The following resources are associated:

'Summarising categorical variables in SAS' and the 'Titanic.csv' dataset

Chi-Squared Test for Association in SAS

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.

	pclass	survived	Residence	name	age
1	3	0	0	Abbing, Mr. Anthony	42
2	3	0	0	Abbott, Master. Eugene Joseph	13
3	3	0	0	Abbott, Mr. Rossmore Edward	16

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). Import the data set into SAS and call it 'Titanic'.

The variables Survived and Residence both have numbers representing different categories. For example, survived has 0 and 1 representing survived and died respectively. Value labels can be defined using the **`proc format`** command which can then be applied to any variable within procedure commands so that output displays labels rather than numbers.

`proc format` means format procedure which is used to format the data variables.
`value` allows numerical values to be renamed.
`surv` & `nat` – These are abbreviations for these formats which are applied to individual variables.

```
proc format;
      value surv 1='Survived' 0='Died';
      value nat 0='American' 1='British' 2='Other';
      value cla 1='1st' 2='2nd' 3='3rd';
run;
```

Summary tables and tests can now be produced with the correct formats through the format line.

Carrying out summary statistics

Categorical data 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. Choose the percentage of the dependant variable within each independent category. The data set can be visualised as a stacked or clustered percentage bar chart in SAS (see '*Summarising categorical variables in SAS*' sheet).

proc freq means frequency procedure which is used to create a frequency table.
data=Titanic is used to determine which dataset is being used.
Tables is the command used within **proc freq** to create a table. Residence*survived means residence (rows) compared against survived (columns) which are the two categorical variables that are being analysed. / is used to enable certain table options. By default, SAS will give row, column and total percentages but only one is needed. Here we want the row percentages (percentage surviving within each nationality) so use nocol to take out the column percentage and nopercent to remove the total percentages.
format will use the formats that were created earlier so that actual names are used within the tables instead of values. Put the variable name first followed by the format name defined earlier.

```
proc freq data=Titanic;
      tables Residence*survived / nocol nopercent;
      format Survived surv. ;
      format Residence nat. ;
run;
```

This creates the summary table opposite, which 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.

Table of Residence by survived

		survived		
		Died	Survived	Total
Residence(Nationality)				
American	Frequency	113	145	258
	Row Pct	43.80	56.20	
British	Frequency	206	96	302
	Row Pct	68.21	31.79	
Other	Frequency	490	259	749
	Row Pct	65.42	34.58	
Total	Frequency	809	500	1309

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

A chi-squared test can be carried out using an extension to the tables row within the **proc freq**. **chisq** will produce a chi-square test as well as the summary statistics above.

```
proc freq data=Titanic;
      tables Residence*survived / chisq nocol nopercent;
      format Survived surv. ;
      format Residence nat. ;
run;
```

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 a p-value of less than 0.001. Therefore the null hypothesis is rejected and we conclude that there is very strong evidence of an association between *Nationality* and *Survival*.

Statistic	DF	Value	Prob
Chi-Square	2	44.8346	<.0001
Likelihood Ratio Chi-Square	2	43.7648	<.0001
Mantel-Haenszel Chi-Square	1	27.8265	<.0001
Phi Coefficient		0.1851	
Contingency Coefficient		0.1820	
Cramer's V		0.1851	

Sample Size = 1309

Reporting

A significant 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.

2x2 tables

Contingency tables are often referred to by the number of categories of the two variables. For example, a 2×2 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 an extra row in the Chi-squared output.

Table of Gender by survived				
		survived		
		Died	Survived	Total
Gender				
Male	Frequency	682	161	843
	Row Pct	80.90	19.10	
Female	Frequency	127	339	466
	Row Pct	27.25	72.75	
Total	Frequency	809	500	1309

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.

Statistic	DF	Value	Prob
Chi-Square	1	365.8869	<.0001
Likelihood Ratio Chi-Square	1	372.9213	<.0001
Continuity Adj. Chi-Square	1	363.6179	<.0001
Mantel-Haenszel Chi-Square	1	365.6074	<.0001

Validity

Chi-squared tests are only valid when you have reasonable sample size, less than 20% of expected values are less than 5 and none have an expected count less than 1. Note that a warning below the output table will appear if any expected values are under 5. Here no warning is displayed which indicates that the analysis is valid and no cells have expected values less than 5.
You can request the expected values using `expected` after the / in the tables line to see which are too small.

```
tables Residence*survived /expected;
```

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** should be used which is given automatically for 2×2 tables in a separate table. For other types of table, add the `fisher` option in the tables command

```
tables Residence*survived /fisher;
```

Fisher's Exact Test	
Cell (1,1) Frequency (F)	682
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	<.0001
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Fisher's exact test uses pure probability calculations based on every combination of category frequencies given the variable totals. It therefore makes no assumptions and has no test statistic.
The two-sided P-Value should be used in the bottom row of the fisher's

table and interpreted in the same way as the Chi-squared test e.g. significant evidence of an association.

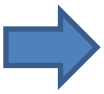
For other tables: If the expected frequencies are a problem, it may be possible to merge categories together and test again with the recoded variable. For example, a Likert response scale question with values from 1 (Strongly Agree) to 5 (Strongly Disagree) could be recoded with 1 representing 1 and 2, i.e. Strongly Agree to Agree, etc and 0 otherwise. It is also possible to group nominal values together provided that the combined group is meaningful and this is advisable for small sample sizes/ expected values. Care should be taken when combining categories to ensure information about different groups is not lost. Alternatively, a Fishers Exact test can be produced.

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 linear relationship between them. If both variables are ordinal, the Mantel-Haenszel row can be reported instead which tests for a linear association. The test has the same requirements for expected values as the standard Chi-squared. For more information on the other tests in the Chi-squared table see: https://stats.idre.ucla.edu/sas/output/proc-freq/

Analysing data already grouped into a table

It is also possible to analyse summary data (taken from a contingency table) in SAS. In this example we have class and survival.

		survived		
		Died	Survived	Total
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

Read in the data allowing one row for each combination of class and survival. The `input` line states the variable names and the `datalines` requires the group codes and frequencies.

```
data Titanicmini;
input class  survived  frequency;
datalines;
1 0 123
1 1 200
2 0 158
2 1 119
3 0 528
3 1 181
run;
```

The `weight` command must be used within any procedure to tell SAS that this is summary rather than raw data and that the `frequency` column indicates the number of individuals for a particular combination of groups. For example there are 123 rows of people in 1st class who died etc. Repeat the Chi-squared steps with the additional `weight` row to get the standard output and interpret in the same way.

```
proc freq data=Titanicmini;
      tables Class*survived / chisq nocol nopercent;
```

```
      format Survived surv. ;
      format Class cla. ;
      weight frequency;
run;
```