Statistical Analysis 5: Chi-squared (χ^2) test for 2-way tables

Research question type: Association of two variables

What kind of variables: Categorical (**nominal** or **ordinal** with few categories)

Common Applications: Questionnaire data from a survey

Example 1:

Research question: Is there an **association** between personality and colour preference?

A group of students were classified in terms of personality (introvert or extrovert) and in terms of colour preference (red, yellow, green or blue). Personality and colour preference are **categorical data**.

Table 1:

Student	Personality	Colour preference
1	Introvert	Yellow
2	Extrovert	Red
3	Extrovert	Yellow
4	Introvert	Green
5	Extrovert	Blue
etc		

Data of this type are usually summarised by counting the number of subjects in each personality/colour group and presented in the form of a table (**cross-tabulation**), sometimes called a contingency table.

The results of a **survey** of 400 students are tabulated below:

Table 2:

		Colour				
		Red	**Yellow**	**Green**	**Blue**	**Totals**
Personality	**Introvert**	20 (10%)	6 (15%)	30 (37.5%)	44 (55%)	100 (25%)
	Extrovert	180 (90%)	34 (85%)	50 (62.5%)	36 (45%)	300 (75%)
	Totals	200 (100%)	40 (100%)	80 (100%)	80 (100%)	400 (100%)

As there are different numbers of students in each group, use of percentages helps to spot any patterns in the data. Table 2 shows **column percentages** in brackets. Table 3 shows **row percentages** in brackets. [You can choose **total** percentages too, when each number is presented as a percentage of the total.] Think about how you would choose which to use.

Table 3:

		Colour				
		Red	**Yellow**	**Green**	**Blue**	**Totals**
Personality	**Introvert**	20 (20%)	6 (6%)	30 (30%)	44 (44%)	100 (100%)
	Extrovert	180 (60%)	34 (11.3%)	50 (16.7%)	36 (12%)	300 (100%)
	Totals	200 (50%)	40 (10%)	80 (20%)	80 (20%)	400 (100%)

Hypotheses:

The 'null hypothesis' might be:

H_0: Colour preference is not related to (associated with) personality

And an 'alternative hypothesis' might be:

H_1: Colour preference is related to (associated with) personality

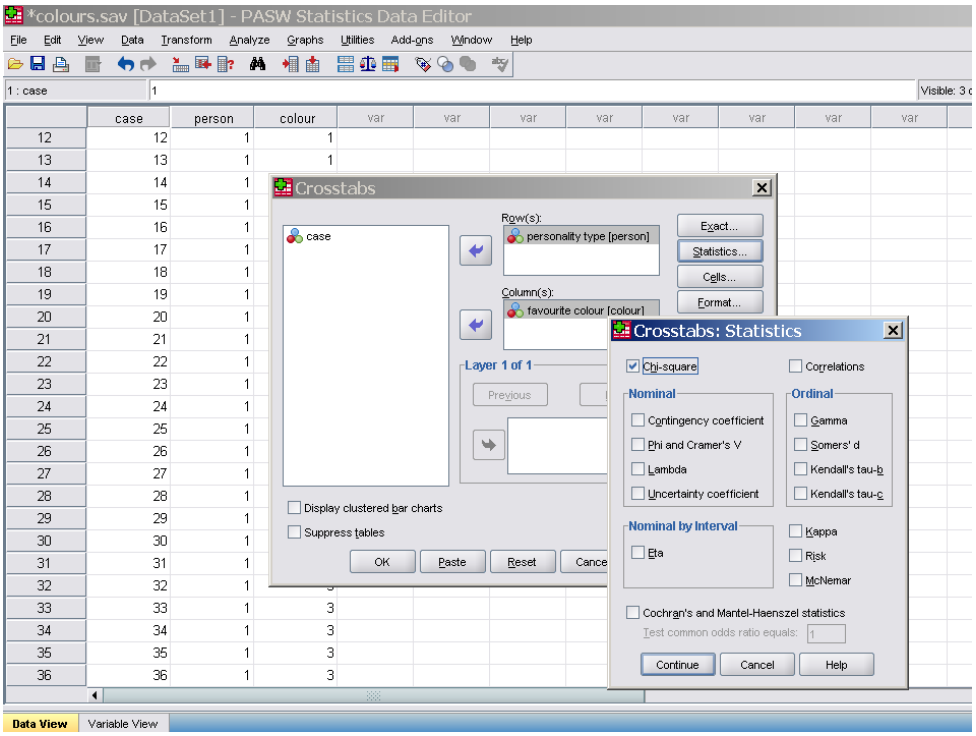
Data can be found in W:\EC\STUDENT\ MATHS SUPPORT CENTRE STATS WORKSHEETS\colours.sav

