



community project

encouraging academics to share statistics support resources

All stcp resources are released under a Creative Commons licence

stcp-karadimitriou-chisqS

The following resources are associated:
Summarising categorical variables in SPSS and the 'Titanic.sav' dataset

Chi-Squared Test for Association in SPSS

Dependent variable: Categorical

Independent variable: Categorical

Common Applications: Association between two categorical variables.

The chi-squared test tests the hypothesis that there is no relationship between two categorical variables. It compares the observed frequencies from the data with frequencies which would be expected if there was no relationship between the two variables.

Data: The dataset *Titanic.sav* contains data on 1309 passengers and crew who were on board the ship 'Titanic' when it sank in 1912. The question of interest is which factors affected survival. The dependent variable is 'Survival' and possible independent values are all the other variables. Here we will look at whether there's an association between nationality (Residence) and survival (Survived).

Variable name	<i>pclass</i>	<i>survived</i>	<i>Residence</i>	<i>Gender</i>	<i>age</i>	<i>sibsp</i>	<i>parch</i>	<i>fare</i>
Name	Class of passenger	Survived 0 = died	Country of residence	Gender 0 = male	Age	No. of siblings/ spouses on board	No. of parents/ children on board	price of ticket
Abbing, Anthony	3	0	USA	0	42	0	0	7.55
Abbott, Rosa	3	1	USA	1	35	1	1	20.25
Abelseth, Karen	3	1	UK	1	16	0	0	7.65

Data of this type are usually summarised by counting the number of subjects in each factor category and presenting it in the form of a table, known as a **cross-tabulation** or a **contingency table**. Row or column percentages are useful for summarising and comparing groups. The data set can be visualised as a stacked or clustered percentage frequency bar chart in SPSS, giving a distribution pattern for each group being compared.

Summary statistics

Use *Analyze* → *Descriptive statistics* → *Crosstabs* to produce a cross-tabulation with row percentages (see Chi-squared section on the next page) or use *Graphs* → *Legacy Dialogs* → *Bar* to produce a stacked bar chart and add percentages as labels. For more information on producing and interpreting charts and percentages in SPSS see the 'Summarising relationships between two categorical variables' resource.

Chi-squared for association in SPSS

From the stacked bar chart opposite, it looks like Americans were more likely to survive. 56% of Americans survived compared to 32% of British and 35% of other nationalities. To see if there is significant evidence of a relationship, a Chi-squared test should be carried out.

Hypotheses

The null hypothesis is

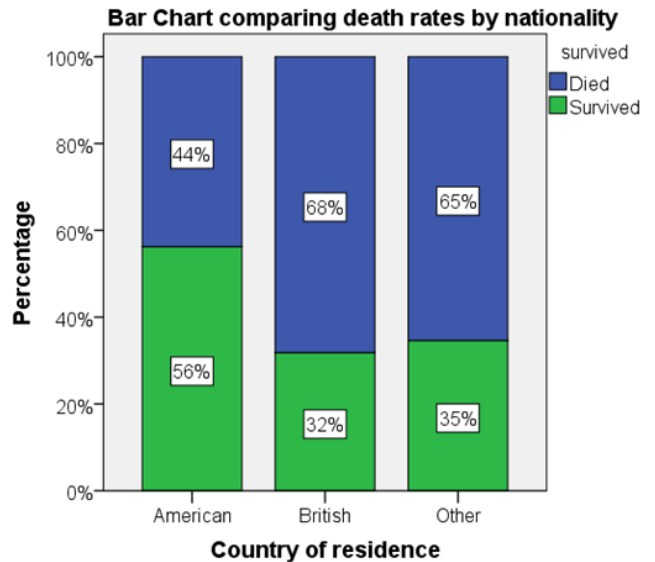
H_0 : Nationality is not associated with survival

The alternative hypothesis is

H_1 : Nationality is associated with survival

The observed frequencies of dying/surviving within each nationality are compared to the frequencies which would be expected if the null, there is no difference in survival between groups, is true.

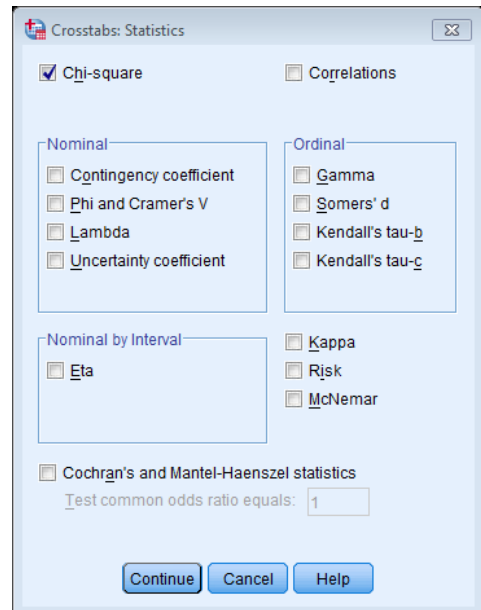
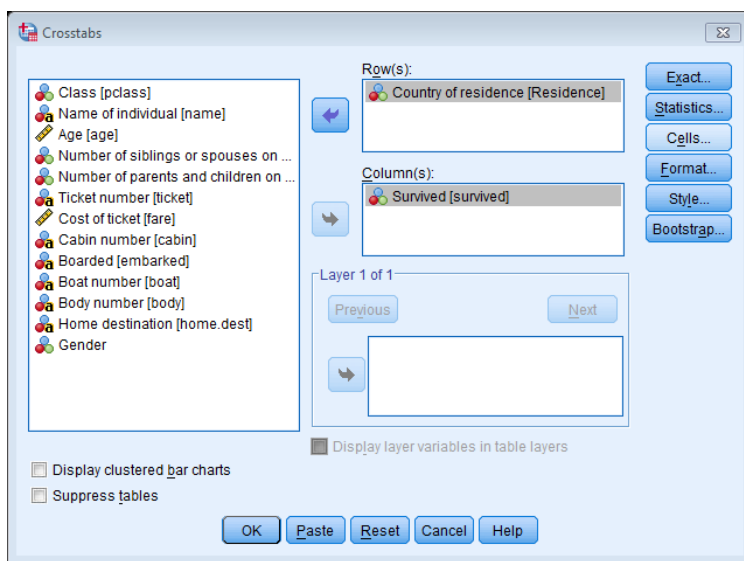
Overall 38% of passengers survived so if there was no association between nationality and survival, approximately 38% of passengers for each nationality would have survived.



