statstutor community project

encouraging academics to share statistics support resources

stcp-marshallowen-6a

The following resources are associated:

Data_for_tutor_training_SPSS_workbook.xls

Solutions_for_tutor_training_SPSS_workbook.pdf

SPSS workbook for New Statistics Tutors

(Based on SPSS Versions 21 and 22)

This workbook is aimed as a learning aid for new statistics tutors in mathematics support centres

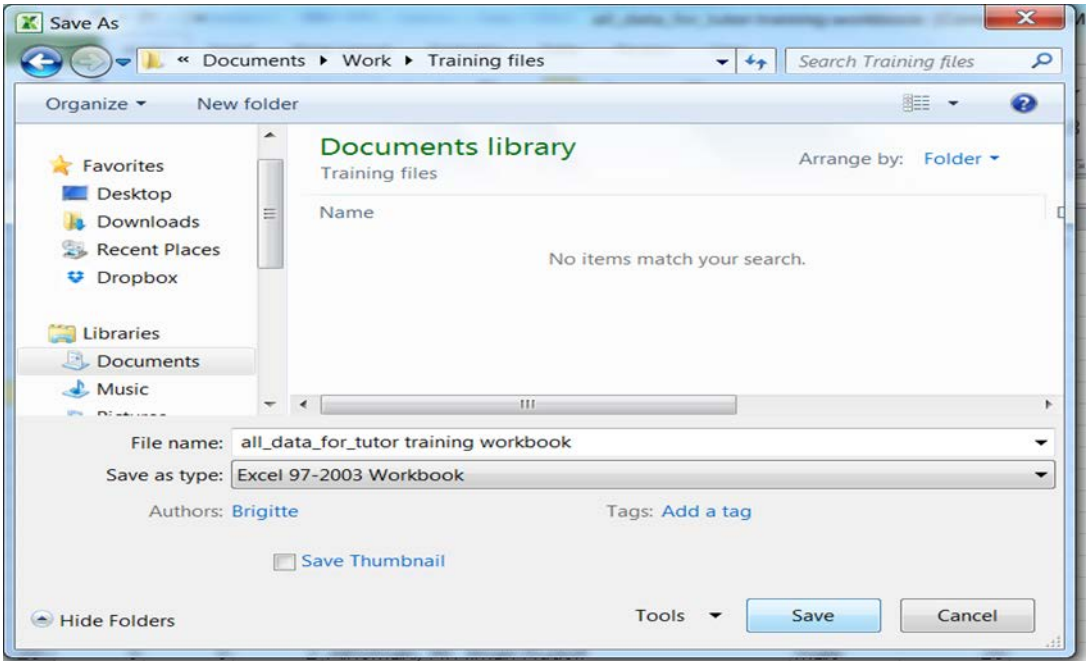
Reviewer: Jean Russell, University of Sheffield

Contents

Data sets used in this booklet

All the data needed for this booklet is contained in the Excel file 'Data_for_tutor_training_SPSS_workbook'.

You will need to save this file on your computer in order to use it.



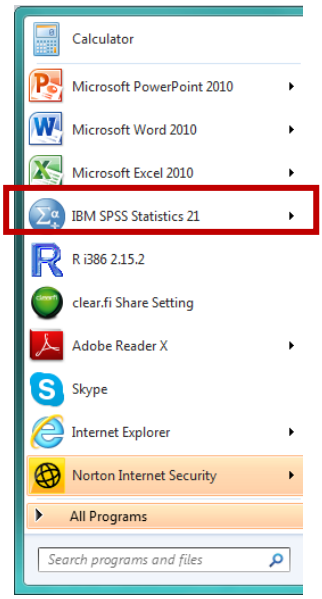
Datasets:

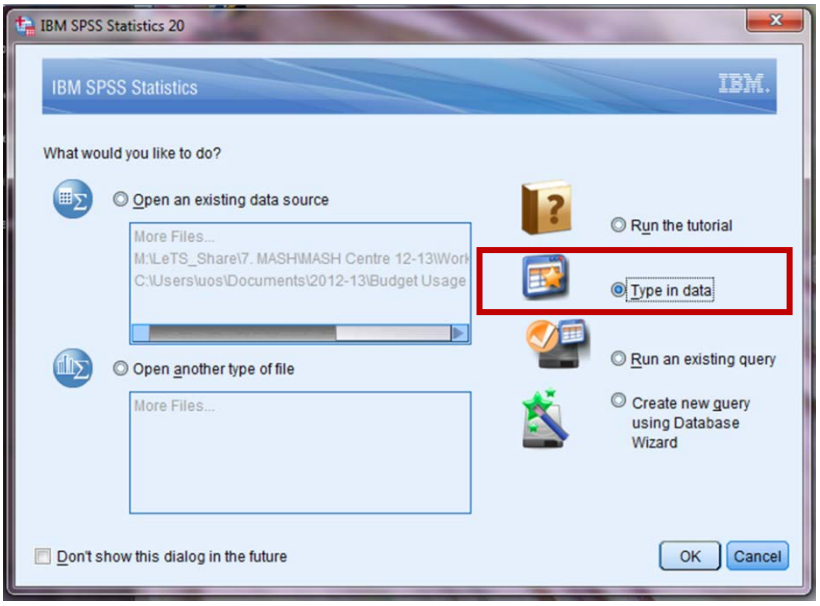
Dataset	Description
Titanic	List of 1309 passengers on board the Titanic when it sank and details about them such as gender, whether they survived, class etc
Diet	78 people were put on one of three diets with the goal being to determine which diet was best.
Birthweight	Details for a number of babies and their parents such as weight and length of babies at birth and weight and height of mother.
Birthweight reduced	This is a reduced set of the above data set. It's used to demonstrate correlation and regression instead of the main set so that graphs are clearer.
Cholesterol	Study investigating whether a certain brand of margarine reduces cholesterol over 3 time points.
Video	This study had 4 methods for explaining a certain condition and wished to find out which method people preferred.

Getting started in SPSS?

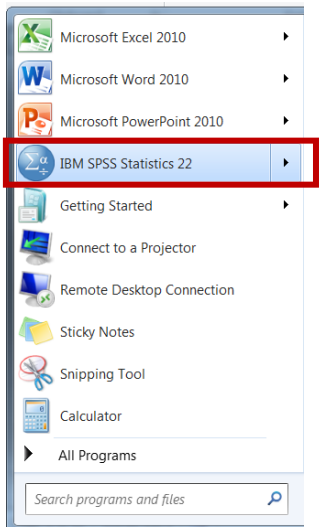
SPSS is similar to Excel but it's easier to produce charts and carry out analysis. To open SPSS versions 21 or 22 you would typically select 'IBM SPSS statistics' from 'All programs'. Before SPSS opens for the first time, an additional screen will probably appear as shown below. You can open a dataset from this screen but in Version 21 it's easiest to just select 'Type in data' every time. Data can be opened after SPSS is opened.