Steps in SPSS (PASW):

SPSS likes numbers, so with data entered in the format of table 1 (data from individuals), using 1 for introvert and 2 for extravert personality; and 1=red, 2=yellow, 3=green, 4=blue, choose

Analyze > Descriptive Statistics > Crosstabs

- Select one variable as the Row variable, and the other as the Column variable (see below)
- Click on the Statistics button and select Chi-square in the top LH corner and Continue.
- Click on the Cells button and select Column percentages (or Row) and Continue.
  [NB You can also ask for Expected Frequencies from the Cells button]
- Click OK
-



Output should look something like below:

personality type * favourite colour Crosstabulation

			favourite colour				
			red	yellow	green	blue	Total
personality type	introvert	Count	20	6	30	44	100
		% within favourite colour	10.0%	15.0%	37.5%	55.0%	25.0%
	extrovert	Count	180	34	50	36	300
		% within favourite colour	90.0%	85.0%	62.5%	45.0%	75.0%
Total		Count	200	40	80	80	400
		% within favourite colour	100.0%	100.0%	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	71.200[a]	3	.000
Likelihood Ratio	70.066	3	.000
Linear-by-Linear Association	69.124	1	.000
N of Valid Cases	400		

p-value

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 10.00.

Results:

From the top row of the last table, Pearson Chi-Square statistic, $\chi^2 = 71.20$, and $p < 0.001$; ie, a very small probability of the observed data under the null hypothesis of no relationship. [**NEVER write p = 0.000].** The null hypothesis is rejected, since $p < 0.05$ (in fact $p < 0.001$).

Conclusion:

Colour preference seems to be related to personality ($p < 0.001$). Go back to the tabulation (Tables 2 and 3). Note that, for instance, the most popular colour for introverts is blue (44% of them preferred blue, Table 3), whilst the most popular colour for extroverts is red (60% of them preferred, Table 3). Also, of all people preferring red, 90% of them are extroverts (Table 2), whilst of all people preferring blue, 55% of them are introverts (Table 2).

Data already grouped into a table:

Grouped data as tabulated in Table 2 can be entered in SPSS as below (with codes as above):

Personality	Favourite colour	Frequency
1	1	20
1	2	6
1	3	30
1	4	44
2	1	180
2	2	34
2	3	50
2	4	36

Before carrying-out the SPSS steps listed above, choose:

Data > Weight Cases and select **Weight cases by** and choose your frequency variable as the **Frequency Variable**.

Then repeat the steps as outlined above to get the same output as before.

Example 2:

Research question: Is there a **association** between the proportion of defectives and the machine used?

A sample of 200 components is selected from the output of a factory that uses three different machines to manufacture these components. Each component in the sample is inspected to determine whether or not it is defective. The machine that produced the component is also recorded. The results are as follows:

		Machine			
		1	**2**	**3**	**Totals**
Outcome	**Defective**	8 (13%)	6 (9%)	12 (17%)	26 (13%)
	Non-defective	54 (87%)	62 (81%)	58 (83%)	174 (87%)
	Totals	62 (100%)	68 (100%)	70 (100%)	200 (100%)

The manager wishes to determine whether or not there is a **relationship (association)** between the proportion of defectives and the machine used. The null and alternative hypotheses can be formulated as above, but in this case it is also equivalent to saying:

H_0: There are no **differences** between machines in the percentage of defectives produced
H_1: There are **differences** between machines in the percentage of defectives produced

Using the instructions outlined above for grouped data, SPSS gives Pearson Chi-Square statistic, $\chi^2 = 2.112$, and $p = 0.348$. Hence, there is no real evidence that the percentage of defectives varies from machine to machine.

Validity of Chi-squared (χ^2) tests for 2-way tables

Chi-squared tests are only valid when you have reasonable sample size.
For 2x2 tables (ie only two categories in each variable):

- If the total sample size is greater than 40, χ^2 can be used
- If the total sample size is between 20 and 40, and the smallest expected frequency is at least 5, χ^2 can be used (see note 'a.' at the bottom of SPSS output to see if this is a problem)
- Otherwise Fisher's exact test must be used (SPSS will automatically give this)

For other tables:

- χ^2 can be used if no more than 20% of the expected frequencies are less than 5 and none is less than 1 (see note 'a.' at the bottom of SPSS output to see if this is a problem)

It is possible to 'pool' or 'collapse' categories into fewer, but this must **only** be done if it is **meaningful** to group the data in this way.