Mine Automation and Data Analytics

Prof. Radhakanta Koner

Department of Mining Engineering

IIT (ISM) Dhanbad

Week - 9

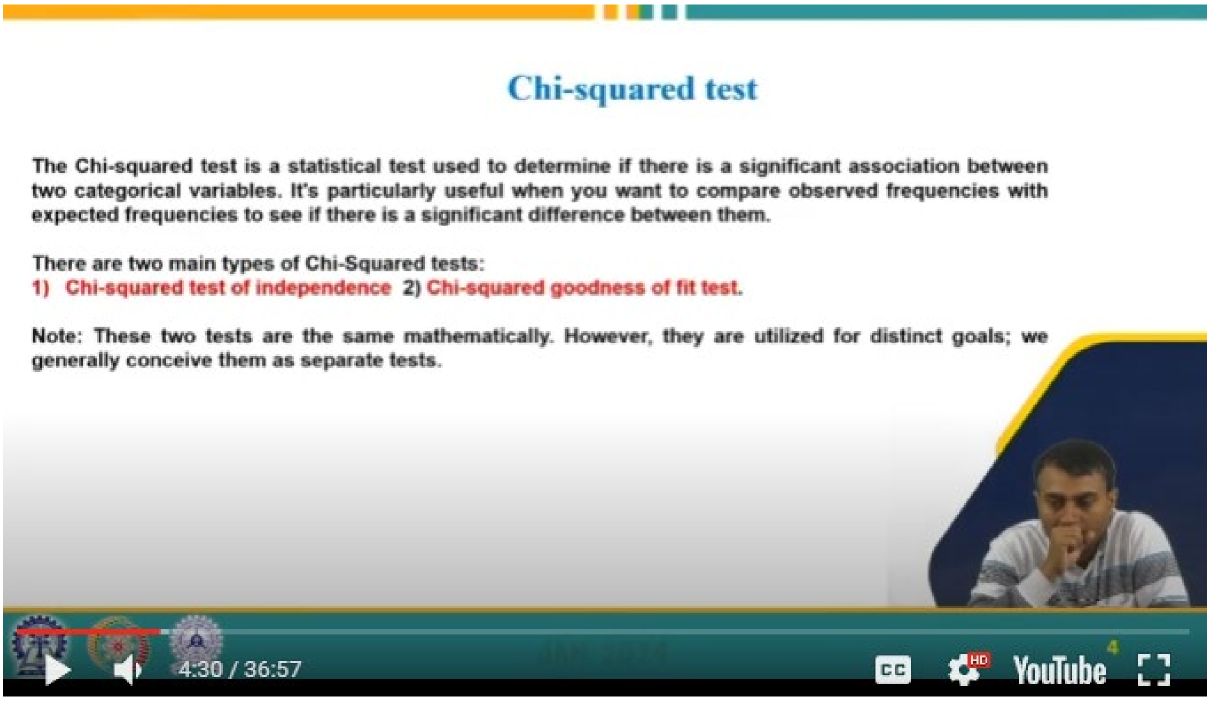
Lecture 44: Chi-Squared Test

Welcome back to my course, mine automation and data analytics. Today, we will discuss one of the hypothesis tests, the chi-square test. So, in this lesson, we will discuss the following: we will first introduce you to a chi-square distribution. As we have observed in other hypothesis testing, we will enumerate you with real-life examples and then what the assumptions required. We will discuss two variants of the chi-square test for independence with examples for this particular test, the chi-square test. Another is the chi-square goodness-of-fit test, with examples. So, what is the chi-square test? So, this test is a statistical test that determines the association between two categorical variables. So, let us take one real example. Elections are coming, and various political parties are working on and propagating their agenda.

Now, let us do some surveys on gender preference in political parties. So here, if you assume that some political agenda means that a political party is X, another political party is Y, and there are males and females. So, there may be an option for people to think that they are not supporting these political parties. You may be collecting a sample that, for example, 1000 people. You have done some sampling, preferably 50-50, on the male and female, and let us see how many females support X, how many females support Y, and independent similarly for the male. So this data is categorical data representing that in thousands of populations, this is their political party choice preference. We can estimate using this chi-square test whether these data strongly represent something or not, for example, whether the gender has any preference for any political party.

So, this particular test would benefit the campaign manager doing this job for their respective political party. So, this is a handy statistical tool. Similarly, for the mining engineering example, we can think of an accident that happened on a mine site or in a mine, whether this accident has any relation with age and experience, the frequency of the accident, and in that particular mine site, the involvement of the miners their age and experience. Does this have any relation? This is a perfect example of how mining engineers and statistical people use this kind of tool for mine safety engineering and ergonomics, so there are many uses, and different research papers are also in the public domain you can go through. So, this chi-square test is relevant and helpful for this situation.

So, this test will compare the observed frequencies and whether these frequencies are significantly different from the expected ones. So that is the goal of this specific test, and based on that data, we will get the from the data table we will match, and based on that, we are going to conclude something. So before going to that, let us introduce the two different types of chi-square tests that we will give you the illustration today. One is the chi-square test of independence, another is the chi-square goodness of fit test. Ideally, the mathematical approach of the mathematical framework is equal for both these tests, but they are helpful for different purposes and that is why we often consider them as a separate tests okay.



Chi-squared test

The Chi-squared test is a statistical test used to determine if there is a significant association between two categorical variables. It's particularly useful when you want to compare observed frequencies with expected frequencies to see if there is a significant difference between them.