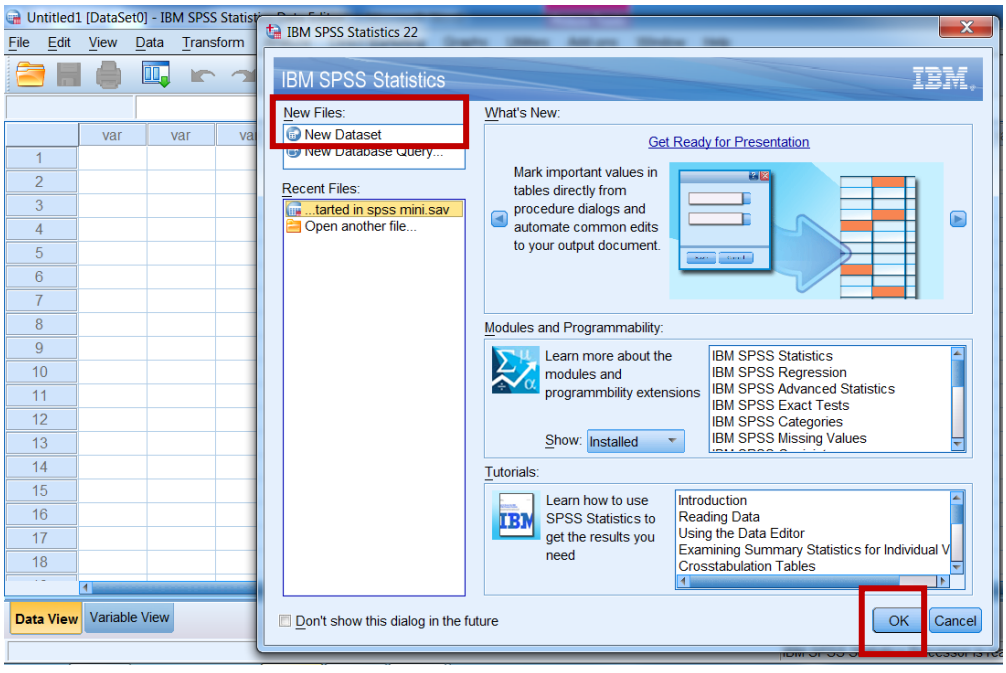
Version 20 - 21:





In version 22, select 'New Dataset' and 'OK'.





Reviewer: Jean Russell, University of Sheffield

Example of data sheet in SPSS

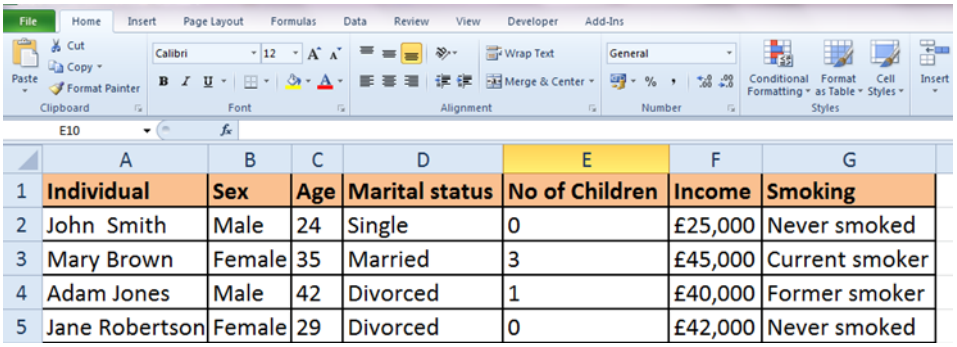
	Individual	Sex	Age	MaritalStatus	Children	Income	Smoking
1	John Smith	Male	24	Single	0	25000	Never smoked
2	Mary Brown	Female	35	Married	3	45000	Current smoker
3	Adam Jones	Male	42	Divorced	1	40000	Former smoker
4	Jane Robertson	Female	29	Divorced	0	42000	Never smoked

Variable headings can only appear at the top in the blue boxes

Unlike Excel, you can only have one dataset on each page of SPSS. A new file must be created for each individual data set.

