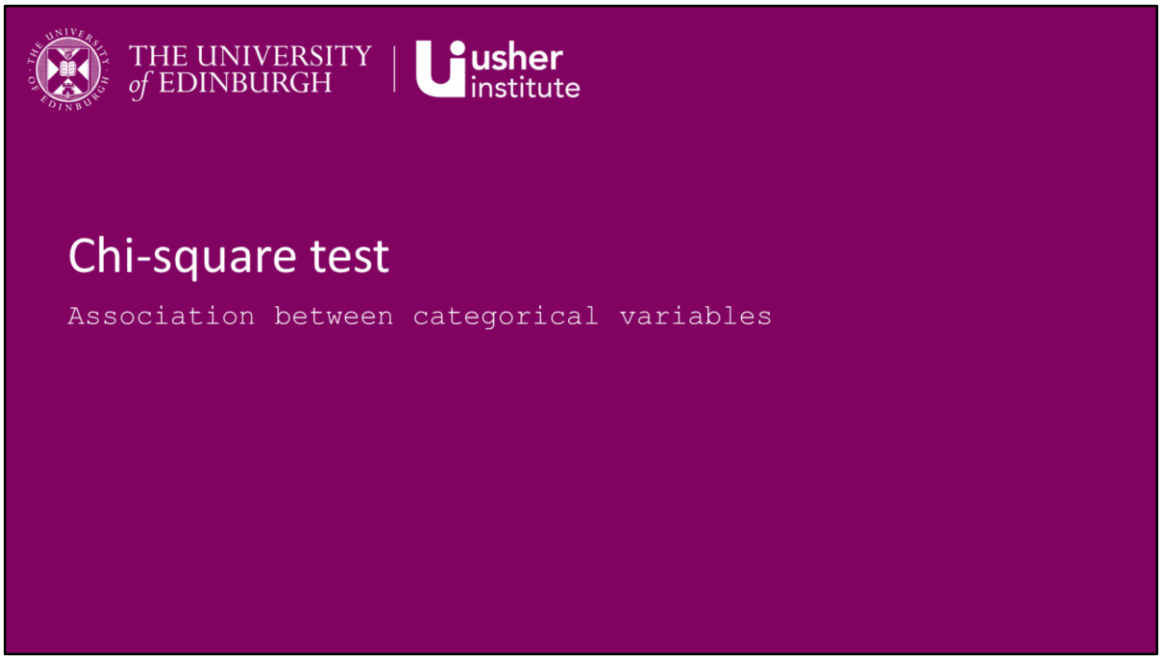


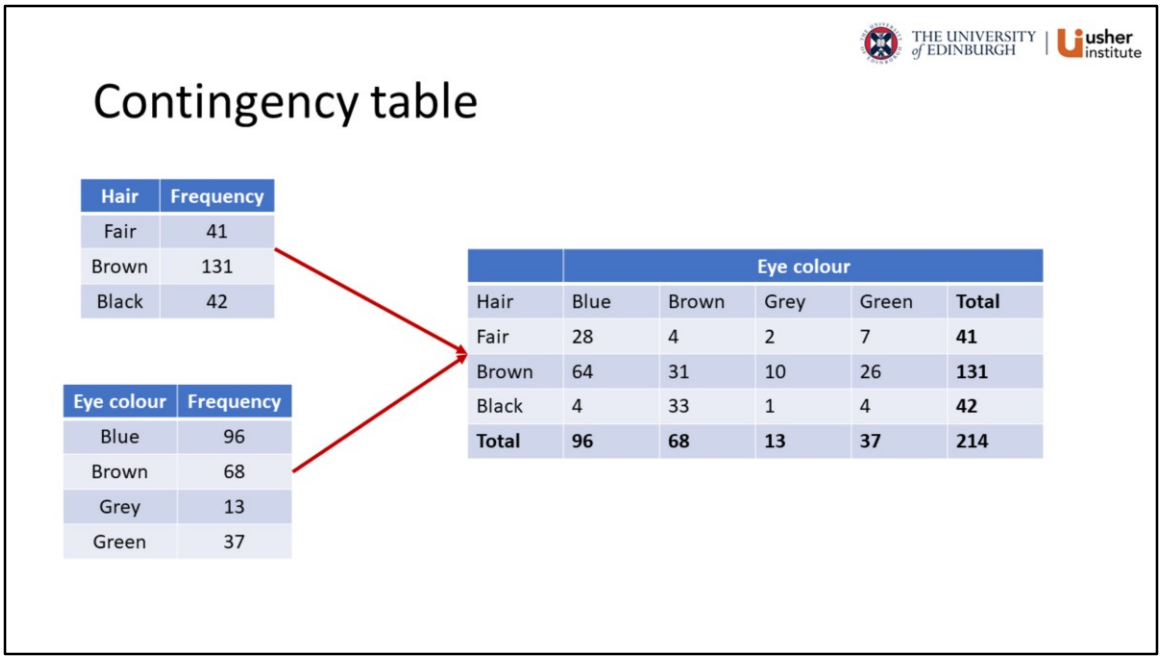
In this video, we are going to introduce the chi-square test, a statistical test to determine the association between two categorical variables. We will see what it is used for and the reasoning behind it.