Carrying out the analysis

Select *Analyze* → *Descriptive Statistics* → *Crosstabs*

- Select one variable as the Row variable, and the other as the Column variable
- Click on the **Statistics...** button and select *Chi-square* and *Continue*.



- Click on the **Cells...** button and select *Expected Counts* and **Continue** (note: expected counts are particularly helpful when the variable is ordinal – see note on validity below).
- You can also calculate row or column percentages in the cells section but choose carefully. Here we wish to calculate the % dying **within** nationality so select 'Row' percentages and **Continue**.
- The select OK to run the analysis.

Results of the Chi-squared

From the top row of the output table we observe the Pearson Chi-Squared statistic, $\chi^2 = 44.835$, degrees of freedom 2, corresponding to $p < 0.001$. Therefore we reject the null hypothesis with 99.9% confidence and conclude that there is very strong evidence of an association between *Nationality* and *Survival*.

Note: the *Asymp. Sig. (2-sided)* value in the Chi-squared row, 0.000, is the p-value rounded to 3 decimal places and should not be quoted in this form.

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	44.835 ^a	2	.000
Likelihood Ratio	43.765	2	.000
Linear-by-Linear Association	27.826	1	.000
N of Valid Cases	1309		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 98.55.

Reporting

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. You need to use summary statistics to discuss what the relationship is.

A Pearson's Chi-Squared test was carried out to assess whether nationality and survival were related. There was significant evidence of an association, ($\chi^2(2) = 44.835$, $p < 0.001$). 56% of Americans survived compared to 32% of British and 35% of other nationalities.

Validity

Chi-squared tests are only valid when you have reasonable sample size, less than 20% of cells have an expected count less than 5 and none have an expected count less than 1. Note "a." below the output table indicates that the analysis is valid and no cells have expected counts less than 5.

2x2 tables

Contingency tables are often referred to by the number of categories of the two variables. For example, a 2x2 table has two categories for each variable e.g. if the association between gender and survival were investigated. If you repeat the steps for the Chi-squared test but use 'Gender' instead of 'Residence' you will get the following output which contains two extra rows.

		Gender		Total
		Male	Female	
Survived	Died	682	127	809
	Survived	161	339	500
Total		843	466	1309

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	365.887 ^a	1	.000		
Continuity Correction ^b	363.618	1	.000		
Likelihood Ratio	372.921	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	365.607	1	.000		
N of Valid Cases	1309				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 178.00.

b. Computed only for a 2x2 table

Chi-squared for association in SPSS

The **Continuity Correction** is an adjustment to the Chi-squared for 2x2 tables and should be reported although for large sample sizes, the Chi-squared test with and without continuity corrections will be similar. If the total sample size is between 20 and 40, no expected values should be below 5. For small samples or when the minimum expected count is under 5, the p-value for **Fisher's exact test** (Exact Sig. (2-sided)) should be used which is given automatically for 2x2 tables.

For other tables:

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, a Fishers Exact test can be computed using by clicking on the **Exact** button within the crosstabs menu and selecting the 'Exact' option.

Ordinal variables: Chi-squared is a test of association, not a test of correlation and assumes the variables are nominal. Even if an association is found between two ordinal variables, we cannot conclude that there is a correlation between them. If both variables are ordinal, the Mantel-Hanzel **Linear by Linear association** row can be reported which tests for a linear association. Spearmans's rank correlation or Kendall's Tau can also be used to investigate this.

Analysing data already grouped into a table

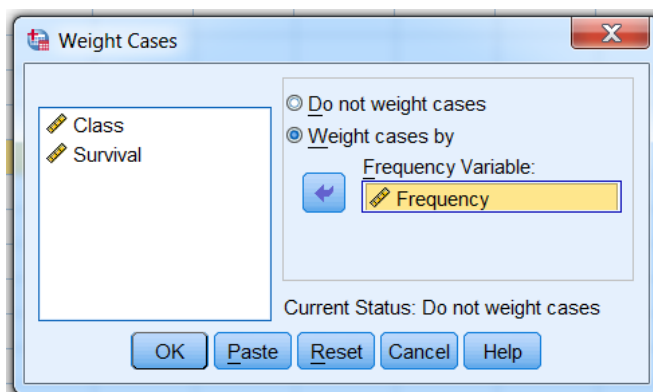
It is also possible to analyse summary data (taken from a contingency table) in SPSS:

		survived		Total
		Died	Survived	
Class	1st	123	200	323
	2nd	158	119	277
	3rd	528	181	709
	Total	809	500	1309



	Class	Survival	Frequency
1	1	0	123
2	1	1	200
3	2	0	158
4	2	1	119
5	3	0	528
6	3	1	181
7			

Select *Data* → *Weight Cases* then select *Weight cases by* and choose your frequency variable as Frequency. SPSS now treats the data as if there were 123 rows of people in 1st class who died etc.



Repeating the Chi-squared steps as outlined above gives the same output as before.