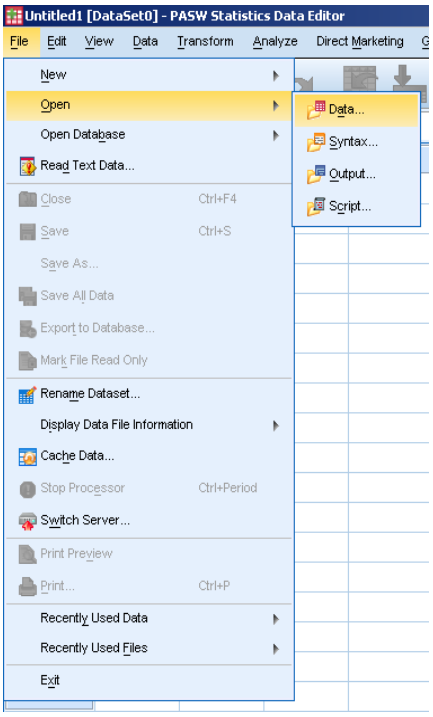
Opening an Excel file in SPSS

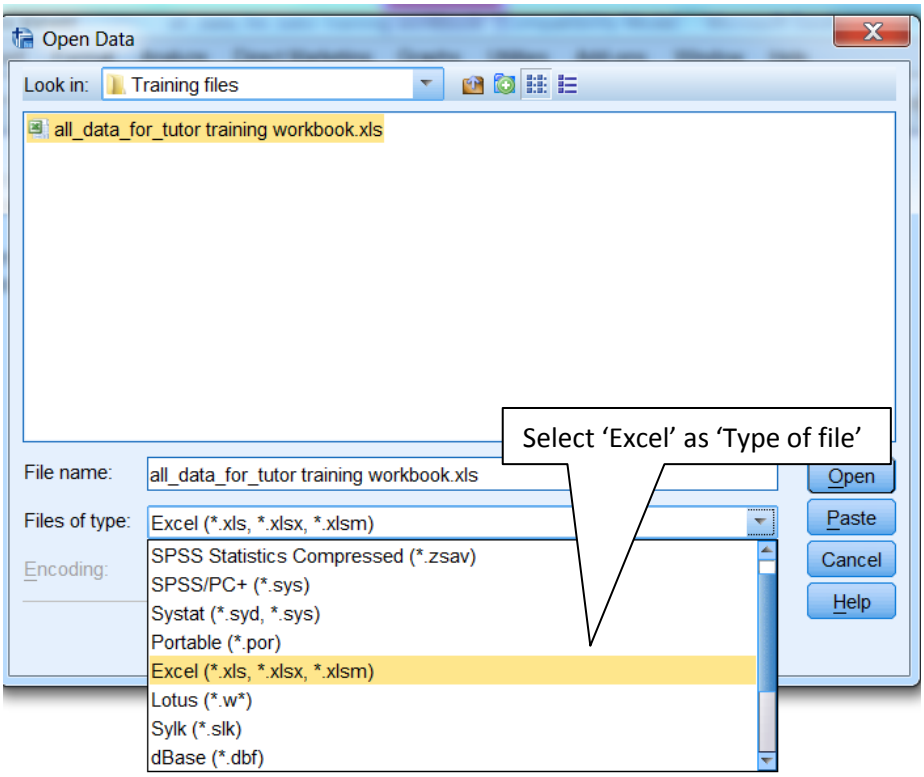
Important note: There must be only one row with headings in for SPSS to open an Excel file correctly.



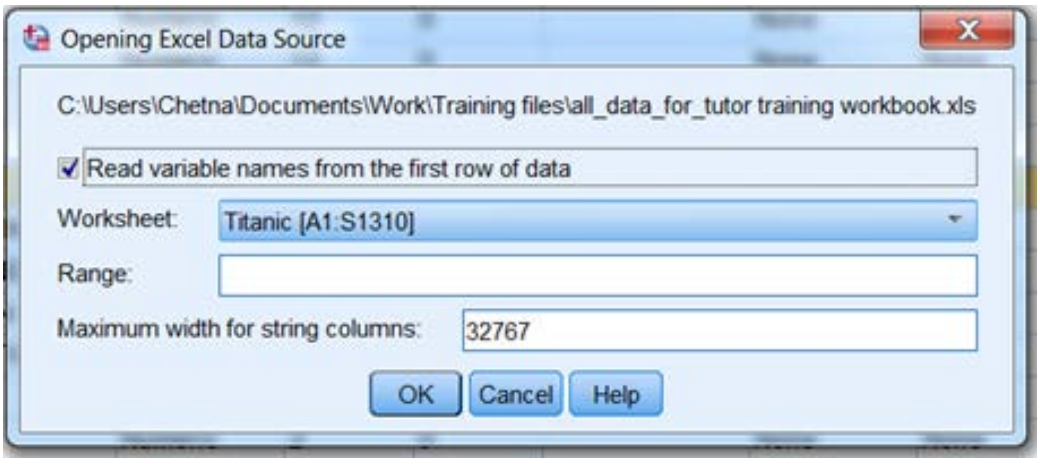
	A	B	C	D	E	F	G
1	**Individual**	**Sex**	**Age**	**Marital status**	**No of Children**	**Income**	**Smoking**
2	John Smith	Male	24	Single	0	£25,000	Never smoked
3	Mary Brown	Female	35	Married	3	£45,000	Current smoker
4	Adam Jones	Male	42	Divorced	1	£40,000	Former smoker
5	Jane Robertson	Female	29	Divorced	0	£42,000	Never smoked

To open any file in SPSS, select *File →Open → Data*. Here we are opening the 'Titanic' data which is currently in Excel. Note: The Excel must not be open on your computer.





SPSS only opens one sheet of data at a time so select the required sheet containing the Titanic data.



Titanic data

The ship 'The Titanic' sank in 1914 along with most of its' passengers and crew. The data set that we have contains information on 1309 passengers.

The Titanic data: Details for passengers travelling on the Titanic when it sank:

Variable name	*pclass*	*survived*	*Residence*	*Gender*	*age*	*sibsp*	*parch*	*fare*
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
Type of variable	Ordinal							

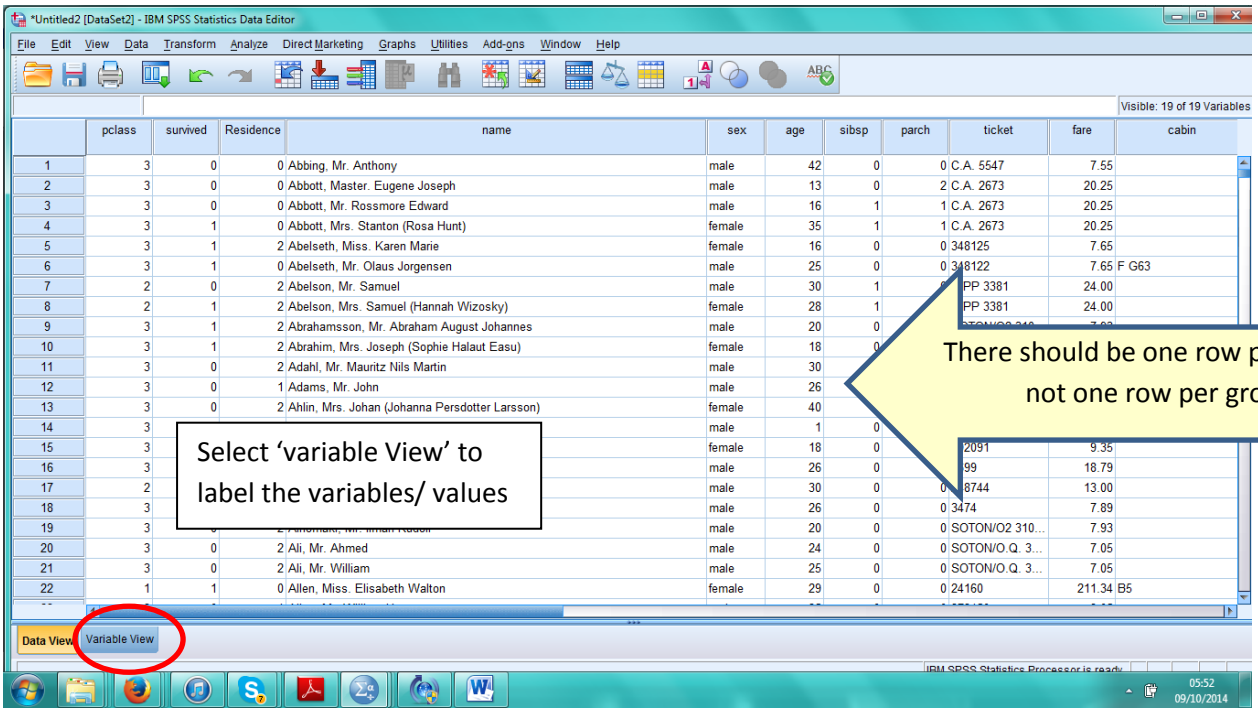
Exercise 1: Were wealthy people more likely to survive on the Titanic?

