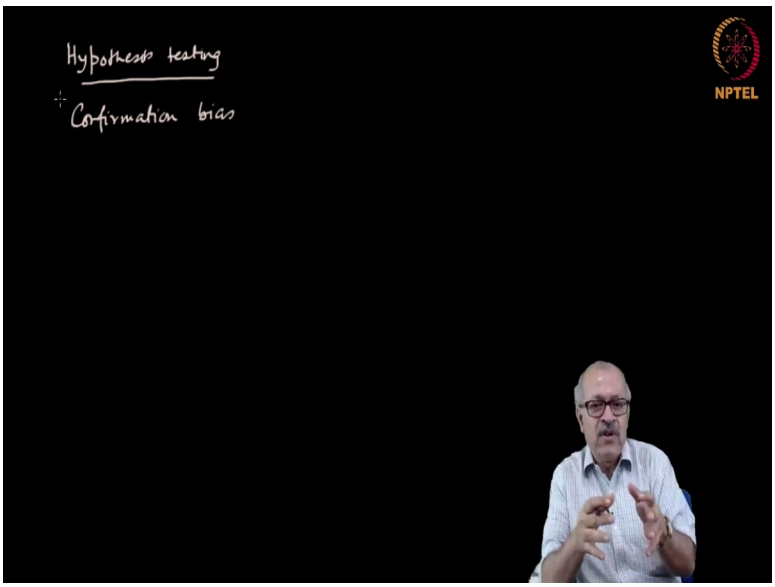
Research Methodology
Prof. Soumitro Banerjee
Department of Physical Sciences
Indian Institute of Science Education and Research, Kolkata

Lecture - 43
Statistical Methods in Hypothesis Testing: Z-Test and T-Test, Part 01

(Refer Slide Time: 00:20)



In the last class we were dealing with Hypothesis Testing. We learnt that we will have a few hypotheses in hand and our main objective is to identify and falsify the wrong ones. We learnt that the main confounding aspect in hypothesis testing is what is known as confirmation bias. I have a hypothesis; I believe in that hypothesis and the whole purpose of my testing is to find confirmation that the hypothesis is true. This is the worst possible mistake a scientist can make.

A scientist has to conduct an experiment with a completely clean mind, without any bias. Therefore, the plan of doing the experiment has to take into account the possibility that the experimenter himself or herself could be biased. And therefore, there are a few things that we build into the experimental plan.

For example, we always start with believing the null hypothesis rather than the alternative hypothesis. Only if we find enough evidence to believe that the null is false, then we go in favour of the alternative hypothesis. That is point number 1. So, we start

by assuming or believing the null hypothesis and work out the consequence of the null hypothesis, and test that, rather than testing the actual hypothesis.

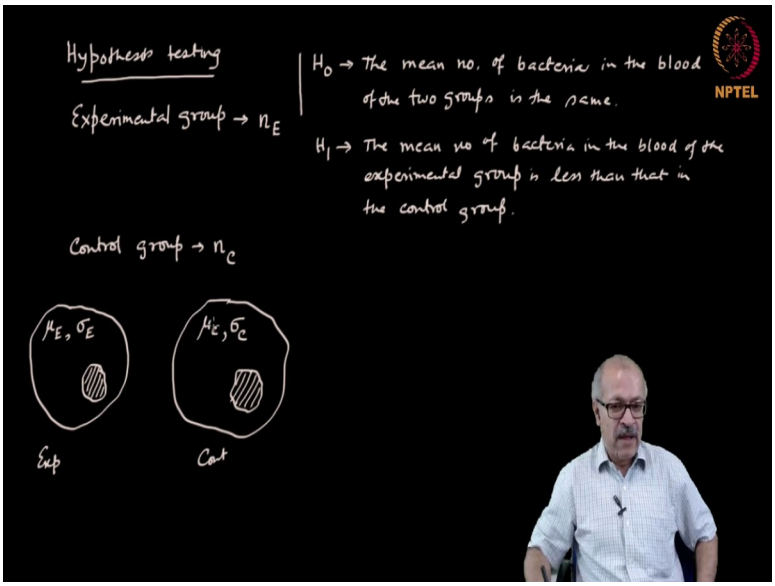
The second point is that, we always try to do either a single blind test, or, in case of experimental subjects being humans, a double blind test.

