statstutor

community project

encouraging academics to share statistics support resources

stcp-gilchristsamuels-8

The following resources are associated:

PersonalityColour Excel file and PersonalityColour SPSS data file

Chi-Squared Test for Two-Way Tables

Research question type: Association of two variables

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

Common Applications: Association between two categorical (nominal or ordinal) variables from questionnaire data

Example: Personality and Colour Preference

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

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

Data of this type are usually summarised by counting the number of subjects in each personality/colour group and presenting it in the form of a table, known as a **cross-tabulation** or a **contingency table**.

The results of a **survey** of 400 students were tabulated as shown:

		Colour				
		Blue	**Green**	**Red**	**Yellow**	**Total**
Personality	**Introvert**	44	30	20	6	100
	Extrovert	36	50	180	34	300
	Totals	80	80	200	40	400

www.statstutor.ac.uk

© Mollie Gilchrist and Peter Samuels
Birmingham City University

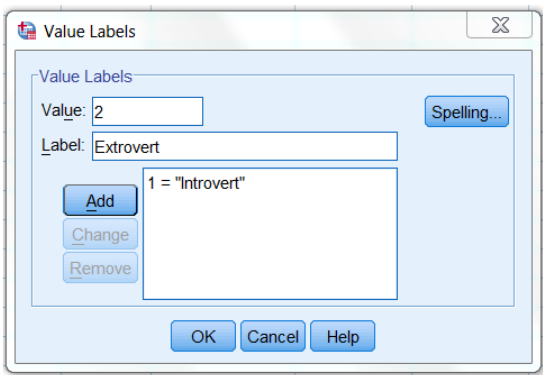
Reviewer: Ellen Marshall
University of Sheffield

Based on material provided by Loughborough University Mathematics Learning Support Centre and Coventry University Mathematics Support Centre

Steps in SPSS

Setting up the data file

In SPSS It is easier to import numbers from Excel then reassign them meaning with the Values option. Individual data should be entered in the format of the first table above, for example coding 1 for *Introvert* and 2 for *Extravert* for the *Personality* variable, and 1 for *Red*, 2 for *Yellow*, 3 for *Green* and 4 for *Blue* for the *Colour* variable.



Value Labels

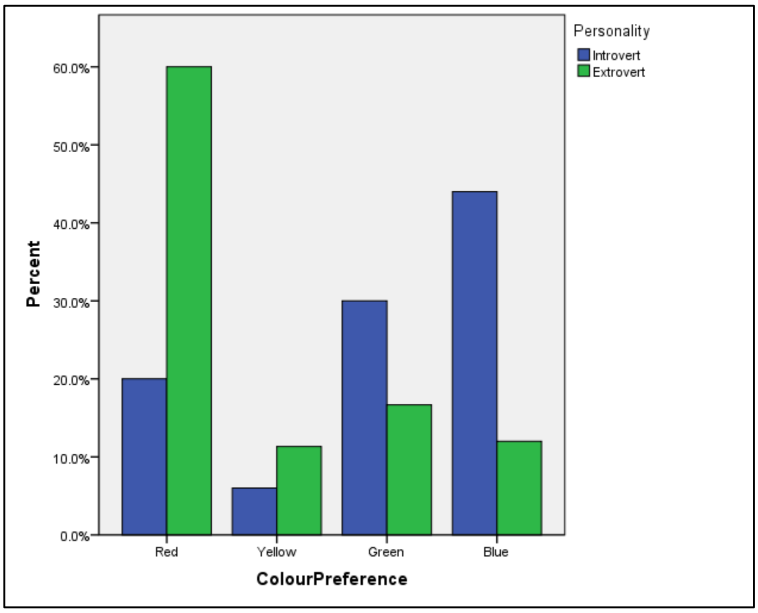
Value: 2

Label: Extrovert

1 = "Introvert"

Then in the Variable View, select the Values field of *Personality* and assign the meanings of 1 and 2 (see above) then select the Values field of Colour and enter the meanings of 1, 2, 3 and 4.

The data set can be visualised as a percentage frequency bar chart in SPSS, giving a distribution pattern for each colour:



Alternatively, the table below shows the row percentages.

This appears to indicate a significant association between colour preference and personality.