Example

Image by Sofie Zbořilová from Pixabay

Image by XxX XXX from Pixabay

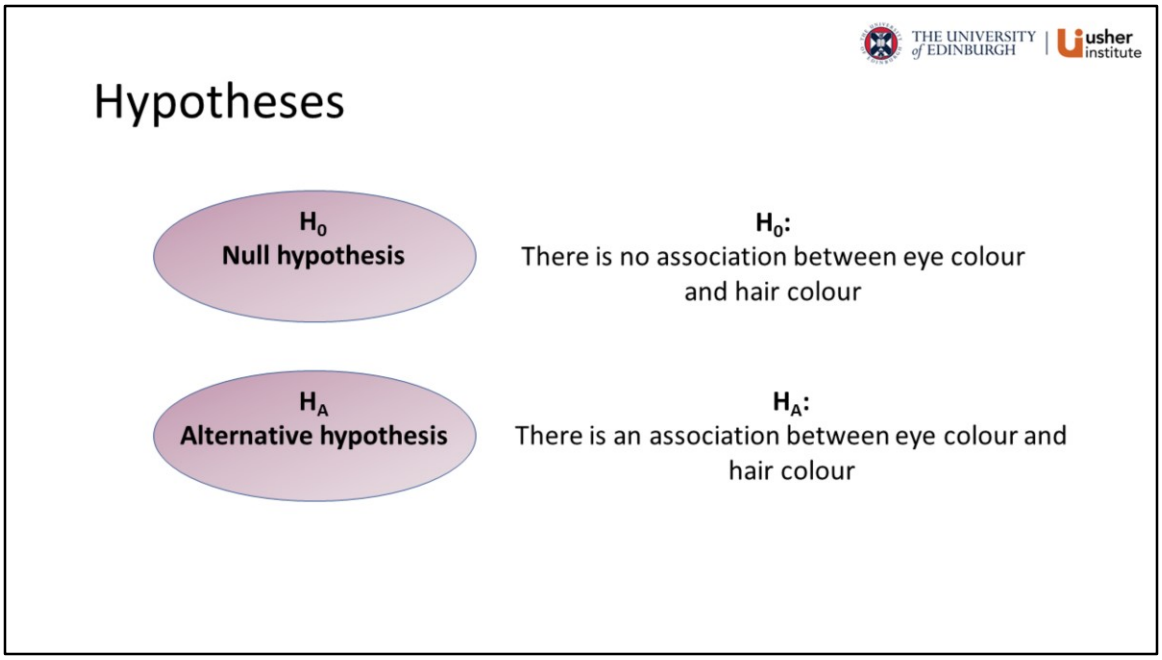
The Chi-square test of independence is a hypothesis test used to measure the association between categorical variables.
This is a fairly intuitive test, which is best demonstrated by working through an example.
214 students were asked to record their eye colour and natural hair colour in a questionnaire. We will use this example over the next few slides to determine the association between hair colour and eye colour, which are both categorical variables.