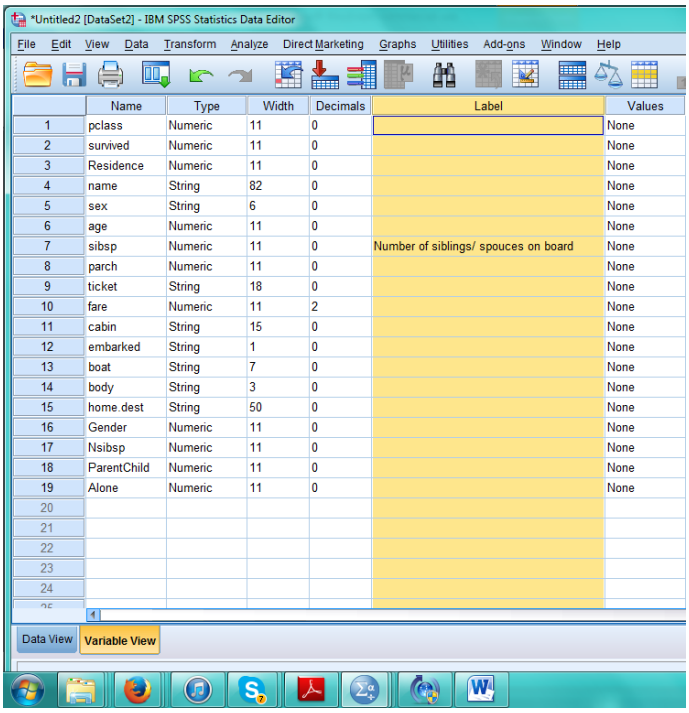
The Titanic data: Details for passengers travelling on the Titanic when it sank:

a) What variable type is each of the variables in the table above?

b) Which variables would you use to investigate the research question: 'Were wealthy people more likely to survive the sinking of the Titanic'?

There are two sheets for each dataset. The 'Data View' sheet is where the numbers are entered and the 'Variable View' sheet is where the variables are named and defined. The option to choose between Data and Variable View is in the bottom left hand corner. For data in categories, type numbers in the Data View sheet and then label the numbers in 'Variable View'.





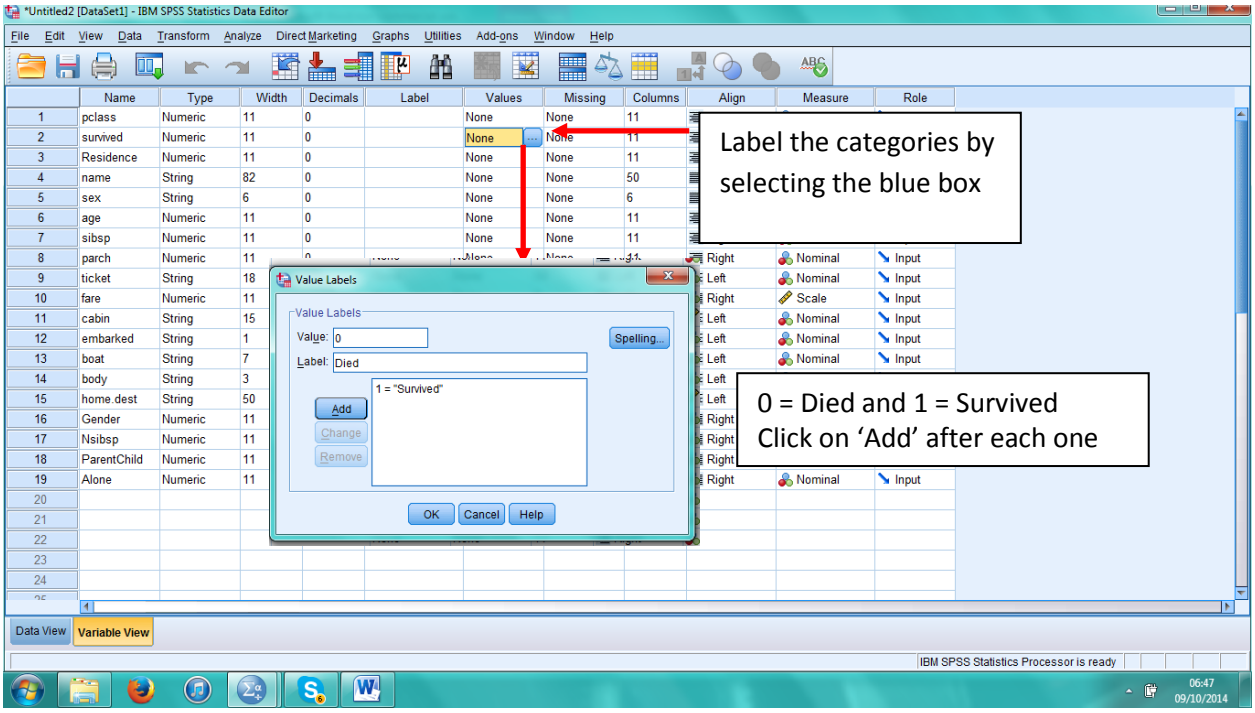
Variable view: Label the variables

The variable name has restrictions. It can have no spaces or use certain characters. Use the 'Label' column to give sensible variable descriptions which will appear in all output. If the label is blank, the variable name will appear in output.

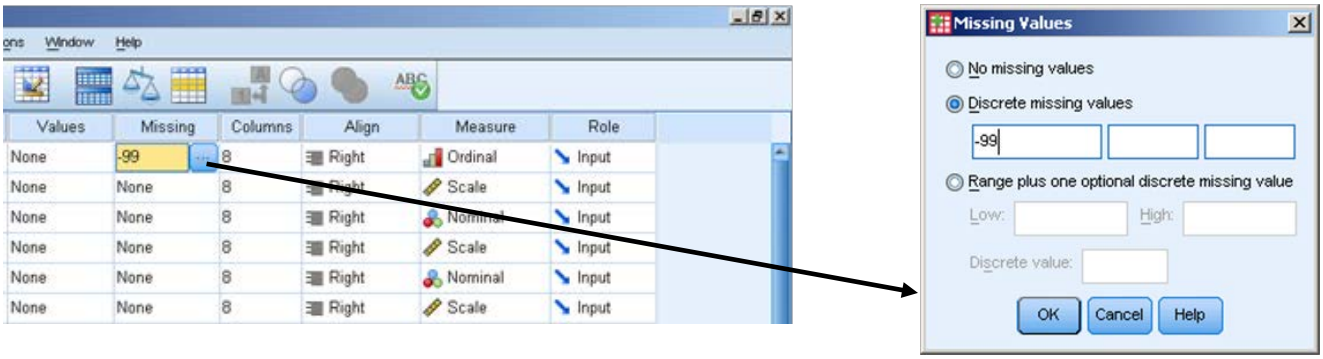
For example sibsp is 'Number of siblings/ spouses on board', parch is 'Number of parents/ children on board' and fare is 'Price of ticket'.