		Colour				
		Blue	**Green**	**Red**	**Yellow**	**Total**
Personality	**Introvert**	44%	30%	20%	6%	100%
	Extrovert	12%	16.7%	60%	11.3%	100%
	Totals	20%	20%	50%	10%	100%

Hypotheses

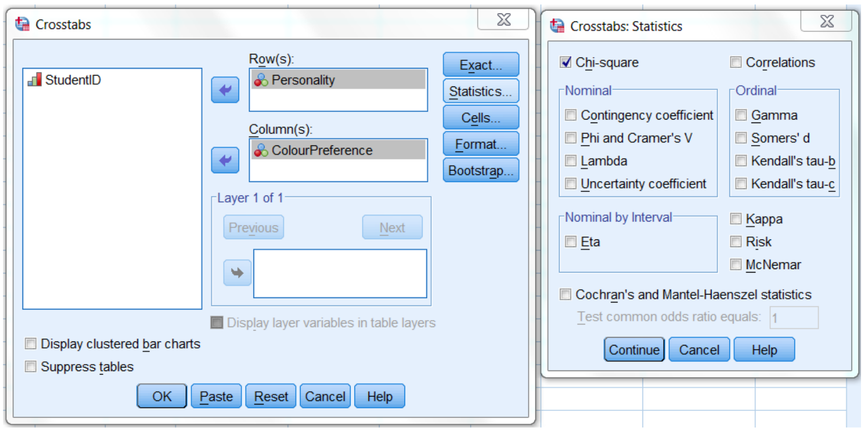
The null hypothesis is:

H_0: Colour preference is not associated with personality

The alternative hypothesis is:

H_1: Colour preference is associated with personality

Carrying out the analysis


Row(s): Personality

Column(s): ColourPreference

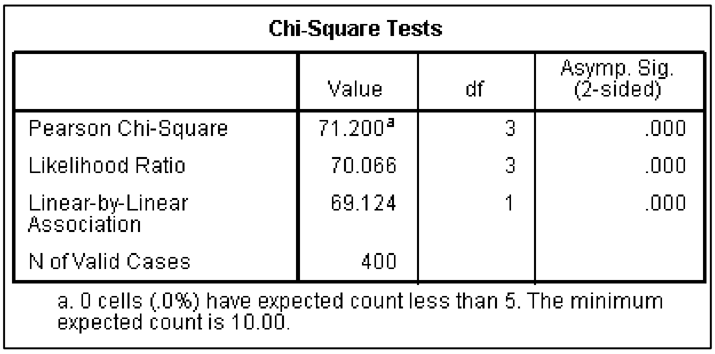
☑ Chi-square

Select Analyze – Descriptive Statistics – Crosstabs:

- Select one variable as the Row variable, and the other as the Column variable (see right)
- Click on the Statistics… button and select Chi-square and Continue.
- Click on the Cells… button and select Expected Counts and Continue, the select OK (note: expected counts are particularly helpful when the variable is ordinal – see note on validity below).

The output should look like this:

Results


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		

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

From the top row of this table we observe the Pearson Chi-Squared statistic, $\chi^2 = 71.20$, corresponding to $p < 0.001$ (note: the *Asymp. Sig. (2-sided)* value in this row, 0.000, is the p-value rounded to 3 decimal places and should not be quoted in this form). Therefore we reject the null hypothesis and conclude that there is very strong evidence of an association between *Personality* and *Colour*.

Note "a." at the bottom of the table indicates that the analysis is valid. We require less than 20% to have an expected count less than 5 and none to have an expected count less than 1.

Analysing data already grouped into a table

It is also possible to analyse grouped data in SPSS:

- Select Data – Weight Cases then select Weight cases by and choose your frequency variable as Frequency
- Repeat the steps as outlined above to get the same output as before

Personality	ColourPreference	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

Validity

Chi-squared tests are only valid when you have reasonable sample size.

For 2×2 tables (i.e. only two categories in each variable):

- If the total sample size is greater than 40, chi-squared can be used
- If the total sample size is between 20 and 40, and the smallest expected frequency is at least 5, chi-squared can be used (see note “a.” at the bottom of SPSS output to see if this is a problem)
- Otherwise Fisher’s exact test (two-sided) should be used (SPSS will automatically give this for 2×2 tables)

For other tables:

- Chi-squared 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)

  If the expected frequencies are a problem and one of the variables is ordinal then it may be possible to merge categories together using Transform – Recode into Different Variables… and testing again with the recoded variable.

  For example a Likert response scale question with values from 1 (Strongly Agree) to 5 (Strongly Disagree) could be recoded with 1 representing 1 and 2, i.e. Strongly Agree to Agree, etc. It is also possible to group nominal values together provided that the combined group is meaningful. However, the two-way table should be kept as large as possible whilst satisfying these validity requirements in order to use the richest possible raw data set.

  Alternatively, nominal variables can also be merged provided that the combined categories are meaningful (e.g. Red and Not Red).

Note

Chi-squared is a test of **association, not a test of correlation**. A positive result from a chi-squared test indicates that there is some kind of relationship between two variables but we do not know what sort of relationship it is. Even if an association is found between two ordinal variables, we cannot conclude that there is a correlation between them. This should be investigated using a Spearman rank correlation test provided there are about 10 values for each variable.

© Mollie Gilchrist and Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Based on material provided by Loughborough University Mathematics Learning Support Centre and Coventry University Mathematics Support Centre