Single blind, because the experimenter—if he or she is biased—then he or she may take the readings wrongly. That possibility exists, and therefore, the person who is taking the readings should not know which sample is coming from the experimental group and which sample is coming from the control group. That is a single blind test, which is conducted so that there is no possibility of the bias of the experimenter influencing the result.

In case of the experimental subjects being humans, with their own psychological inclinations, it is necessary that they subject themselves also do not know whether they belong to the experimental group or the control group.

We were taking a particular example, but it is applicable almost everywhere. We took the example of drug discovery, where we took a particular disease which results in some bacteria build up in the blood. So, in order to test it, you would draw the blood and you would count the number of bacteria or parasite like plasmodium that are there in the blood per unit volume. That gives a number, some kind of a numerical test, of the hypothesis. That way we try to eliminate all possibilities of confirmation bias. That is our method of going ahead.

(Refer Slide Time: 04:14)



So, we start now with the statistical test. There is an experimental group and there is a control group. You have initially collected samples and then we have divided them into the experimental group and the control group. It is not necessary always that the size of the groups should be exactly the same.

Let us assume that this number is, say, n in the experimental group and it is m in the control group. Now notice that, as far as the experimental design is concerned, the way we have designed the experiment, the null hypothesis and the alternative hypothesis can be stated as follows. The null hypothesis, H0: we can state it as the mean number of bacteria in the blood of the two groups is the same and the alternative hypothesis is that they are different.

So, we measure the mean number of bacteria in the blood of the experimental group. Now, we have administered the drug in the experimental group and if the drug is effective, the mean number of bacteria will be less. So, this the alternative hypothesis is that the number is less than that in the control group. This is how the actual measured things are to be tested.

Now let us try to figure out what we are actually doing. There is a population of people afflicted with that disease and we are trying to figure out, if we apply the drug to every member of that population, what will be the effect of the drug? We can think of a whole

population of people who are afflicted with the disease, and who have been administered with the drug. That is a population of people.

If we can get the blood samples from all of them, then we can definitely get the actual values. But we cannot get the blood sample of all of them. We have actually drawn a number of samples from the population. So, there was a population, and we have drawn samples. From that sample, we are trying to figure out the result.

So, the point I am making is that, there is a population out there. What is the character of the population? The population of people afflicted with the disease and who have been administered the drug. That is the population. There is another population, the control population, who are afflicted with the disease, but have not been given the drug.

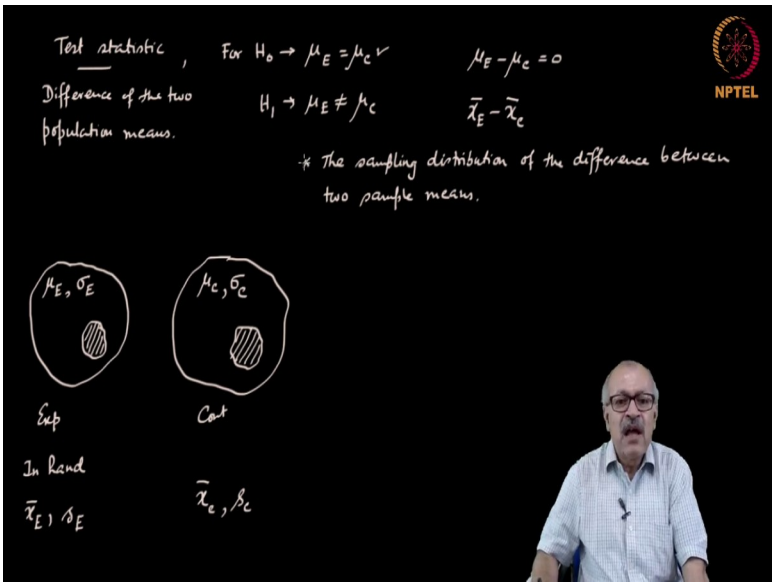
So, there are two populations. Let me depict it this way. Suppose we have a population of people, the experimental; and there is another control population. This is the population of people, and within that, we have drawn a sample. There is a population of people of the control group and here also we have drawn a sample.