There are two main types of Chi-Squared tests:
1) Chi-squared test of independence 2) Chi-squared goodness of fit test.

Note: These two tests are the same mathematically. However, they are utilized for distinct goals; we generally conceive them as separate tests.

Let's start with the chi-square test for independence, okay? So where we are going to use it this testing method we are going to use this method to determine the significant association with the two categorical variables that I just gave you the example. One example is the gender preference with the political party or the influence of age and experience with the accident or the frequency of the accident. So these are an example, though we are going to illustrate you with some examples that will help you understand clearly what the chi-square test for independence is. So, this test is beneficial for examining the relationship between these two variables.

We are researchers, and we are also working on this domain. We are so based on whether these two categorical variables have a significant influence or significance dependent on each other. So, we can use this particular tool and this method to reach a fruitful conclusion or a truth. The truth was not known by this testing method. We are reaching the truth. So before proceeding to that stage, we have to frame the hypothesis like the other hypothesis testing method.

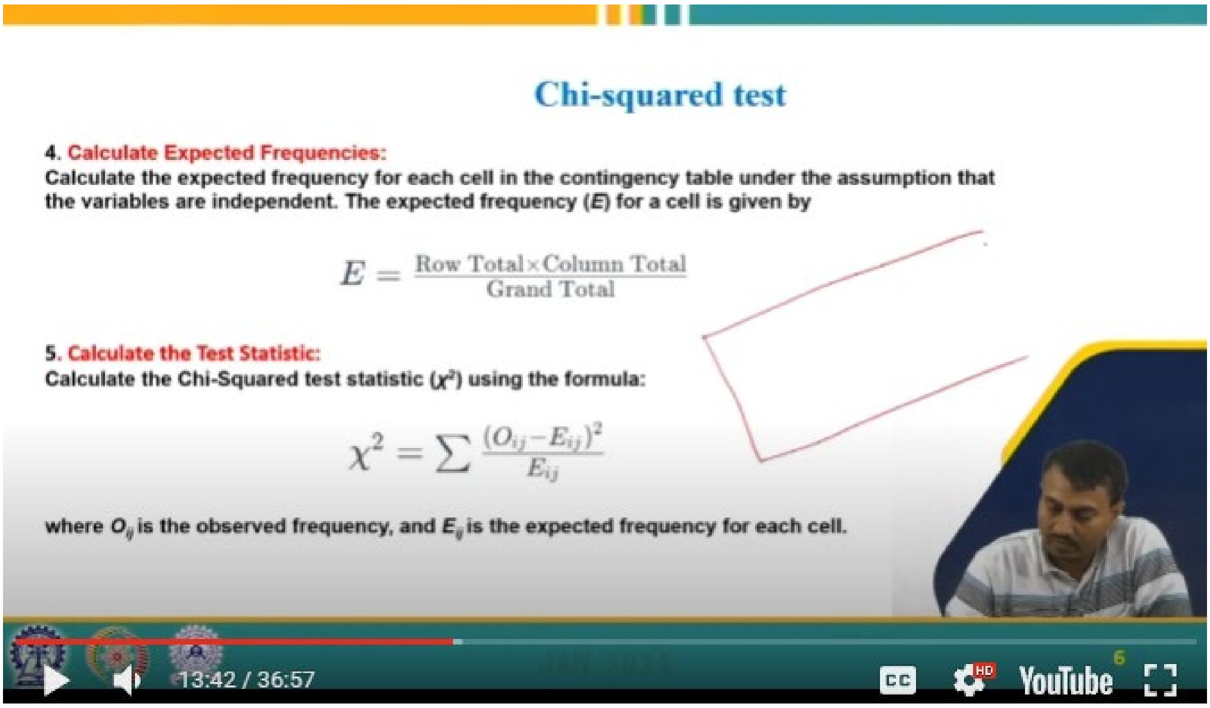
So, how do you frame the hypothesis? So here we are following specific nomenclature. So one is the null hypothesis, similar to the other testing method other hypothesis testing which is hypothesis H0. So, this H0 for the chi-square test for independence represents that there is no significant association between two categorical variables. So, we start with the assumption that gender does not have any preference with the preference of the political parties. So, the notion is a population is equally distributed 50-50, and gender does not have any preference for any political party.

Let us assume that. Get the data by sampling it, then analyzing it. So if the analysis based on the statistics it is found that we are not able to reject the null hypothesis, then there is no significant association, or if you can reject the null hypothesis, then we have to accept the null hypothesis that there is a significant association between the two categorical variables okay. So these can be tested in widespread choice are 0.05, 0.01 and 0.1 significance level α and we have to collect and organize the data and here we are introducing a new concept called contingency table.

$$E = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

So, I have given you the example of the gender preference with the political party. So let us assume that X is a political party. Next, the Y, another column then another is independent okay, another column then categorize on the row wise male and female. So how many males prefer X, how many males prefer Y, how many males prefer independence, and in the second row, how many females prefer X political party Y and the independent?



