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

**Lecture - 45**
**Hypothesis Testing: The Chi-Square Test - Part 01**

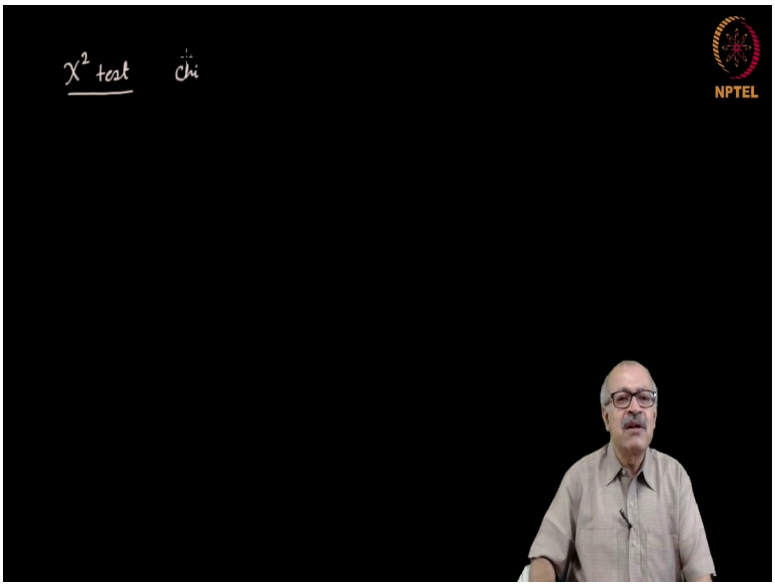
In the last class, we learnt that whenever we want to test a hypothesis, we have to base ourselves on the null hypothesis. And the null hypothesis is always of the equality type. If you define what you test as an equality, it becomes a very easy procedure to test whether that can be accepted. Because then you get a distribution of the means, which will be have a certain character which you can test. We have learnt that.

So, we have learnt how to test hypotheses when you have a situation where there is a experimental group and a control group and you are trying to figure out whether there is a statistically acceptable difference between the two groups. If the number of samples in each group exceeds 25, then we can anticipate that the distribution of the means or the difference of means will be a normal distribution, and on that basis we developed z test.

We also said that, due to some reason, we may be unable to get that many data points. There can be various possible factors, a particularly rare disease for example, where that many patients cannot be found. A dwindling population of a particular species which is going to go extinct and in that case you cannot get that many samples of that particular species. A situation where getting one data point incurs a lot of expense and you cannot do that. Travelling to Antarctica, for example, you cannot do that every now and then.

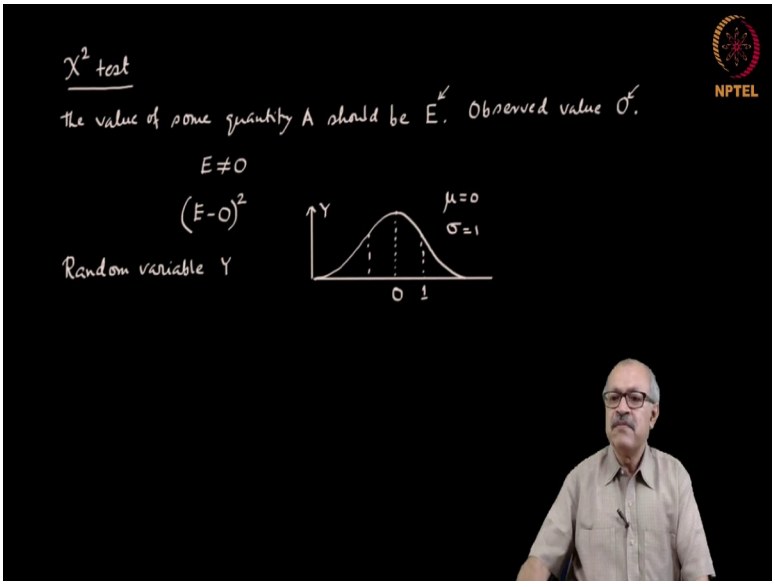
So, in those cases we do use a different statistics which is also acceptable, the t statistics and by that we proceed. Now, there is a third test that is often employed in statistical hypothesis testing. It is called the chi square test.

(Refer Slide Time: 02:49)




Chi square test: that is what we will be doing today. Chi is a Greek letter written somewhat like X but curly $\chi$. So, that is how I would write it. It is written as chi, but pronounced as kai not chi.

(Refer Slide Time: 03:36)




Now the basic idea is, suppose you have got a hypothesis or a theory, which predicts that the value of some quantity A. The value of some quantity A should be some expected value E. So, by theory this is expected. And you do the experiment and you get an

observed value, say, O. This is what you have got. And obviously, E should not be equal to O; you will not get them equal.