And we actually have this data. But we are trying to figure out the character of the population, the population of people who have the disease and who have been given the drug, and the population of people who have the disease and who have not been given the drug. But we are trying to make an estimate of that using a smaller sample. That is the problem that we have in hand.

Let us define. In the experimental population, there will be a mu, mean of the experimental group. Similarly there will be the mu of the control group. And within the experimental group there will be a standard deviation. There will be a standard deviation in the control group also.

What does this mean? It means that, if you somehow were able to take blood from all the people who are afflicted with the disease and who have been given the drug—all the people—then also there will be a mean and there will be a standard deviation. Everybody will not react to the administration of the drug in the same way. So, there will be a variation and there will be a mean as well as a standard deviation. Similarly, in the control group.

(Refer Slide Time: 11:06)



But we do not have that data. What do we actually have in hand? What we actually have in hand are the average of the experimental group and the average of the control group, the sample standard deviation of the experimental group and the sample standard deviation of the control group. These are the things that we have, and we are trying to figure out the character of these. That is the problem.

These are the things that we have in hand and these are things that we are trying to arrive at. What should be the test statistic: the thing that we are trying to measure, whose characteristic we are trying to test?

Now, what is our real intention? We are trying to figure out how these two groups differ. As I said, we will first believe in the null hypothesis and then we try to figure out if the data induce us to reject the null hypothesis.

What does that mean, if the null hypothesis is true? According to the numbers that we have obtained, for the null hypothesis we can say that the mu_E should be equal to mu_C, and for the alternative hypothesis we have to say that mu_E is not equal to mu_C. This is what we have to test. Ultimately all those written descriptions boil down to these numbers. So, this is what we are trying to test, and we start by believing the null hypothesis.

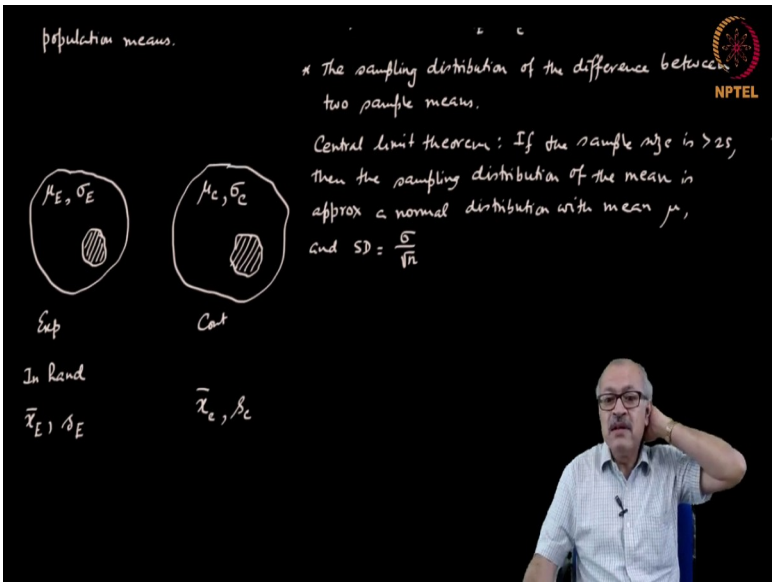
Now, if the null hypothesis is true, then mu_E minus mu_C should be equal to 0. We start with the assumption that the null hypothesis is true, and therefore, mu_E minus mu_C is equal to 0. That indicates that the difference of the two population means is the test statistic. So, the test statistic is the difference of the two population means. That is what we have to check.