Chi-squared test

4. Calculate Expected Frequencies:
Calculate the expected frequency for each cell in the contingency table under the assumption that the variables are independent. The expected frequency (E) for a cell is given by

$$E = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

5. Calculate the Test Statistic:
Calculate the Chi-Squared test statistic (χ^2) using the formula:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

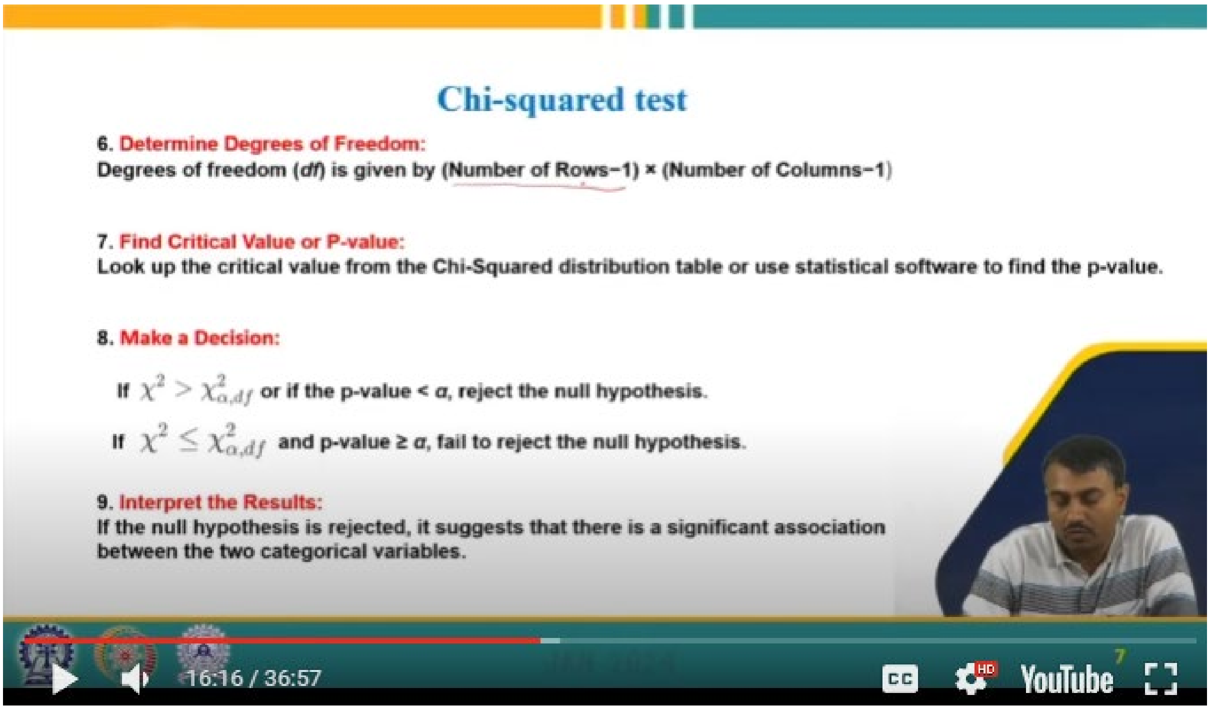
where O_{ij} is the observed frequency, and E_{ij} is the expected frequency for each cell.

That represents the contingency table and in statistical community also this table is many people called this two-way table okay. These display the frequencies of each combination of these two categorical variable okay. And how do we calculate the expected frequencies? Expected frequency is calculated based on this formula for each cell: E is equal to the row total, row total so male was first, then the row total, and X is the first parameter first political party column total okay. So suppose 100 is 150, okay? The total may be 500, okay? So total is total is 500.

So, the row total and then the column total are multiplied and divided by the total, which basically represents the expected frequency for each contingency table cell. So each cell of the contingency table was, as I said, independent okay, and this is male, female, then there should be total okay, here is also total. So this is the row total multiplied by the column total and divided by this total, okay? This is an example. Now, this is the expected frequency for each cell.

So after this calculation, we will put this as okay here. Some observations are there, and some observations are OK, and based on that, we will calculate the expected frequency. Then, we will calculate the Chi-square test statistics. X square is equal to suppose it was found earlier 60, it is 50, and it is the data, and it has found that expected was 55, this is expected 55 for example, okay. So this 60 minus 60, 55 square divided by 55 is okay, and we will put that value in each cell, okay? We then sum over all cell values, which basically gives you the Chi-square test statistic value.

So here, similar to t-distribution, we have a degree of freedom. The degree of freedom is given by the number of rows minus one into a number of columns minus 1. So here it was: 3 minus one into two minus one is equal to 2 into one is equal to 2, degree of freedom for the example that I just enumerated. Then, find the critical value or p-value. So this critical value can be seen from the Chi-square distribution table, or if we want to find out the p-value, then we have to use the software okay.



Chi-squared test

6. **Determine Degrees of Freedom:**
Degrees of freedom (*df*) is given by (Number of Rows−1) × (Number of Columns−1)

7. **Find Critical Value or P-value:**
Look up the critical value from the Chi-Squared distribution table or use statistical software to find the p-value.

8. **Make a Decision:**

If $\chi^2 > \chi^2_{\alpha,df}$ or if the p-value < α, reject the null hypothesis.

If $\chi^2 \leq \chi^2_{\alpha,df}$ and p-value ≥ α, fail to reject the null hypothesis.

9. **Interpret the Results:**
If the null hypothesis is rejected, it suggests that there is a significant association between the two categorical variables.

So, this p-value, this Chi-square distribution table, is available. So based on that, we have to reach a decision. So, how do we reach a conclusion? So for the Chi-square test for independence, if this value is found to be greater than the Chi-square distribution table based on that value and the significant level, suppose 0.05 and the degree of freedom is 2. If this value is found particular from this table value and the calculated value is greater than, then, we reject the null hypothesis.