Normally you will never get them to be equal. But by that, can you say that the theory is wrong, because you did not get the expected value? No, you cannot say that, because, due to various reasons—more important being the random fluctuations, random errors—you might get a value O, which is different from the value E.

And that does not provide you sufficient reason to say that the theory is wrong. But there would be some situations where you get a different value and you would be able to say that the theory is wrong. So, one has to have some kind of a criterion, on the basis of which we might be able to say that, according to our results of the experiment, the theory is false, or according to our experiment, we cannot say the theory is false.

What should be the generally acceptable criteria, because in all cases you will not get E equal to O. So, that is the problem we will be tackling today. Now when you have E, that means expected value from the theory, differing from the observed value from the experiment.

Then, if it is still true that the theory is true, in that case the quantity E minus O, this quantity should be behaving like a random variable. Right? I would not like the quantity here to be negative. Let us try to get something positive. Let us put a square here, so that this number is a square, a positive number, and that positive number should behave like a random variable.

So, we need to check whether it does behave like a random variable. That means, if you do repeated experiments, you get repeatedly different values of O, but E-O should behave like a random variable.

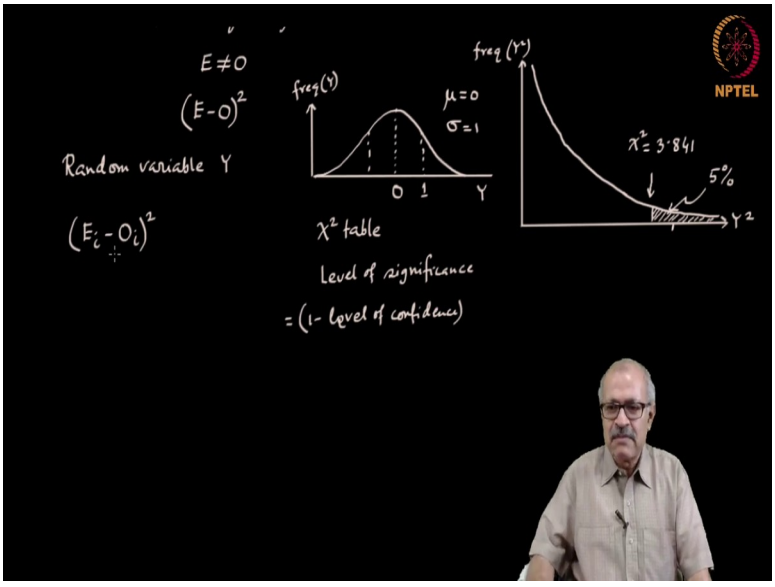
So, every time you make a measurement of E minus O, it should have different values. And the different values should behave like a random variable. If it behaves like a random variable, you are more or less certain that the changes, the differences between E and O are due to random fluctuations. You do not have enough reason to say that the theory is wrong.

But if there is a systematic fluctuation, then you would say the theory is wrong and in that case it does not behave like a random variable. In order to decide whether it is behaving like a random variable, you have to understand the behaviour of a random variable.

So, let us tackle that issue first; that means, we consider a random variable, say, Y. The random variable is distributed. We consider that it is a random variable, so it is distributed as a Gaussian function, as a normal distribution. It is distributed as a normal distribution, with a mean and standard deviation. The mean is 0 and standard division 1.

So, mean 0, standard deviation 1, a random distribution. So, this is how Y is distributed. If Y is distributed like this, how should Y square be distributed? Because, here we are dealing with a square of a random variable, we are trying to understand how would the square of a random variable behave. So, here our variable is Y. Now we are trying to figure out how should the square of the random variable behave.

(Refer Slide Time: 09:27)




Now, we are trying to plot the frequency of Y square here. This is my Y and this is the frequency of Y, which is distributed like a normal distribution with mean 0 and standard deviation 1.

How would the Y square be distributed? Notice that a large number of points of Y are very close to 0 because, it has a peak at 0. And since you are taking a square of that,

square of Y, therefore, it will become even smaller. So, if it is even smaller, than there would be a very large number of points very close to 0. Since Y square is positive, I can only draw the positive side. So, it will have a very large value at very close to 0 and then, it will taper off for higher values of Y.

By that logic, you can anticipate the behaviour to be something like this. So, this will be the distribution the frequency of this, frequency of Y square. As you can see, as the Y increases or Y square increases, the frequency of Y square decreases. Because all the negative sides will become positive, because you are taking square and very close to 0 it will have a very large value and as you take a square of a small number it becomes even smaller. So, you have very large distribution very close to 0 and it will taper off. So, this will be the character.