Labelling values

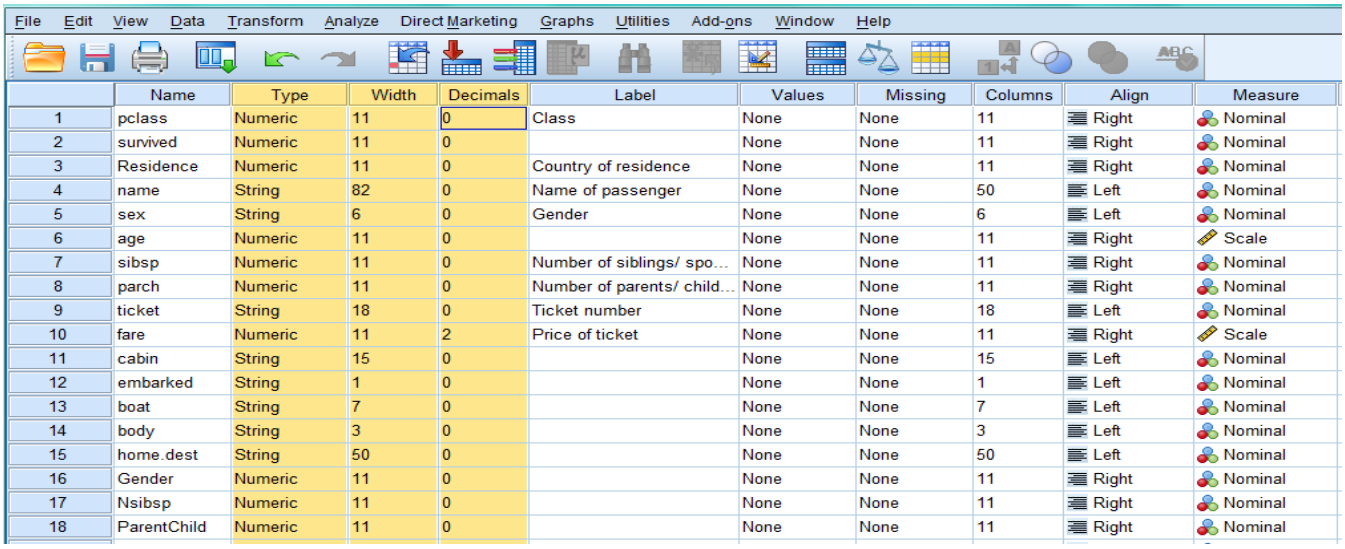
It is best to have your categories coded as numbers for analysis in SPSS but for your output, people need to know what the numbers mean. Go to the 'Values' column in 'Variable View', let the mouse hover until you see a blue square. Clicking the square gives the 'Value labels' box. In the value box, put the number and the label for that number in the label box. Click on 'Add' after each label and 'Ok' when finished.



Also, when using secondary data, watch for odd values, such as -99 indicating a missing value. These can be identified in the missing column so they are not taken into account in any analysis.



Reviewer: Jean Russell, University of Sheffield



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	pclass	Numeric	11	0	Class	None	None	11	Right	Nominal
2	survived	Numeric	11	0		None	None	11	Right	Nominal
3	Residence	Numeric	11	0	Country of residence	None	None	11	Right	Nominal
4	name	String	82	0	Name of passenger	None	None	50	Left	Nominal
5	sex	String	6	0	Gender	None	None	6	Left	Nominal
6	age	Numeric	11	0		None	None	11	Right	Scale
7	sibsp	Numeric	11	0	Number of siblings/ spo...	None	None	11	Right	Nominal
8	parch	Numeric	11	0	Number of parents/ child...	None	None	11	Right	Nominal
9	ticket	String	18	0	Ticket number	None	None	18	Left	Nominal
10	fare	Numeric	11	2	Price of ticket	None	None	11	Right	Scale
11	cabin	String	15	0		None	None	15	Left	Nominal
12	embarked	String	1	0		None	None	1	Left	Nominal
13	boat	String	7	0		None	None	7	Left	Nominal
14	body	String	3	0		None	None	3	Left	Nominal
15	home.dest	String	50	0		None	None	50	Left	Nominal
16	Gender	Numeric	11	0		None	None	11	Right	Nominal
17	Nsibsp	Numeric	11	0		None	None	11	Right	Nominal
18	ParentChild	Numeric	11	0		None	None	11	Right	Nominal

Variable Type: SPSS only analyses Numeric variables for some functions. String means it's a word. The **width** is the number of numbers/ letters allowed for that

Decimals: When typing in data, the default number of decimals is 2. Change this to 0 for categorical and discrete data.

The **Measure** column is where the data type is entered. **Continuous/ discrete are called Scale in SPSS.** SPSS won't allow certain analyses for the wrong type of variable.

Quick exercise: Give the variables and values suitable labels, check variables with numbers are numeric and choose the right data type for each variable.

Note: There are two variables for gender. 'Sex' is a string variable (words) whereas 'Gender' has 0 for males and 1 for females so should be used during analysis.