So if we reject the null hypothesis, then it suggests that there is a significant association between the two categorical values, and else if this calculated value is found to be less than equal to the value found from the Chi-square table distribution table, then we fail to reject null hypothesis instead we accept the null hypothesis that there is no significant association between the two categorical variables. So, this is the Chi-square critical value. This particular column is popularly used with a 0.05 significance level, and here is the degree of freedom. So earlier, it was 2, and it was found to be 5.991 okay for the degree of freedom 3, 7.815. So, what is the assumption? There are some assumptions, and these assumptions must be satisfied and must be followed sincerely. we have to follow these assumptions in a better way. So what is one? One is the independence of observation. Independent observation, for example, that the example I have just clarified to you, is the gender preference with the political party okay.

So who is collecting the data? Whether we are collecting the data through a very reputed agency, which has very good skill in gathering that kind of data, or that agency has a perfect clear image in the population, this particular agency is very, very reliable. The data will not be leaked to anybody. So people will give their choice and express their choice voluntarily okay. Also, they are ensured that there should not be any bias. That one observation just in front of one population, one particular person you are taking the data, another in front of that, no.

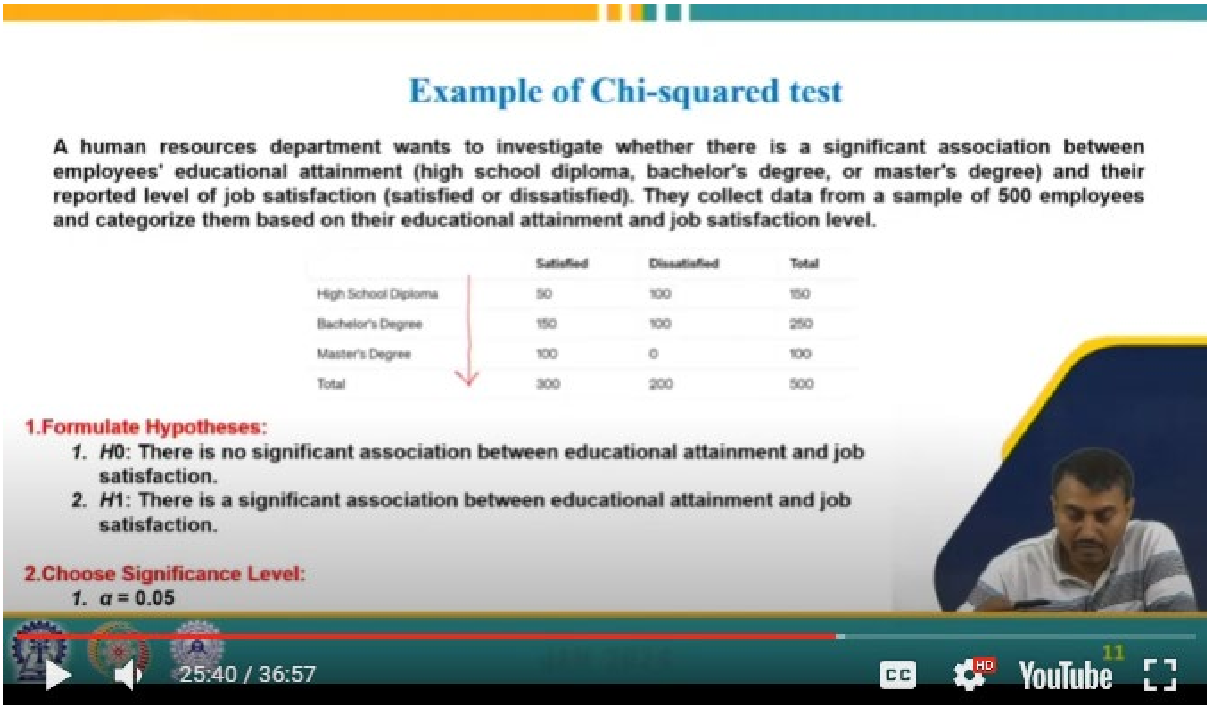
You should take the data in a very isolated mode so that another person choice is not influencing one observer or one person's choice. So, we must maintain that observation should not affect the occurrence or value of the other observation. That is the independence of observation. In random sampling, we have to ensure that data is collected throughout the country because there are different states that states have different preferences. So you have to exhaustively collect the data from all parts of the country to reach a generalized conclusion that these data represent the population of this country.

This helps minimize bias and ensure that the results are generalizable. Yeah, that's what we have to provide. That is why the sampling plan should be laid out before: what are the pockets of areas you are identified in a whole country, and what is the critical size, so you must basically select first? Then, the categorical data, this chi-square test, is designed for the categorical data, meaning that variables or analyses are divided into two distinct categories or groups. So, it is not appropriate for numerical data or continuous variables. The expected frequencies in each contingency table cell should be greater than or equal to 5.

The frequency data we try to avoid the because if the value is small, the value is 0 or near that going to give you some unreliable results. So, in those cases, Fisher's exact test is more appropriate. With a large sample size, mathematically, there is no strict restriction or requirement for the sample size. Still, it is always better when we predict the gender preference with the political party choice. So the population size should be large. A large population size will increase and give a more reliable prediction or the statistical correlation of the inter-relationship or association.