Now, notice the logic. Logic is that, if Y is a random variable, the behaviour of Y square would be something like this. Then, if we can find out, say, some point beyond which the area is less than 5 percent, the area is 5 percent of the whole area, if you get a Y square value beyond that, say somewhere here, then the frequency of that will be very small and the probability of getting such a Y square value will be very small.

So, the probability of getting a Y square value beyond this is 5 percent and therefore, if you get a Y square value beyond this, then it is probability is less than 5 percent. So, we can state with a 95 percent confidence that such a behaviour is not expected from a random variable.

We were testing whether this behaves as a random variable or not. We will check the behaviour of this against the behaviour of a random variable and if this falls somewhere here, beyond the 5 percent range, then we would say that it is very unlikely that this difference has been caused by a random variation. It is unlikely, because the probability of such value to occur is less than 5 percent. So, we say it is unlikely. So, that is the essential idea of the test.

Now, we have a situation, where we have this character of a random variable. This you do not have to plot, because some people have already plotted it and have calculated the areas under the curve.

Beyond some particular value of Y square, how much is the area? Another particular value of Y square, how much is the area? All these are actually tabulated in the chi square table. I will show how the chi square table looks, but the point is that the chi square table would be referred to find out what value of chi square is necessary in order to reject a hypothesis.

So, this is how we would proceed. Now, if you consult the chi square table, then you will find that 5 percent area lies beyond the chi square value of 3.841. So, if in any experiment we get a value beyond this, then we will be able to say that difference is unlikely. We are unlikely to get a difference beyond this if the differences are only because of random fluctuation, because it does not behave like a random variable. The probability of that being obtained using a random variable is only 5 percent or even less than that.

Now, you know that we have been talking about the level of confidence, and there is another thing: the level of significance. The level of significance that is equal to 1 minus the level of confidence. In the chi square table, it is normally given in terms of the level of significance, which means that, if it is given in terms of 5 percent, that would mean that the level of confidence is 95 percent.

So on that basis let us go ahead. One major advantage of this chi square method in comparison to the other methods that we have come across is the following. In the case I have illustrated, there was a particular expected value, and a particular value obtained. But a theory might predict different values for different situations. A single theory can predict different values for different situations. In that case, say, there are n different situations, and for each situation there is the expected one and then you do the measurement for that situation and you find another the observed one.

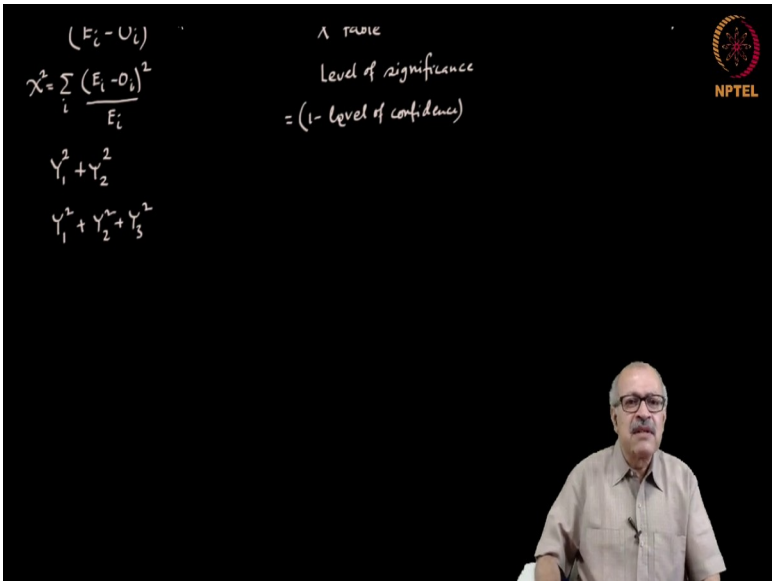
Which means, you actually have for each value of measurement, E_i minus O_i and its square. E_i means the ith situation. O is a value which is obtained; that means, in measuring the observed value you have already taken 25 readings and therefore, you have obtained a mean value, you have obtained the standard deviations—all that have gone into the measurement of O.

But that is for a particular situation, for which there was a particular prediction. And since for different situations the same theory can have different predictions, there can be

i different situations and therefore, there will be i different observations and therefore for each one the difference will be quantified like this.

So, in that case we are talking about that E_1 minus O_1, E_2 minus O_2, each one will be distributed like a random variable. So, we are essentially talking about summation of the squares of random variables. It is something like this.

(Refer Slide Time: 19:27)




If we define our chi square as these summed up and we are trying to figure out whether that the whole thing is behaving like a random variable. We are talking about a sum of E_i minus O_i square, sum over i.