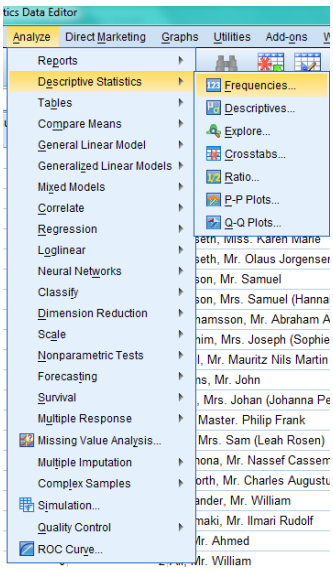
Variable Value	0	1	2
Gender	Male	Female	
Survived	Died	Survived	
Country of residence	America	Britain	Other

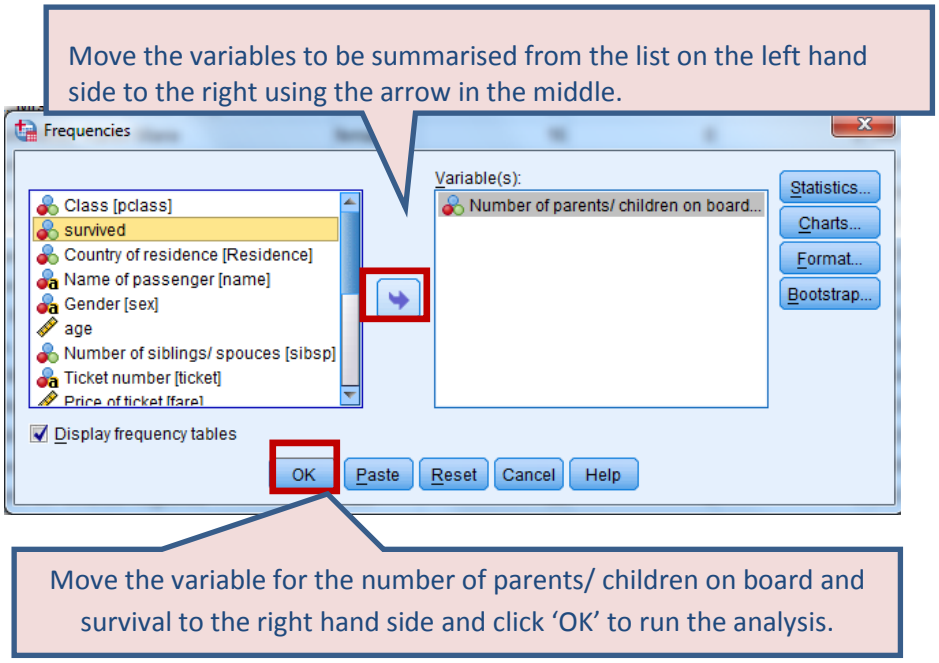
Once the data is in SPSS, **save the SPSS data file using *File → Save as***. Save again after making changes to the data.

Summarising categorical data

The simplest way to summarise a single categorical variable is by using frequencies or percentages.

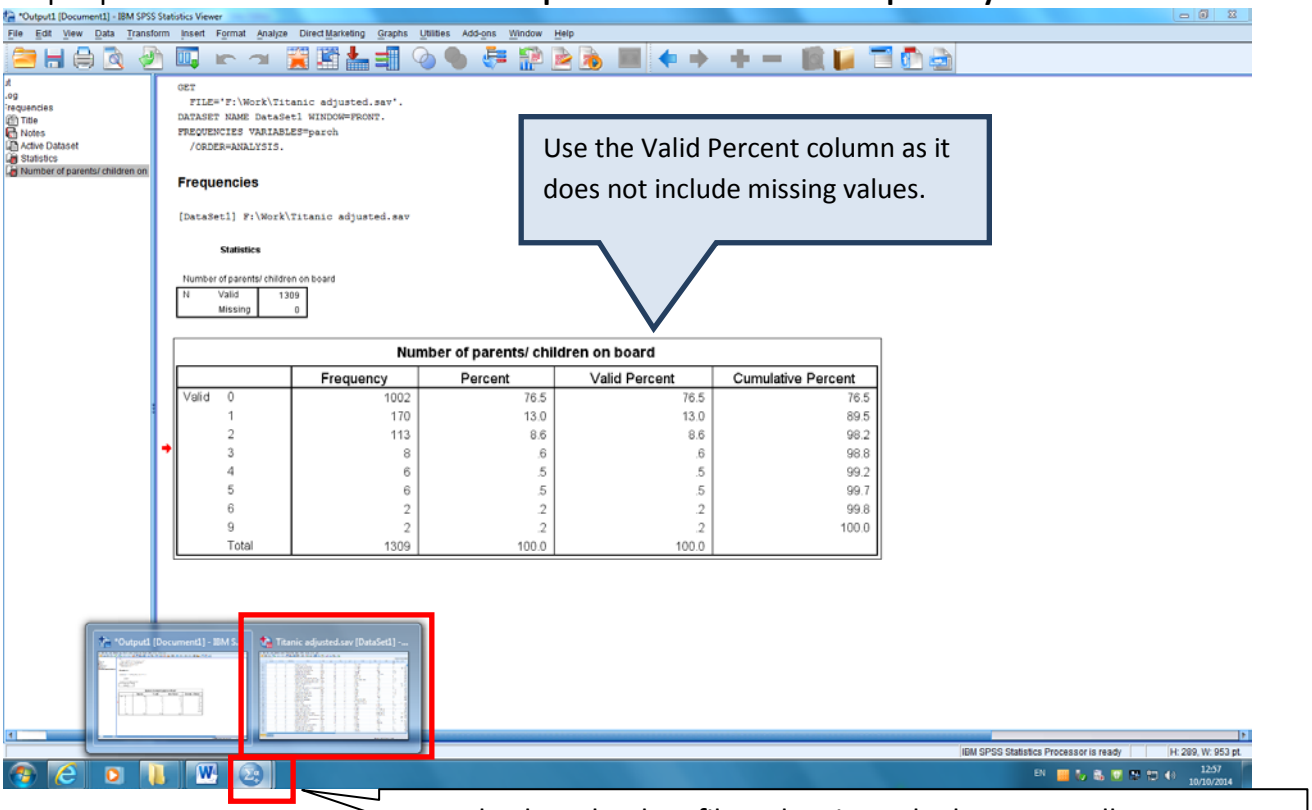
Analyse → Descriptive statistics → Frequencies





Output in SPSS

Charts, tables and analysis appear in a separate 'output' window in SPSS. The output window is brought to the front of the screen when analysis/ charts etc are requested. The left hand column shows all of the output produced in that session. **The output file has to be saved separately to the data file.**



As SPSS produces a lot of output for analysis and you may produce several charts before you decide which one is best, copying the output you require for your project and pasting into a Word document is preferable.