Here, we are not using the correlation because that correlation is not appropriate here. So we are always trying to find the association that is okay. Mutual exclusivity and exhaustiveness, that is, this category within each variable should be mutually exclusive that each observation should belong to only one category, and exhaustive means all possible categories would be represented in the analysis okay. So, these ensure that every observation is accounted for and avoid ambiguity in interpretation. And no cell count should be 0, so when it is 0, that might lead to some undefined results and computational issues, so we must avoid that.

So, we must adhere to all these assumptions to ensure that the chi-square test produces an accurate and meaningful result for finding out the association between the two categorical variables. So, if any of these assumptions are violated, then there might be some issue or compromise in the test result. So, some kind of necessary adjustment might be required for that kind of case. So, let us start with the example. So here we have given one instance: educational level, high school diploma then bachelor degree then master degree.



Example of Chi-squared test

A human resources department wants to investigate whether there is a significant association between employees' educational attainment (high school diploma, bachelor's degree, or master's degree) and their reported level of job satisfaction (satisfied or dissatisfied). They collect data from a sample of 500 employees and categorize them based on their educational attainment and job satisfaction level.

	Satisfied	Dissatisfied	Total
High School Diploma	50	100	150
Bachelor's Degree	150	100	250
Master's Degree	100	0	100
Total	300	200	500

1.Formulate Hypotheses:

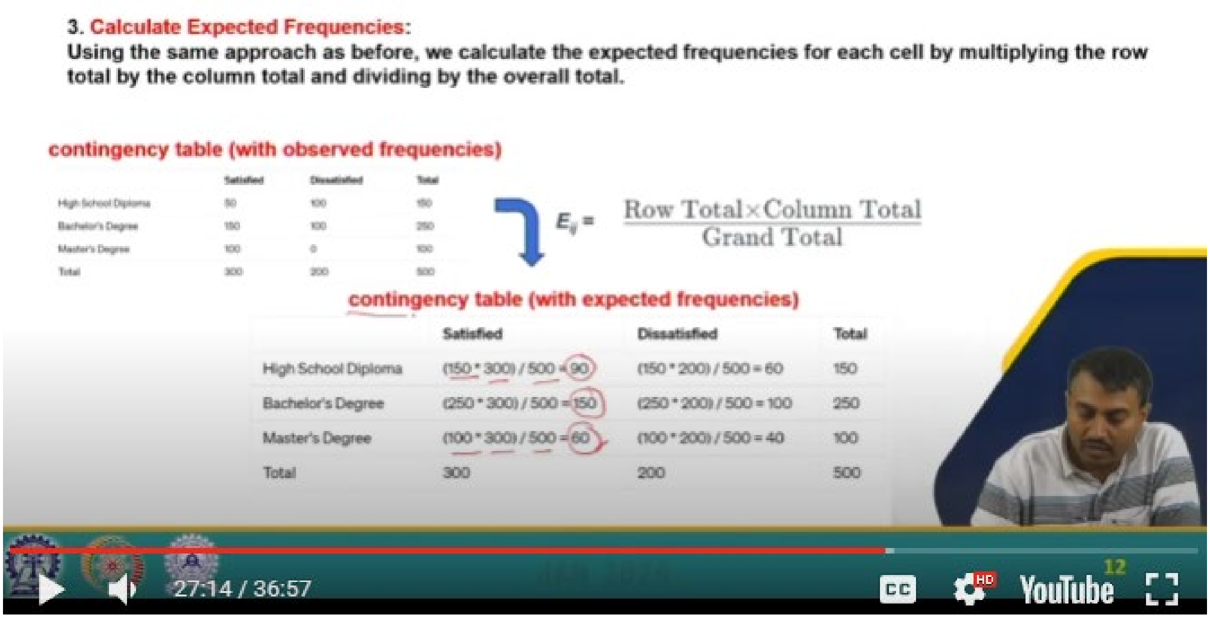
1. *H*0: There is no significant association between educational attainment and job satisfaction.
2. *H*1: There is a significant association between educational attainment and job satisfaction.

2.Choose Significance Level:

1. $\alpha = 0.05$

This is the educational level, and the human resource department has collected data about job satisfaction. Amongst these educational levels, educational level of attainment is okay, high school, bachelor level, and master degree. So they have categorized it into two categories: satisfied and dissatisfied, they have collected the data, and the total data is 500 okay. So, we have to find out that the null hypothesis is that there is no significant association between educational attainment and job satisfaction. This is the assumption that the null hypothesis and the alternative hypothesis is the starting point that there is a significant association between educational attainment and job satisfaction.

We also had to test at a 0.05 alpha significance level. So, how do you calculate? So, the calculation technique is to divide the row total into a column total. So, the row total is 150, and the column total is 300. So 150 into 300 divided by the total 500.