As we will be looking at two categorical variables at the same time, we first need to understand how to present this information. A frequency table shows the number of observations for each category of a single categorical variable. We can combine this information for two categorical variables by using a table where each cell shows the number of observations for a specific combination of categories from each categorical variable. This type of frequency table is called a cross-tabulation or a contingency table because it shows the frequency of each category in one variable, contingent upon the specific level of the other variable.

The two tables on the left represent the frequency distribution of hair and eye colour, respectively, for our sample of students. The larger table on the right represents the contingency table of these two variables, with row and column totals on the last column and last row, respectively.

By simply looking at the data, can you see any associations between eye colour and hair colour?



The null hypothesis for the chi-square test in this case would be that there is no association between eye colour and hair colour. The alternative hypothesis would then be that there is an association between eye colour and hair colour.

Observed *vs* Expected

	Eye colour				
Hair	Blue	Brown	Grey	Green	**Total**
Fair	28 (18.4)	4 (13.0)	2 (2.5)	7 (7.1)	**41**
Brown	64 (58.8)	31 (41.6)	10 (8.0)	26 (22.6)	**131**
Black	4 (18.8)	33 (13.4)	1 (2.5)	4 (7.3)	**42**
Total	**96**	**68**	**13**	**37**	**214**

$$expected = \frac{row\ total\ \times column\ total}{grand\ total}$$

We begin by assuming that the row and column totals are fixed. So we would assume that there are 96 students with blue eyes, 41 students with fair hair, etc. Assuming the null hypothesis of no association to be true, we would expect that the 214 students would be randomly allocated in the different cells of the table, leaving the totals unchanged. Therefore, the distribution of the students into the cells of the table would follow the proportions of the population. For example, we would expect the 96 students with blue eyes to have hair colours in the ratio 41:131:42, the same ratio as the overall sample.

Using this reasoning, we can calculate the expected number of students in each cell. So for the students with blue eyes, we would calculate:
(41/214) x 96 students to have fair hair = 18.4
(131/214) x 96 students to have brown hair = 58.8
(42/214) x 96 students to have black hair = 18.8
These expected figures have been added to the table and are shown in red and in brackets below the observed numbers of students.
The general formula to calculate expected values is to multiply the row total by the column total, and then divide the result by the total number of observations.

Chi-square statistic

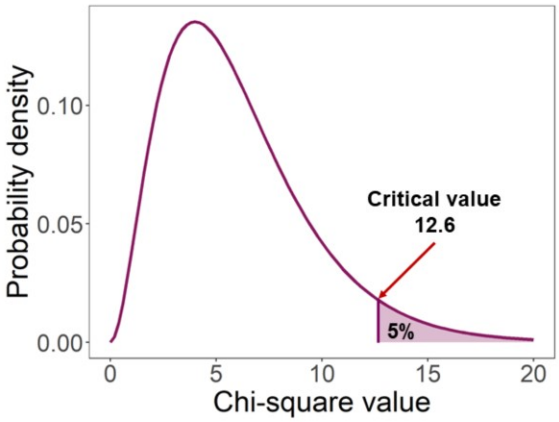
	Eye colour				
Hair	Blue	Brown	Grey	Green	**Total**
Fair	28 (18.4)	4 (13.0)	2 (2.5)	7 (7.1)	**41**
Brown	64 (58.8)	31 (41.6)	10 (8.0)	26 (22.6)	**131**
Black	4 (18.8)	33 (13.4)	1 (2.5)	4 (7.3)	**42**
Total	**96**	**68**	**13**	**37**	**214**

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected} = 58.2$$

The differences between the observed and expected numbers in each cell indicate how far the observed numbers differ from what you would expect if there was no association. The larger the difference, the more likely it is that there is an association. The test statistic, which we call chi-square, is based on these differences. Each difference between the observed and expected values is calculated and squared, and then divided by the expected value. Taking the square of the differences prevents the negative and positive differences from cancelling each other out. Finally, all these differences are added together to calculate the chi-square statistic. Here, the chi-square value is 58.2.

p-value

$df = (number\ of\ rows - 1) \times (number\ of\ columns - 1)$
$df = 2 \times 3 = 6$