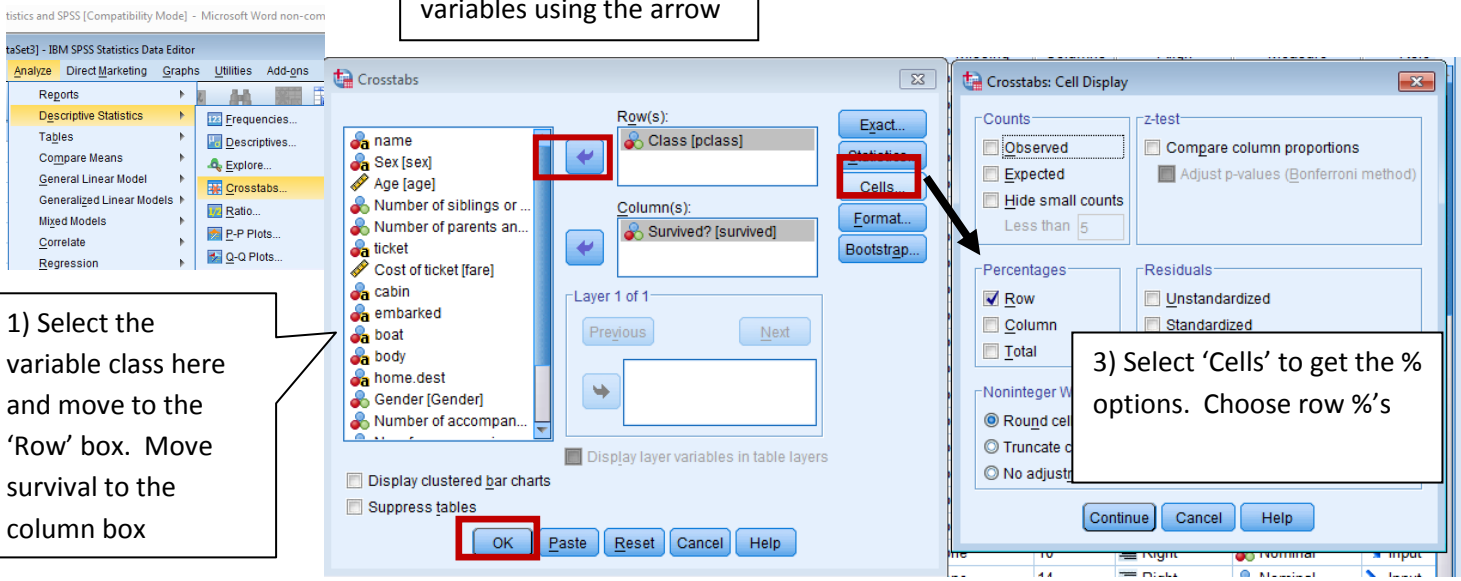
Quick question: What percentage of people survived the sinking of the Titanic?

Research question 1: Were wealthy people more likely to survive on the Titanic?

To break down survival by class, a cross tabulation or contingency table is needed. Percentages are usually preferable to frequencies but remember to include counts for small sample sizes. Choose either row or column percentages carefully.

Analyse → Descriptive statistics → Crosstabs

2) Move selected variables using the arrow



1) Select the variable class here and move to the 'Row' box. Move survival to the column box

3) Select 'Cells' to get the % options. Choose row %'s

4) Select 'OK' when finished and the chart appears in the output

Quick question: What percentage of people survived the sinking of the Titanic in each class?

Class	% Died	% Survived
1st		
2nd		
3rd		

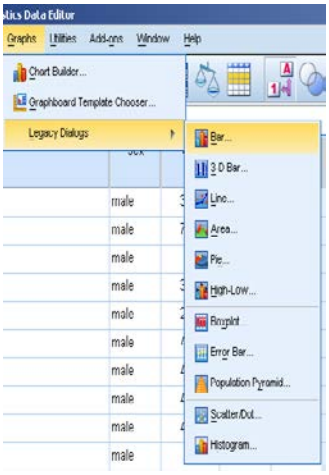
Bar Charts

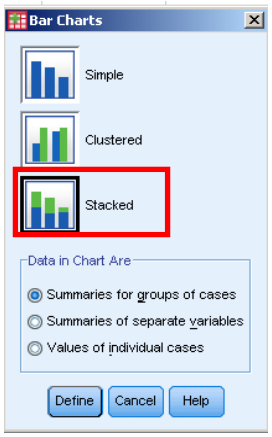
Plotting graphs in SPSS is much easier than in Excel. All graphs can be accessed through

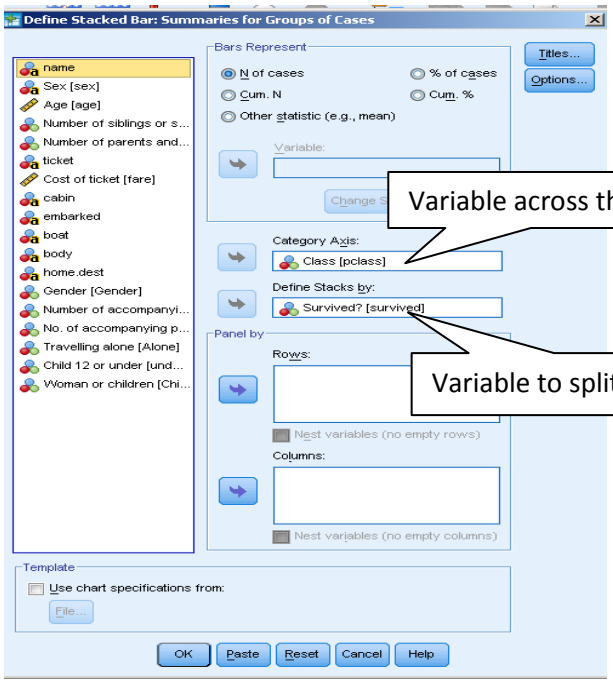
Graphs → Legacy Dialogs

There is a chart builder option but the legacy dialogs options are more user friendly.
To display the information from the cross-tabulation graphically, use either a stacked or clustered bar chart. Both of these can be accessed through