3. Calculate Expected Frequencies:
Using the same approach as before, we calculate the expected frequencies for each cell by multiplying the row total by the column total and dividing by the overall total.

contingency table (with observed frequencies)

	Satisfied	Dissatisfied	Total
High School Diploma	50	100	150
Bachelor's Degree	150	100	250
Master's Degree	100	0	100
Total	300	200	500

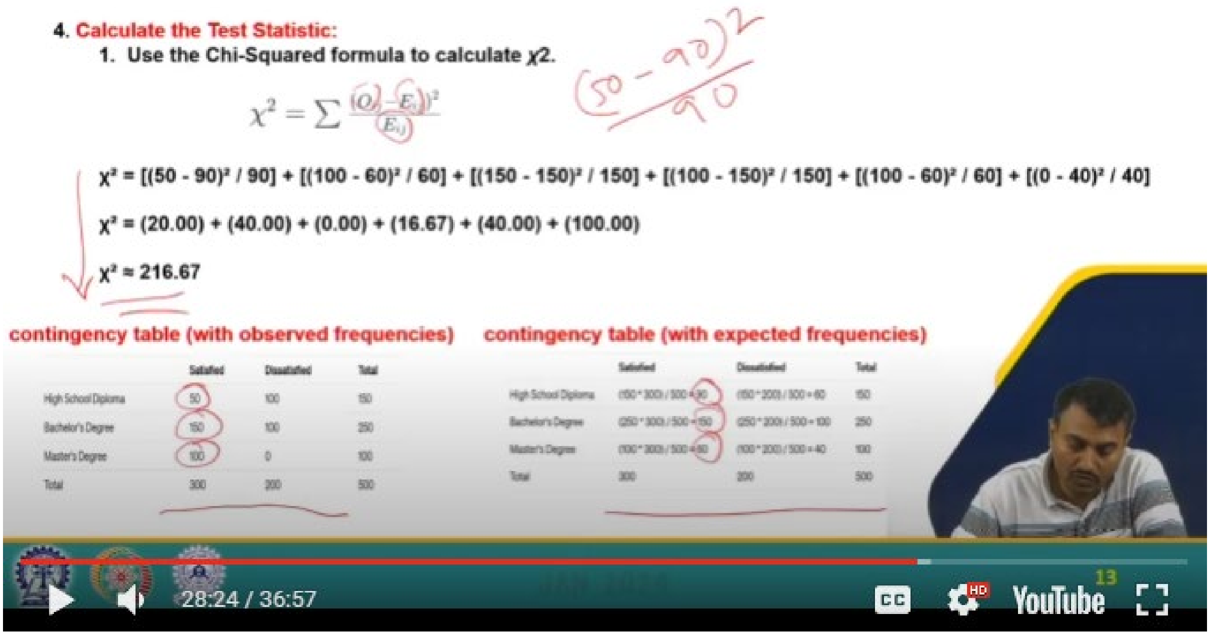
$$E_{ij} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

contingency table (with expected frequencies)

	Satisfied	Dissatisfied	Total
High School Diploma	(150 * 300) / 500 = 90	(150 * 200) / 500 = 60	150
Bachelor's Degree	(250 * 300) / 500 = 150	(250 * 200) / 500 = 100	250
Master's Degree	(100 * 300) / 500 = 60	(100 * 200) / 500 = 40	100
Total	300	200	500

So, for the high school diploma, it is 90. Bachelor similarly, the row total is 250, and the column total is 300 divided by the total of 500, 150. Master degree, this is 100, is 300, 100, 300 divided by 560, okay. Similarly, this is done. So, this is the contingency table we have just prepared now with expected frequencies. So this is the observed frequency, the first data, and the data given, and we have calculated it okay.

So this is 50, here you can see it is 90, here it is 150, here it is 150, here it is 100, here it is 60, okay. Now, the calculation of the chi-square is the test statistics. So it is 50 minus 90 divided by 90; this is the expected, the expected, and the observed. So 50 minus 90 square divided by the 90.



4. Calculate the Test Statistic:
   1. Use the Chi-Squared formula to calculate $\chi 2$.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$\frac{(50-90)^2}{90}$$

$\chi^2 = [(50 - 90)^2 / 90] + [(100 - 60)^2 / 60] + [(150 - 150)^2 / 150] + [(100 - 150)^2 / 150] + [(100 - 60)^2 / 60] + [(0 - 40)^2 / 40]$

$\chi^2 = (20.00) + (40.00) + (0.00) + (16.67) + (40.00) + (100.00)$

$\chi^2 \approx 216.67$

contingency table (with observed frequencies)

	Satisfied	Dissatisfied	Total
High School Diploma	50	100	150
Bachelor's Degree	150	100	250
Master's Degree	100	0	100
Total	300	200	500

contingency table (with expected frequencies)

	Satisfied	Dissatisfied	Total
High School Diploma	(150 * 300) / 500 = 90	(150 * 200) / 500 = 60	150
Bachelor's Degree	(250 * 300) / 500 = 150	(250 * 200) / 500 = 100	250
Master's Degree	(100 * 300) / 500 = 60	(100 * 200) / 500 = 40	100
Total	300	200	500

Similarly, for all okay. So, this has been calculated and is equal to 216.67. So the degree of freedom was 1 to 3, this is 2. So three minus 1, 2, and 2 minus 1, 1 into 2 equals 2, and this is two is 5.991. So, the degree of freedom is 2, and the significant alpha level is equal to 0.05, which is 5.99. The value we have calculated, 216.67, is much greater than the critical value of 5.99, so we reject the null hypothesis. So, we conclude that there is a significant association between the employee's educational attainment and their reported level of job satisfaction in the company. So, this example demonstrates the application of the chi-square independent test to analyze the relationship between the two categorical values. Another variant we use is the chi-square goodness of fit test. So, this test determines whether an observed distribution or frequency matches the expected frequency for a categorical value. For example, some product companies manufacture and claim something.

