both the samples. We can supply the confidence level as 95 percent. And then, we can provide the alternate hypothesis, which is greater. Prop.test(x=c(sum(Pre_Bino), sum(Post_Bino), n=c(50,50), conf.level=0.95, alternative= "greater")). So, here we are testing whether the proportion of promoters in the post-sample has increased from the pre-sample. Let us test that.

Please notice that the chi-square value and corresponding p-value indicate. The p-value is, in fact, quite low indicating that the proportion of promoters in the post sample is much higher as compared to post samples. So, we are able to reject the null with a lot of confidence and a very high level of significance that the proportion of promoters is considerably higher in the post-sample as compared to the pre-sample.

Although not needed, we can also test the other hypothesis that is of less. That means the promoters are less in post sample as compared to the pre-sample. And we can test that hypothesis very easily. Look at this; the p-value is one which indicates that this hypothesis was spurious. We already knew that promoters in the post-sample are more as compared to the pre-sample. So, therefore, the other way around that is promoters are more in pre-sample as compared to post-sample.

That hypothesis is spurious; we cannot reject the null, as we have already discussed. So, in conclusion, we can say that this hypothesis as a testing exercise that because of the training and development program, there is a very statistically significant increase in promoters after the training development program as compared to before the training and development program. **(Video Ends: 01:04:46)**