But we do not really have this in hand. We have x bar E and x bar C, the actual measured values from the smaller sample. So, since we do not have this in hand. What we actually have in hand is the x E bar minus x C bar. That is what we have in hand. So, this is our test statistic. Suppose I have taken 100 samples, out of that some number in the experimental group another number in the control group. I have drawn the blood, I have counted the number of bacteria in the blood samples per unit volume and that has given me some value.

If I now do the experiment again, repeat it by collecting a different sample and do the same way, I will get a different value of x_E minus x_C. The third time I will get a different value. So, this will have a distribution. But if the null hypothesis is true, then the mu_E minus mu_C should be equal to 0 and therefore, that distribution should have a mean 0.

If we take such measurements again and again, every time we will get a different value of this. So we will get a distribution. That will be the sampling distribution of the difference between the two sample means. So, it will be this distribution.

(Refer Slide Time: 17:06)



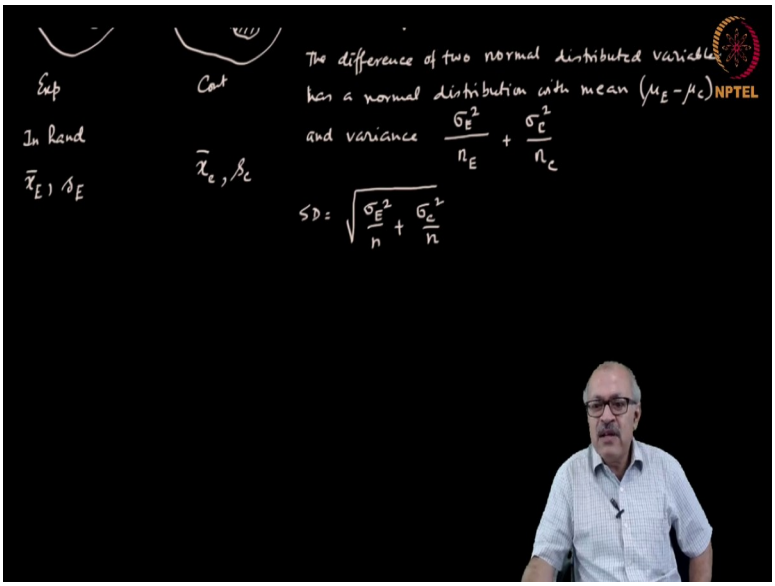
Will that be a normal distribution? Now, the central limit theorem says that if you conduct an experiment with at least 25 data points taken, then the sampling distribution of the means will be a normal distribution, with mean of the normal distribution at the population mean and the standard deviation equal to the standard deviation of the population divided by the square root of the number of data points.

If the sample size is greater than 25, then the sampling distribution of the mean is approximately a normal distribution with mean at the population mean and standard deviation is equal to sigma by square root of n. This is what we learnt.

This is for the data taken. But here we are talking about a difference. Now, there exists another theorem in statistics—we are not going to the details of it, but I can state that—if both x_E and x_C are distributed in a normal distribution, then the difference also has a normal distribution.

Now, x_E will be distributed in a normal distribution. That is guaranteed by the fact that the number n_E is greater than 25. The number n_C is also greater than 25. Therefore, each of these will be distributed in a normal distribution. If these are distributed in a normal distribution, then the subtraction, the difference, will also be distributed in a normal distribution. But what will be the mean of that?

(Refer Slide Time: 20:13)