So there should be some frequency, the expected product that they are advertising, and if you randomly sample whatever you have found out, you can match or test the difference between these observed and the predicted frequency using the chi-square goodness of fit test. So, this test is often employed when you want to compare the observed frequency to the expected frequency for one categorical variable, one categorical variable. So, rather than comparing two categorical variables in a chi-square test for independence. So, one categorical variable, one example that we will tell you, is the color of the eye. So, we collected data from about 200 people and found this data.

Blue eye is 50, brown eye is 100, green eye is 30, and gray eye is 20. This is the data we found in the sample. Now, we want to test whether these observed frequencies match the expected distribution of eye color in the population, hypothesizing that the eye colors are distributed equally. Blue is 25%, brown is 50%, 15%, green, and gray is 10%. So, in formulating the hypothesis, the observed eye color frequency matched the expected distribution.

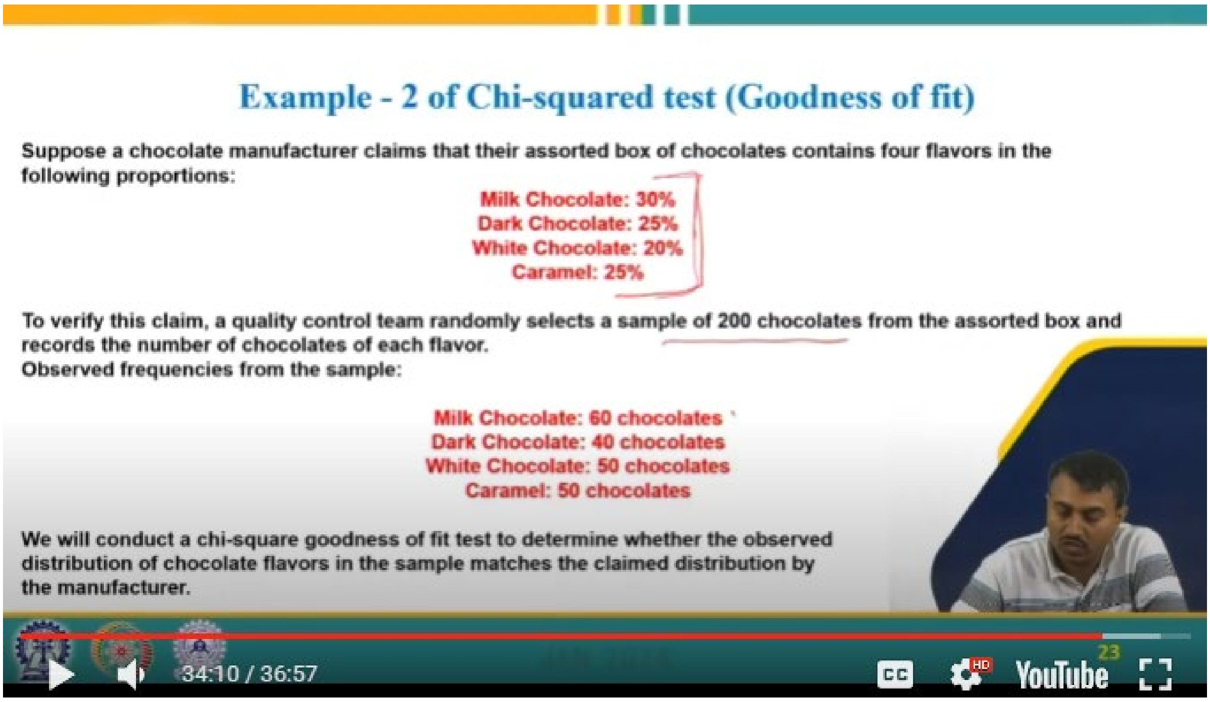
This is the null hypothesis. The alternative observed eye color frequency does not match the expected frequency, and the significance level is 0.05; alpha equals 0.05. So, since we expect the eye color to be distributed equally, we calculate the predicted frequency as follows. So 25% into 200, that is 50, 50% is 200, 100, and so on.

So, this is the expected data. Whatever we have found out from the data we collected, all are similar, so it is 0. The calculation technique remains the same. This is the observed frequency. The observed frequency is 50, the expected frequency is 50, and the expected frequency is 50.

So, the calculation mathematical framework remains the same. So, this is the chi-square critical value. So we have the four categories of blue eye color. So the degree of freedom is 4-1, 3.

So, the significance level at 0.05 degree 3 is 7.815, less than the critical value of 7.81, which we have found to be 0. So, we need to reject the null hypothesis. So, we conclude that insufficient evidence suggests that the observed eye color frequencies differ significantly from the expected distribution. So thus, we accept the hypothesis that eye colors are distributed equally in the population.

So, in this case, the chi-square goodness of fit test indicates that the observed frequency matches the expected frequency, supporting the hypothesis of an equal distribution of eye color in the population. Suppose we are given another example, and suppose the chocolate manufacturer claims that these are the combinations of the chocolate they pack on a box of chocolate. Milk chocolate 30%, dark chocolate 25%, white chocolate 20%, caramel 25%, total 55 and 45..100. To verify this claim, a quality control team randomly selected a sample of 200 chocolate from the assorted box, and we found the milk chocolate is 60, dark chocolate is 40, white chocolate is 50, caramel is 50. So, we will conduct a chi-square goodness of fit test to determine whether the observed distribution of the chocolate flavors in the sample matches the manufacturer's claim for distribution.



