SWAYAM NPTEL COURSE ON

MINE AUTOMATION AND DATA ANALYTICS

By

Prof. Radhakanta Koner

Department of Mining Engineering

Indian Institute of Technology (Indian School of Mines) Dhanbad

Module 9 : Hypothesis Testing

Lecture 22B : Chi-Squared Test

CONCEPTS COVERED

1) Introduction to chi-squared test
2) Assumptions of chi-squared test
3) Introduction to chi-squared test for independence
- Solved Example of the above chi-squared test - 1
4) Introduction to chi-squared goodness of fit test
- Solved Example of the above chi-squared test -1
- Solved Example of the above chi-squared test -2

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.

Chi-squared test

First, we will discuss the Chi-Squared test for independence.

We use the chi-squared test for independence to determine whether there is a significant association between two categorical variables. This test is particularly useful when we want to examine the relationship between two variables to see if they are related or independent of each other.

1. Formulate the Hypotheses:
Null Hypothesis (H0): There is no significant association between the two categorical variables.
Alternative Hypothesis (H1): There is a significant association between the two categorical variables.

2. Choose the Significance Level (α):
Common choices are 0.05, 0.01, or 0.10.

3. Collect and Organize Data:
Organize the data into a contingency table, which displays the frequencies of each combination of the two categorical variables.

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.

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.

Chi-squared Critical Values

Degrees of freedom (*df*)	Significance level (α)							
	.99	.975	.95	.9	.1	.05	.025	.01
1	-------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892

Assumptions of Chi-squared test

The chi-squared test relies on several assumptions to ensure its validity and reliability. These assumptions include:

Independence of Observations: The observations or data points used in the chi-squared test should be independent of each other. In other words, the occurrence or value of one observation should not influence the occurrence or value of another observation.

Random Sampling: The data should be collected through a random sampling process to ensure that the sample is representative of the population from which it is drawn. This helps to minimize bias and ensure that the results are generalizable.

Categorical Data: The chi-squared test is designed for categorical data, meaning that the variables being analyzed are divided into distinct categories or groups. It is not appropriate for numerical data or continuous variables.

Expected Frequencies: The expected frequencies in each cell of the contingency table (or cross-tabulation) should be greater than or equal to 5. When expected frequencies are too small, the chi-squared test may produce unreliable results. In such cases, alternative tests like Fisher's exact test may be more appropriate.

Assumptions of Chi-squared test

Large Sample Size: While there is no strict requirement for sample size, the chi-squared test tends to perform better with larger sample sizes. As the sample size increases, the distribution of the test statistic approaches a chi-squared distribution, making the test results more reliable.

Mutual Exclusivity and Exhaustiveness: The categories within each variable should be mutually exclusive (i.e., each observation should belong to only one category) and exhaustive (i.e., all possible categories should be represented in the analysis). This ensures that every observation is accounted for and avoids ambiguity in interpretation.

No Cell Count Should Be Zero: None of the cells in the contingency table should have an observed or expected frequency of zero. A zero count in any cell can lead to undefined results or computational issues when calculating the test statistic.

Adhering to these assumptions helps to ensure that the chi-squared test provides accurate and meaningful results when assessing the association between categorical variables. If any of these assumptions are violated, the reliability and validity of the test may be compromised, and alternative methods or adjustments may be necessary.

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. $\alpha = 0.05$

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

4. **Calculate the Test Statistic:**

1. Use the Chi-Squared formula to calculate $\chi 2$.

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

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

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

$\chi^2 \approx 216.67$

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

Chi-squared Critical Values

Degrees of freedom (*df*)	Significance level (α) .99	.975	.95	.9	.1	.05	.025	.01
1	-------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892

5. **Determine Degrees of Freedom:**
Degrees of freedom (df) = (number of rows - 1) * (number of columns - 1) = (3 - 1) * (2 - 1) = 2

6. **Find Critical Value or P-value:**
At a significance level of $\alpha = 0.05$ and 2 degrees of freedom, the critical value from the chi-squared distribution table is approximately 5.99.

7. **Compare χ^2 to Critical Value:**
Since χ^2 (216.67) is much greater than the critical value (5.99), we reject the null hypothesis.

8. **Interpretation:**
We conclude that there is a significant association between employees' educational attainment and their reported level of job satisfaction in the company.

This example demonstrates the application of the chi-squared independence test to analyze the relationship between two categorical variables (educational attainment and job satisfaction) and interpret the results accordingly.

Chi-squared test

Chi-Square Goodness of Fit test (Second main type of Chi-Squared test)

The chi-square goodness of fit test is used to determine whether an observed frequency distribution matches an expected frequency distribution for a categorical variable.

This test is often employed when you want to compare observed frequencies to expected frequencies for one categorical variable, rather than comparing two categorical variables as in the chi-square test for independence.