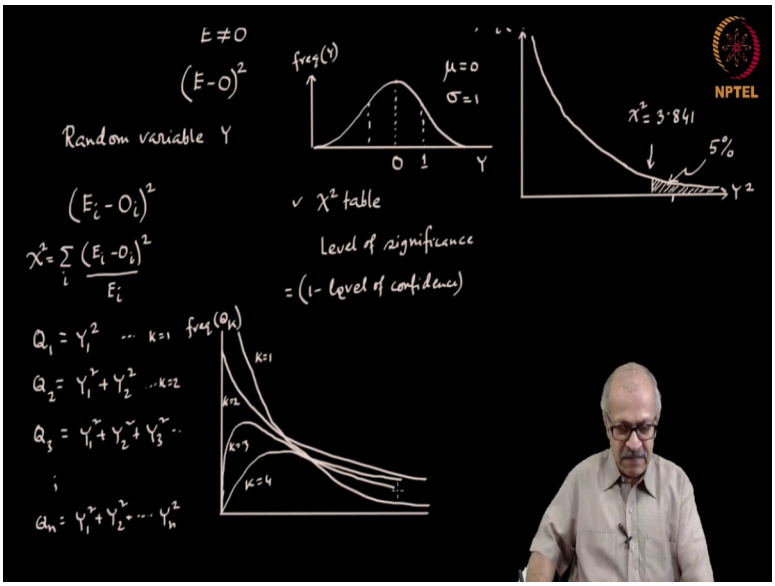
Not only that. See, if the expectation value E_i is large, then you expect this difference also to be large, but if the expectation value is small, a small number like 0.01, then, obviously the difference is also expected to be small. So, this absolute magnitude of the difference does not really make sense, you have to normalize it with respect to the expected value.

So, if you define the chi square like this, you would notice that it is nothing but a sum. What are we assuming? We are assuming that these differences have happened due to random fluctuations. So, it is nothing but a summation of n number of random variables. Therefore, in order to understand the behaviour of that, what shall we do? We will try to understand the behaviour of summations of random variables.

Now suppose, there are two such random variables, say, Y_1 and Y_2. We have to take a square of them because, we want positive numbers. So, Y_1 is the random variable distributed this way this way and Y_2 is another random variable distributed exactly in the same way. Then how would this summation behave? Then if there are two different situations regarding which the theory has a prediction, you will be able to match with that.

If the theory has prediction regarding three different situations, then you would need to take Y_1 square plus Y_2 square plus Y_3 square. This you have to take. That way it continues. Normally we give it a particular name.

(Refer Slide Time: 22:18)




We say that, let us say, a variable Q_1 is defined as simply Y_1 square, where Y_1 is a random variable. Q_2 is Y_1 square plus Y_2 square, where Y_1 and Y_2 are both independent random variables. Q_3 is equal to Y_1 square plus Y_2 square plus Y_3 square, but all three are independent random variables.

This way, without going through any experiment, independently we can study the behaviour of these quantities Q_1, Q_2, Q_3 in a graph something like this. If you go on it is Q_n is Y_1 square plus Y_2 square plus dot dot dot dot Y_n square. And the number that we have, the number of these that we take, is called the degree of freedom, K. Now, how will they look? Let us try to figure that out.

So, we are trying to get the distribution of these values of Q_1, Q_2, Q_3. Earlier we had only obtained the behaviour of Y_1. In this case K is equal to 1. This is K is equal to 2, so on and so forth. You understand that this K is equal to n.

So, how many of these are there, that signifies the degrees of freedom. So, here we are plotting that and we are now trying to find out the frequency of that. We have already worked out the character of the first one, which is K is equal to 1. Its character was something like this, we have seen that.

So, that is K is equal to 1. Now K is equal to 2. The logic of this one was that here the centre point is 0, there will be a large crowding of numbers close to 0. Now, when you take two of them, each one will have a large crowding of numbers close to 0, but when this one is 0, this one need not be 0, when this one is 0, this one need not be 0. So, a very large crowding near 0 will be relatively less and so, it will be something like this.

So, that will be the character of K is equal to 2. Now, when you take three of them, then when this one picks up a value close to 0, these are random variables, so there will be very little probability that both of these will be very close to 0. And so, actually the density of points or the probability of points being close to 0 will be very small. So, it will actually start from 0. It will be something like this. So, that is K is equal to 3. Similarly, K is equal to 4 will be something like this, so on and so forth.

My point is that this is the frequency of Q_K, where K is the degree of freedom. These have actually been plotted and from there the table has been extracted. So, we already have the data available: how a random variable or a collection of random variables together behave. Depending on the experimental situation we will apply one of them.