Example - 2 of Chi-squared test (Goodness of fit)

Suppose a chocolate manufacturer claims that their assorted box of chocolates contains four flavors in the following proportions:

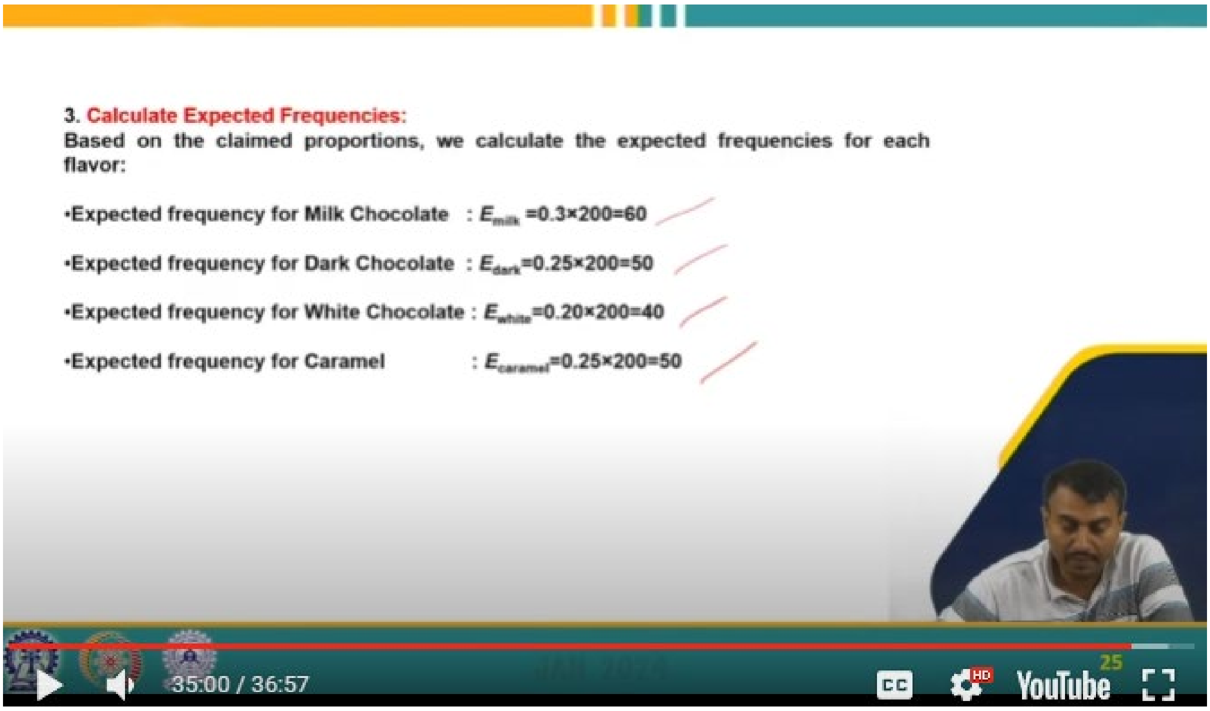
Milk Chocolate: 30%
Dark Chocolate: 25%
White Chocolate: 20%
Caramel: 25%

To verify this claim, a quality control team randomly selects a sample of 200 chocolates from the assorted box and records the number of chocolates of each flavor.
Observed frequencies from the sample:

Milk Chocolate: 60 chocolates
Dark Chocolate: 40 chocolates
White Chocolate: 50 chocolates
Caramel: 50 chocolates

We will conduct a chi-square goodness of fit test to determine whether the observed distribution of chocolate flavors in the sample matches the claimed distribution by the manufacturer.

So, the null hypothesis is that the observed distribution of chocolate flavors matches the claimed distribution of the manufacturer, and the observed distribution of the chocolate flavor does not match the claimed distribution by the manufacturer at alpha, which is equal to 0.05. So based on that, we calculate 30%, 25%, 20%, and 25%; these are the data we found, the frequency, and again 60, 60, then 40, 50, 40, 50 and 50, 40, 50, 50.



3. Calculate Expected Frequencies:
Based on the claimed proportions, we calculate the expected frequencies for each flavor:

- Expected frequency for Milk Chocolate : E_{milk} =0.3×200=60
- Expected frequency for Dark Chocolate : E_{dark}=0.25×200=50
- Expected frequency for White Chocolate : E_{white}=0.20×200=40
- Expected frequency for Caramel : $E_{caramel}$=0.25×200=50

So, these are found to be the data chi-square, which is equal to 4.5. And how many numbers of variants? 1, 2, 3, 4.

So, the degree of freedom is 4-1. So 4-1 is 3. This is the value 7.815. So, our value is 4.5, and the table value we found from the table is 7.815. So, we failed to reject the null hypothesis. So, we conclude that there is not enough evidence to suggest that the observed distribution of chocolate flavor significantly differs from the manufacturer's claim for distribution.

So thus, we accept the manufacturer's claim regarding the distribution of flavor in the assorted box of chocolate. So, in this example, the chi-square goodness of fit indicates the observed frequency aligned with the expected frequency based on the manufacturer's claim. So these are the references, and let me conclude in a few sentences. So, we have introduced the chi-square test for the hypothesis testing and the two variants. One is the test for independence, and the other is the goodness of fit test, along with examples. We have elaborated on the assumptions required for this test and provided real-life examples of its applications. Thank you.