Chi-Squared Tests for Association I Cheat Sheet

Chi-squared tests are used to test if two variables are **independent** from one another. In other words, it tests whether there is a statistically significant relationship between two variables, which are usually categorical variables.

Contingency Tables

A contingency table can be used to show the observed frequency distribution or expected frequency of two variables. Observed frequency is denoted by O_i. Expected frequency is denoted by E_i and can be calculated for each cell in a contingency table using the following formula:

$$E_i = \frac{\text{row total} \times \text{column total}}{\text{overall total}}$$

Example 1: The contingency table below shows the observed frequency distribution between gender of customer and the colour of shoes bought. Construct a contingency table showing the expected frequency.

	Male	Female
Black	39	28
Blue	9	8
Green	6	5
Red	5	6
White	34	40
Yellow	7	13

Find the row total, column total and overall total from the contingency table.

Black: $39 + 28 = 67$
Blue: $9 + 8 = 17$
Green: $6 + 5 = 11$
Red: $5 + 6 = 11$
White: $34 + 40 = 74$
Yellow: $7 + 13 = 20$

Male: $39 + 9 + 6 + 5 + 34 + 7 = 100$
Female: $28 + 8 + 5 + 6 + 40 + 13 = 100$

Overall total: $100 + 100 = 200$

Find the expected value for each cell using $E_i = \frac{\text{row total} \times \text{column total}}{\text{overall total}}$.

	Male	Female
Black	$\frac{67 \times 100}{200} = 33.5$	$\frac{67 \times 100}{200} = 33.5$
Blue	$\frac{17 \times 100}{200} = 8.5$	$\frac{17 \times 100}{200} = 8.5$
Green	$\frac{11 \times 100}{200} = 5.5$	$\frac{11 \times 100}{200} = 5.5$
Red	$\frac{11 \times 100}{200} = 5.5$	$\frac{11 \times 100}{200} = 5.5$
White	$\frac{74 \times 100}{200} = 37$	$\frac{74 \times 100}{200} = 37$
Yellow	$\frac{20 \times 100}{200} = 10$	$\frac{20 \times 100}{200} = 10$

Chi-Squared Values and Degrees of Freedom

Chi-squared value is the test statistic used for hypothesis testing in a chi-squared test. A low chi-squared value shows a high correlation between the two variables investigated. It is calculated from the observed frequency O_i and expected frequency E_i using the formula:

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

The degree of freedom, which can be written as v, is a parameter of the chi-squared distribution and can be found from the number of rows and columns of the contingency table. For a contingency table with n rows and m columns, the degree of freedom is given by:

$$v = (n-1)(m-1)$$

The critical value at a given significance level is dependent on the degrees of freedom and is given in the formula book. If the chi-squared value calculated is greater than the critical value, there is sufficient evidence to suggest the two variables investigated are dependent.

Hypothesis Testing Using Chi-Squared Tests

When using chi-squared tests for hypothesis testing, the null hypothesis always states that two variables are independent. By assuming that, expected frequencies can be calculated.

To calculate chi-squared value, it is important to ensure that the expected frequency in each cell is greater than 5, as the chi-squared distribution is an approximation which is invalid for $E_i \leq 5$. When this happens, two or more columns can be merged.

Sources of Association

A contingency table showing $\frac{(O_i-E_i)^2}{E_i}$ will reveal the sources of association if hypothesis testing shows that the variables are dependent. Cells with larger values for $\frac{(O_i-E_i)^2}{E_i}$ are likely to be the sources of association and can be interpreted in context.

Example 2: The contingency table below shows the favourite movie genre of different age groups. Test, at 5% significance level, whether age group and favourite movie genre are two independent variables. Suggests the sources of association.

	≤ 20	$21-30$	$31-40$	>40
Action	17	14	11	3
Comedy	8	11	8	7
Horror	9	22	14	5
Romance	6	6	21	1

State the null and alternative hypothesis.

H_0: Age group and favourite movie genre are independent.
H_1: Age group and favourite movie genre are dependent.

Construct a contingency table showing the expected values using $E_i = \frac{\text{row total} \times \text{column total}}{\text{overall total}}$.

	≤ 20	$21-30$	$31-40$	>40
Action	$\frac{45 \times 40}{163} = 11.043$	$\frac{45 \times 53}{163} = 14.910$	$\frac{45 \times 54}{163} = 14.908$	$\frac{45 \times 16}{163} = 4.417$
Comedy	$\frac{34 \times 40}{163} = 8.344$	$\frac{34 \times 53}{163} = 11.055$	$\frac{34 \times 54}{163} = 11.264$	$\frac{34 \times 16}{163} = 3.337$
Horror	$\frac{50 \times 40}{163} = 12.270$	$\frac{50 \times 53}{163} = 16.258$	$\frac{50 \times 54}{163} = 16.564$	$\frac{50 \times 16}{163} = 4.908$
Romance	$\frac{34 \times 40}{163} = 8.344$	$\frac{34 \times 53}{163} = 11.055$	$\frac{34 \times 54}{163} = 11.264$	$\frac{34 \times 16}{163} = 3.337$

Notice that the last column has expected values of ≤ 5, so the last two columns need to be merged.

	≤ 20	$21-30$	>30
Action	11.043	14.910	$14.908 + 4.417 = 19.325$
Comedy	8.344	11.055	$11.264 + 3.337 = 14.601$
Horror	12.270	16.258	$16.564 + 4.908 = 21.472$
Romance	8.344	11.055	$11.264 + 3.337 = 14.601$

Find v using the table with merged columns and state the critical value at 5% significance level. This is the value such that $P(X \leq CV) = 0.95$.

$v = (3-1)(4-1) = 6$
$CV = 12.592$

Construct the contingency table for $\frac{(O_i-E_i)^2}{E_i}$, keeping in mind that O_i for the last two columns should also be merged. $\sum \frac{(O_i-E_i)^2}{E_i}$ can be calculated straightaway for hypothesis testing, but the table is needed to study the sources of association.

	≤ 20	$21-30$	>30
Action	$\frac{(17-11.043)^2}{11.043} = 3.21$	$\frac{(14-14.910)^2}{14.910} = 0.06$	$\frac{(14-19.325)^2}{19.325} = 1.47$
Comedy	$\frac{(8-8.344)^2}{8.344} = 0.01$	$\frac{(11-11.055)^2}{11.055} = 0.00$	$\frac{(15-14.601)^2}{14.601} = 0.01$
Horror	$\frac{(9-12.270)^2}{12.270} = 0.87$	$\frac{(22-16.258)^2}{16.258} = 2.03$	$\frac{(19-21.472)^2}{21.472} = 0.28$
Romance	$\frac{(6-8.344)^2}{8.344} = 0.66$	$\frac{(6-11.055)^2}{11.055} = 2.31$	$\frac{(22-14.601)^2}{14.601} = 3.75$

$$\sum \frac{(O_i-E_i)^2}{E_i} = 3.21 + 0.01 + 0.87 + 0.66 + 0.06 + 0.00 + 2.03 + 2.31 + 1.47 + 0.01 + 0.28 + 3.75$$
$$= 14.66$$

Compare χ^2 with the critical value and state your conclusion.

$\chi^2 = 14.66 > 12.592$
$\therefore$ Reject H_0. There is sufficient evidence to suggest that age group and favourite movie genre are dependent.

Look at the contingency table for $\frac{(O_i-E_i)^2}{E_i}$. Large values indicate sources of association. Refer back to the expected and observed frequency of these cells.

Action movies seemed to be more popular in the age group ≤ 20, while romance movies are more popular in the age group > 30.