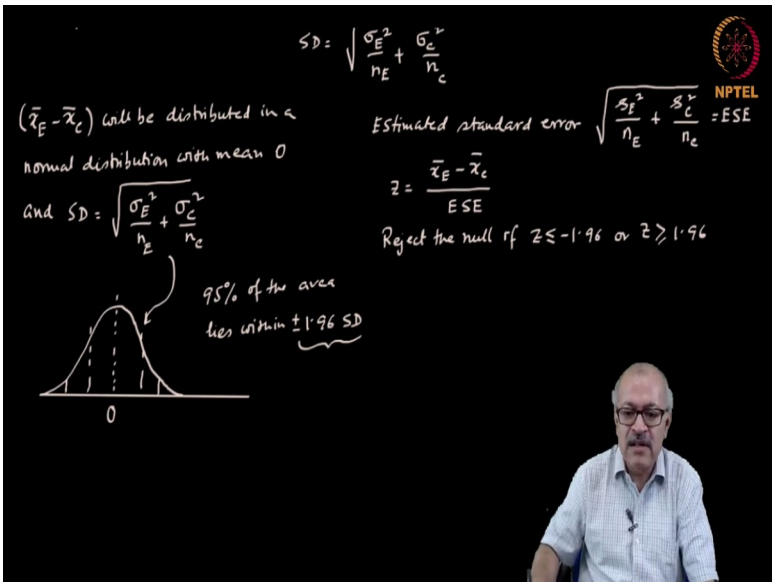


Let me write: the difference of two normal distributed variables has a normal distribution whose mean will be the difference of the two means. The variance? The first distribution's variance by n_E plus the second distribution's variance by n_C. Therefore, the standard deviation SD will be square root of sigma E square by n plus sigma C square by n. So, that would be the standard deviation of the difference distribution.

Let me just repeat. If you have two variables that are each distributed in a normal distribution, then their difference is also distributed in a normal distribution with mean at the difference between the means of the two distributions, and the variance equal to this.

Now, this is all very good. We are, sort of, guaranteed that we are dealing with a normal distribution. But we do not have this value. But we know that if the null hypothesis is true, what we started with, then this value is 0.

(Refer Slide Time: 22:56)



It means that these quantities, x E bar minus x C bar, if you make repeated observations, these will be distributed in a normal distribution with mean 0. And why the mean is 0? Because this is true; this is 0. Because we have started by believing the null hypothesis. If the null hypothesis is true, then it is 0 and standard deviation this.

That gives us a way of testing, because if it is a normal distribution like this, then it will have mean 0 and the standard deviation equal to this. Then we know that 95 percent of this area lies within plus or minus 1.96 standard deviations. And since we know the standard deviation, we can find out within what range will 95 percent remain.

Now, if the value that I have got, the value of x_E minus x_C, this value that I have got, is beyond this, then it is a very unlikely thing to happen. Because 95 percent of the time this will be in this range. Only 5 percent of the time it has any possibility of being outside. If I have got this value, then it is a reasonably good evidence that the null hypothesis is wrong.

If the null hypothesis is true, then this would be the distribution and then it would be very unlikely to get a value which is beyond this range of confidence. So, if I get a value which is beyond this range of confidence, then we can have a 95 percent confidence in saying that the null hypothesis is false. Otherwise we will say that we do not have sufficient evidence to reject the null hypothesis. So that is the essential logical structure.

One problem we still have. That is, we do not have the values of these, because these are something that belongs to the population. We have the sample values. But as always, we will substitute the population values by the sample values, so that we get something in number. If the n's are large, then more or less this assumption is a valid assumption. So, we estimate the standard error.

The estimated standard error will be the square root of: we will substitute this by s_E square by number of experimental group plus s_C square by the number of the control group. So, that is the estimated standard error. That is the ESE we are trying to find out. That will be the standard deviation of this distribution.

Now, we are trying to find out whether the data I have got is beyond 1.96 ESE. The general way of doing it is to define the z. z is equal to what we have got: x E bar minus x C bar by the estimated standard error. And we will reject the null hypothesis if z is very large.

So, reject the null if z is below minus 1.96 or z is above 1.96. This is the standard test of the hypothesis, where we are saying that, with 95 percent confidence we can reject the null hypothesis; or we cannot reject the null hypothesis with a 95 percent confidence.

So, 95 percent confidence is the yardstick. As I have said, in different fields you might need different extents of confidence. Depending on that you will have to set this number. But in general, for 95 percent confidence this is 1.96.

Let me give an example to illustrate how this is actually done.