

Pearson's Chi-square test of independence of two categorical variables

Say we want to find out if drinking preferences are related to nationality, i.e. are not independent of nationality. We take a random sample of people from all three countries of the UK and record their drinking preferences.

There seem to be differences in drinking preference between the nationalities in our sample. For the English, all drinks are about equally popular except spirits. For the Scottish, spirits and beer are most popular and for the Welsh beer and soft drinks are very popular but spirits and wine are not. But is this strong evidence for a relationship in the entire population of the UK? We have only taken a small sample, so the differences could be due to random chance.

Table 1: Sample data of drinking preferences in UK countries

Drinking preference	Nationality			Row totals
	English	Scottish	Welsh	
Beer	35	32	45	112
Soft drinks	24	16	37	77
Spirits	12	34	6	52
Wine	29	18	12	59
Column totals	100	100	100	

What would the table look like if there was really no relationship between nationality and drinking preference? The proportion of our sample in each drinking preference category would be the same for each nationality across the entire UK, i.e. in the same proportions as the totals shown in the 'Row totals' column. Table 2, below, shows these 'expected' (*Exp*) values. It also shows, in brackets, the differences between the 'observed' (*Obs*) values from our sample (as shown in the table above) and the expected values.

Table 2: Expected values and calculated differences

Drinking preference	Nationality		
	English	Scottish	Welsh
Beer	37 (-2)	37 (-5)	37 (+8)
Soft drinks	26 (-2)	26 (-10)	26 (+11)
Spirits	17 (-5)	17 (+17)	17 (-11)
Wine	20 (+9)	20 (-2)	20 (-8)
Column totals	100	100	100

To determine whether these differences are likely to be due purely to chance, we calculate the chi-square test statistic. This statistic is calculated as follows:

$$\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$$

Working this out using the table is simple:

$$\chi^2 = (-2)^2/37 + (-5)^2/37 + (+8)^2/37 + (-2)^2/26 + \dots \text{ etc.} \approx 43.9$$

We then refer to the Chi-Square distribution to determine the likelihood of this value occurring if there was really no relationship in the population from which the sample came.

We first need to determine the degrees of freedom (Df) of the distribution, which is calculated as follows:

$$Df = (\text{Number of categories in first variable} - 1) \times (\text{Number of categories in second variable} - 1) = 3 \times 2 = 6$$

We then calculate the p-value associated with our chi-square statistic of 43.9. The function we can use in Excel to do this is called CHIDIST:

$$\text{p-value} = \text{CHIDIST}(\chi^2, Df) = \text{CHIDIST}(43.9, 6) = 0.0000000783$$

This p-value represents the probability that we would get the observed counts in our table if there was really no relationship in the population from which the sample came. A value near to 1 indicates that it is highly likely that this could happen, a value near to zero indicates that it is very unlikely that this could happen. Our value is very small so we know that it is very unlikely that we would get the observed results by chance alone. We therefore infer that a relationship between drinking preference and nationality is very likely to exist in the population from which our sample comes. (NB: For this test to be accurate, the expected counts in each cell should be a minimum of 5.)