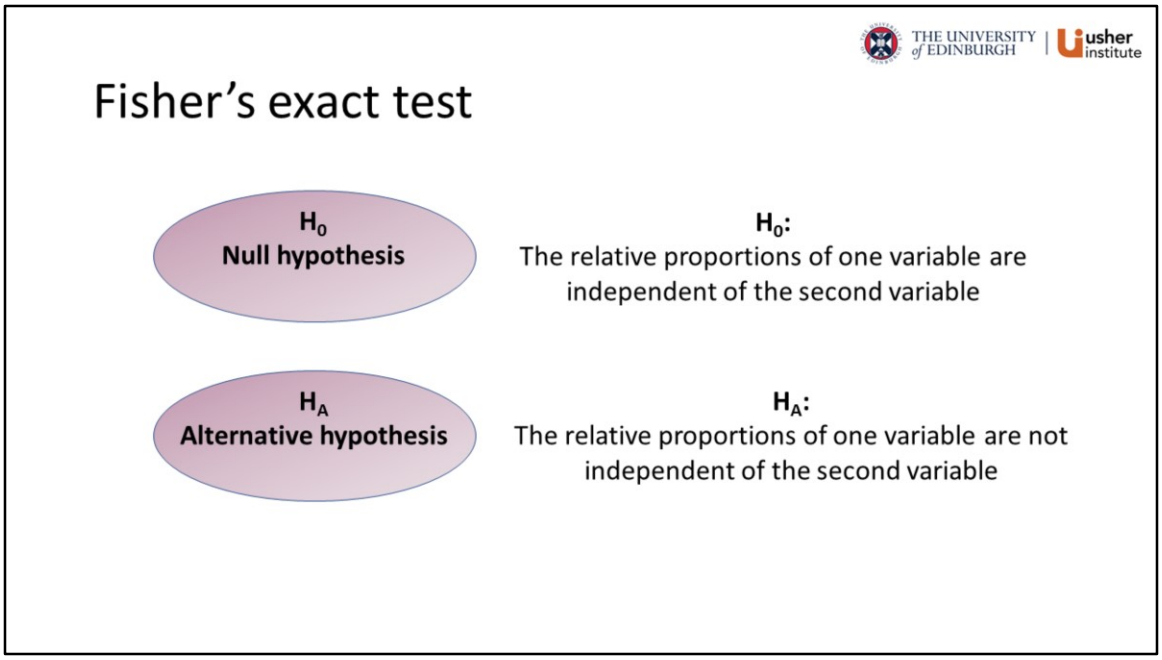
$\chi^2 = 58.2$
$df = 6, p = 8.6 \times 10^{-11}$

If H_0 is true, the chi-square statistic follows a theoretical probability distribution called the chi-square distribution, which depends on the number of degrees of freedom.
The number of degrees of freedom is equal to the number of rows minus 1 times the number of columns minus 1. For our example, we have 3 rows and 4 columns, therefore there are 6 degrees of freedom.
The plot represents the chi-square distribution with 6 degrees of freedom, and the pink area represents the area for which the probability is 5%. The critical chi-square value for 6 degrees of freedom at the 5% significance level is 12.6, therefore the probability of getting a chi-square statistic greater than 12.6 is 5%.
Since our chi-square statistic is equal to 58.2, which is much greater than 12.6, we can reject the null hypothesis at the 5% significance level, and conclude that there is a statistically significant association between hair colour and eye colour.
Statistical software can give you the exact p-value, which is very small, at 8.6×10^{-11} and much lower than our 5% significance level.

Assumptions

- Observations independent
- Expected numbers not too small
  - Less than 20% of cells have expected values <5
  - No cell is empty
  - Can use Fisher's exact test if expected numbers are small