Example - 1 of Chi-squared test (Goodness of fit)

Suppose we have observed eye color frequencies for a sample of 200 individuals:

Blue eyes: 50 individuals
Brown eyes: 100 individuals
Green eyes: 30 individuals
Gray eyes: 20 individuals

We want to test whether these observed frequencies match the expected distribution of eye colors in the population, hypothesizing that eye colors are distributed equally:

Blue eyes: 25%
Brown eyes: 50%
Green eyes: 15%
Gray eyes: 10%

1.**Formulate Hypotheses:**

*H*0: The observed eye color frequencies match the expected distribution.

*H*1: The observed eye color frequencies do not match the expected distribution.

2.**Choose Significance Level:**

1. $\alpha = 0.05$

3. **Calculate Expected Frequencies:**

Since we expect the eye colors to be distributed equally, we can calculate the expected frequencies as follows:

- Expected frequency for blue eyes: $E_{blue} = 0.25 \times 200 = 50$
- Expected frequency for brown eyes: $E_{brown} = 0.50 \times 200 = 100$
- Expected frequency for green eyes: $E_{green} = 0.15 \times 200 = 30$
- Expected frequency for gray eyes: $E_{gray} = 0.10 \times 200 = 20$

4. **Calculate the Test Statistic:**

1. Use the Chi-Squared formula to calculate $\chi 2$.

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

$\chi^2 = [(50 - 50)^2 / 50] + [(100 - 100)^2 / 100] + [(30 - 30)^2 / 30] + [(20 - 20)^2 / 20]$

$\chi^2 = 0 + 0 + 0 + 0$

$\chi^2 \approx 0$

JAN 2024

Chi-squared Critical Values

Degrees of freedom (*df*)	Significance level (α) .99	.975	.95	.9	.1	.05	.025	.01
1	-------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892

5. **Determine Degrees of Freedom:**
Since we have 4 categories of eye color, the degrees of freedom (df) = 4 – 1 = 3

6. **Find Critical Value or P-value:**
At a significance level of 0.05 and 3 degrees of freedom, the critical value from the chi-square distribution table is approximately 7.815.

7. **Compare χ^2 to Critical Value:**
Since χ^2 (= 0) is less than the critical value of 7.815, we fail to reject the null hypothesis.

8. **Interpretation:**
We conclude that there is not enough evidence to suggest that the observed eye color frequencies differ significantly from the expected distribution. Thus, we accept the hypothesis that eye colors are distributed equally in the population.

In this case, the chi-square goodness of fit test indicates that the observed frequencies match the expected frequencies, supporting the hypothesis of an equal distribution of eye colors in the population.

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.

1. **Formulate Hypotheses:**

    Null Hypothesis (H0): The observed distribution of chocolate flavors matches the claimed distribution by the manufacturer.

    Alternative Hypothesis (H1): The observed distribution of chocolate flavors does not match the claimed distribution by the manufacturer.

2. **Choose Significance Level:**
    1. $\alpha = 0.05$

3. **Calculate Expected Frequencies:**

Based on the claimed proportions, we calculate the expected frequencies for each flavor:

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

4. **Calculate the Test Statistic:**

1. Use the Chi-Squared formula to calculate $\chi 2$.

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

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

$\chi^2 = 0 + 2 + 2.5 + 0$

$\chi^2 \approx 4.5$

Chi-squared Critical Values

Significance level (α)

Degrees of freedom (*df*)	.99	.975	.95	.9	.1	.05	.025	.01
1	-------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892

5. **Determine Degrees of Freedom:**
Since we have 4 categories of chocolate flavors, the degrees of freedom (df) is $4 - 1 = 3$

6. **Find Critical Value or P-value:**
At a significance level of 0.05 and 3 degrees of freedom, the critical value from the chi-square distribution table is approximately 7.815.

7. **Compare χ^2 to Critical Value:**
Since χ^2 (= 4.5) is less than the critical value of 7.815, we fail to reject the null hypothesis.

8. **Interpretation:**
We conclude that there is not enough evidence to suggest that the observed distribution of chocolate flavors significantly differs from the claimed distribution by the manufacturer. Thus, we accept the manufacturer's claim regarding the distribution of flavors in the assorted box of chocolates.

In this example, the chi-square goodness of fit test indicates that the observed frequencies align with the expected frequencies based on the manufacturer's claim.

REFERENCES

- Introduction to Probability and Statistics for Engineers and Scientists, Sixth Edition, Sheldon M. Ross
- Statistical Methods Combined Edition (Volume I& II), N G Das

CONCLUSION

- We discussed the chi-squared test along with computation steps.
- We discussed the assumptions of chi-squared test
- We discussed the chi-squared test for independence, along with one example
- We discussed the chi-squared goodness of fit test, along with two solved examples


THANK YOU