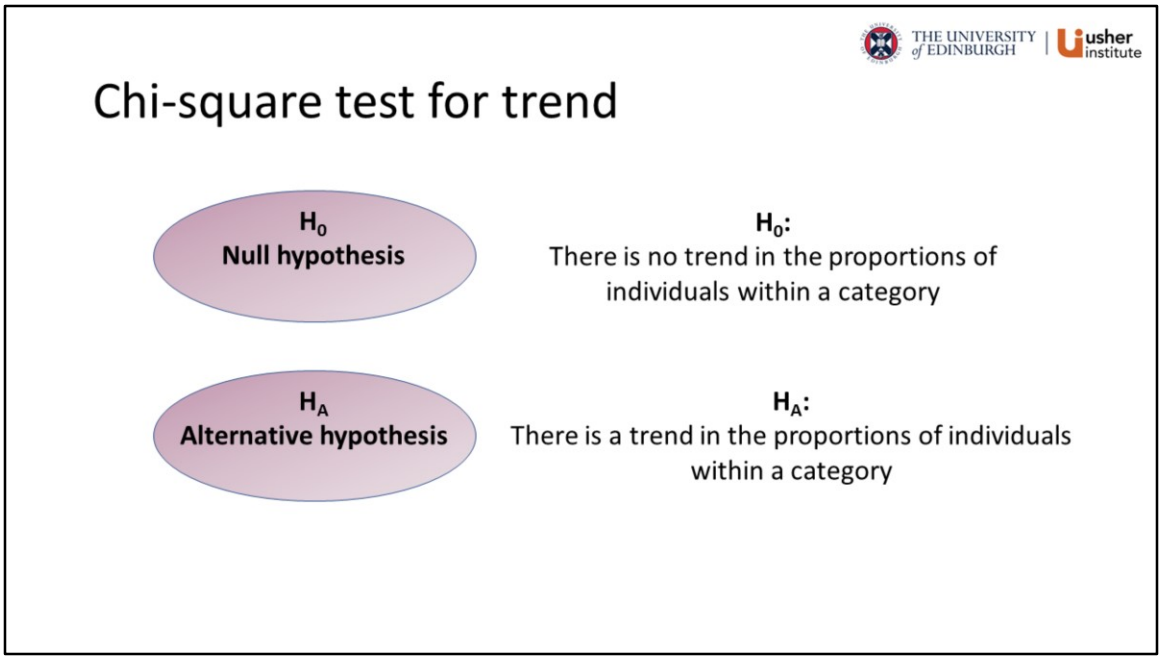
The chi-square test makes two assumptions that need to be met. The observations should be independent from one another, and the data shouldn't be too sparse across the cells. An approximate guide is that no more than 20% of the cells should contain less than 5 expected observations and no cells should be empty. One way to increase the numbers of observations in cells is to group some categories, if appropriate.
If the numbers are still too small for a $\chi 2$ test, there is another test, called the Fisher's exact test, which can be used.



Fisher's exact test is a non-parametric test used to determine whether or not there is a significant association between two categorical variables. Although in practice it is employed when sample sizes are small, it is valid for all sample sizes. It is typically used as an alternative to the chi-square test when one or more of the cell counts in a contingency table is less than 5.

Unlike most statistical tests, Fisher's exact test does not use a mathematical function that estimates the probability of a value of a test statistic. Instead, it uses the hypergeometric distribution to calculate the probability of getting the observed data, under the null hypothesis of independence.

Similarly to the chi-square test, the null hypothesis is that the relative proportions of one variable are independent of the second variable.



It is also possible to use a slight variation on the $\chi 2$ test if one of the variables is ordinal, with a natural order. This is called the $\chi 2$ test for trend. The null hypothesis for that test is that there is no trend in the proportions of individuals within a given category in the population, and the alternative hypothesis is that there is a trend. The chi-square statistic is calculated in a different way and follows a chi-square distribution with one degree of freedom under the null hypothesis.
The chi-square test for trend has more power than the classic chi-square test when the suspected trend is correct, but the ability to detect unsuspected trends is sacrificed. An example of ordinal variable that could be analysed in this way would be weight categories in relation to a binary variable such as wheezing after exercise.

Summary

- **Contingency table**
  - Combined frequency table for two variables
- **Chi-square test**
  - H_0: no association
  - Difference between observed and expected
  - Chi-square statistic
- **Assumptions**
  - Independent observations
  - Sufficient numbers in each cell

In this video, we have seen how a contingency table can summarise the frequency distributions of two categorical variables. We have seen that the null hypothesis for the chi-square test is that there is no association between two categorical variables. The chi-square test statistic is based on the differences between observed and expected counts and follows the chi-square distribution. Finally we have seen that observations should be independent and there should be sufficient numbers of observations to perform a chi-square test.