Graphs → Legacy Dialogs → Bar

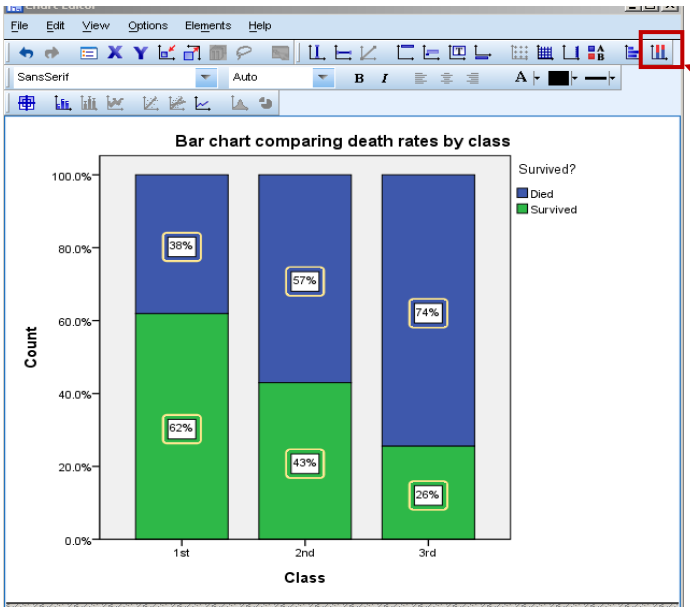






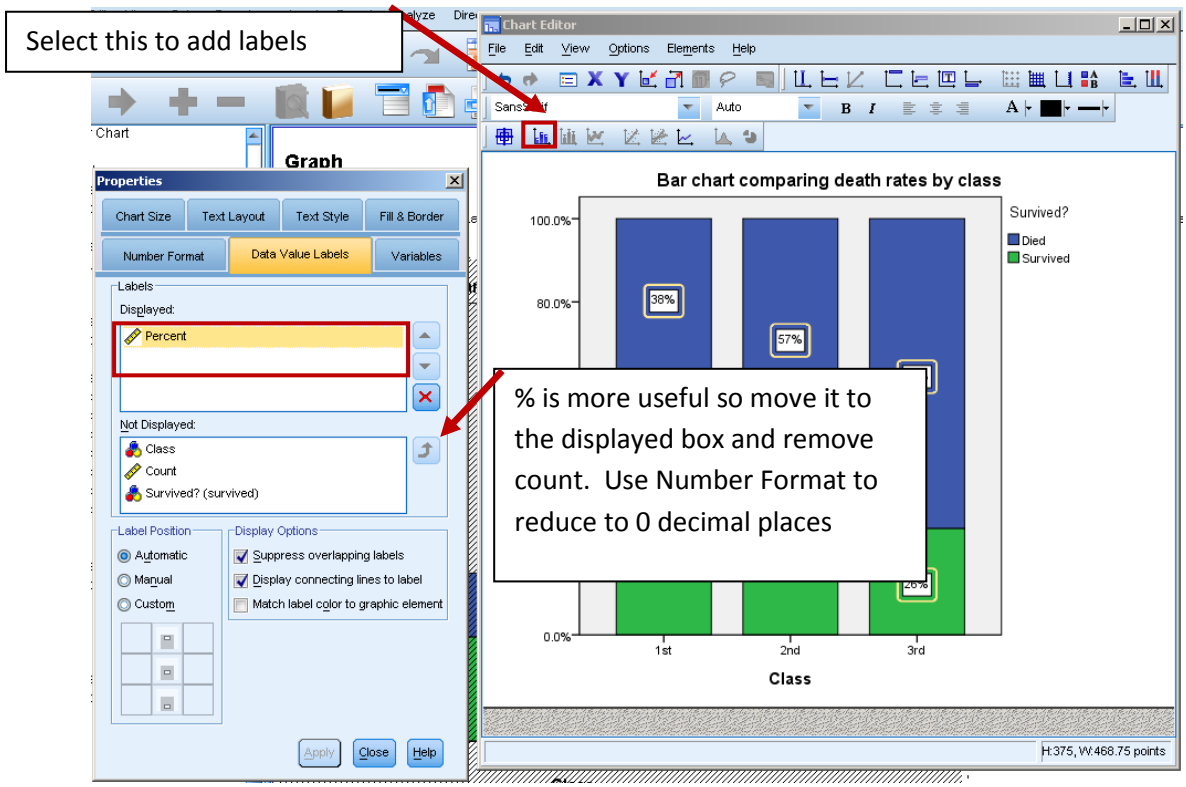
Tidying up a bar chart

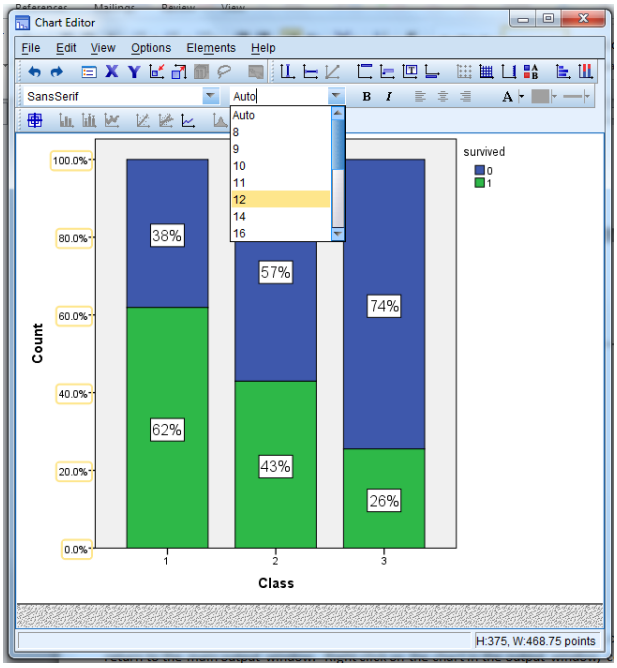
Double click on the chart to open an editing window.



Selecting this turns the bars into 100% for each class

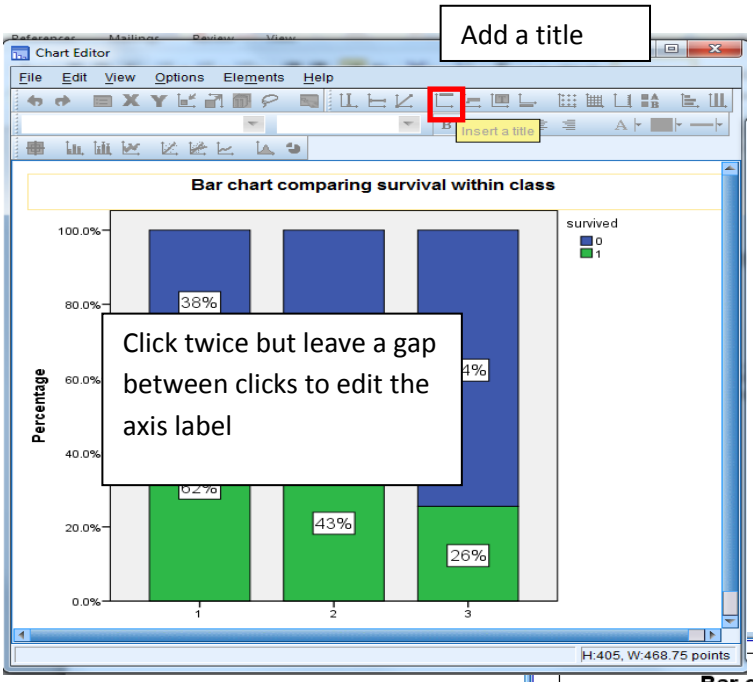
Reviewer: Jean Russell, University of Sheffield





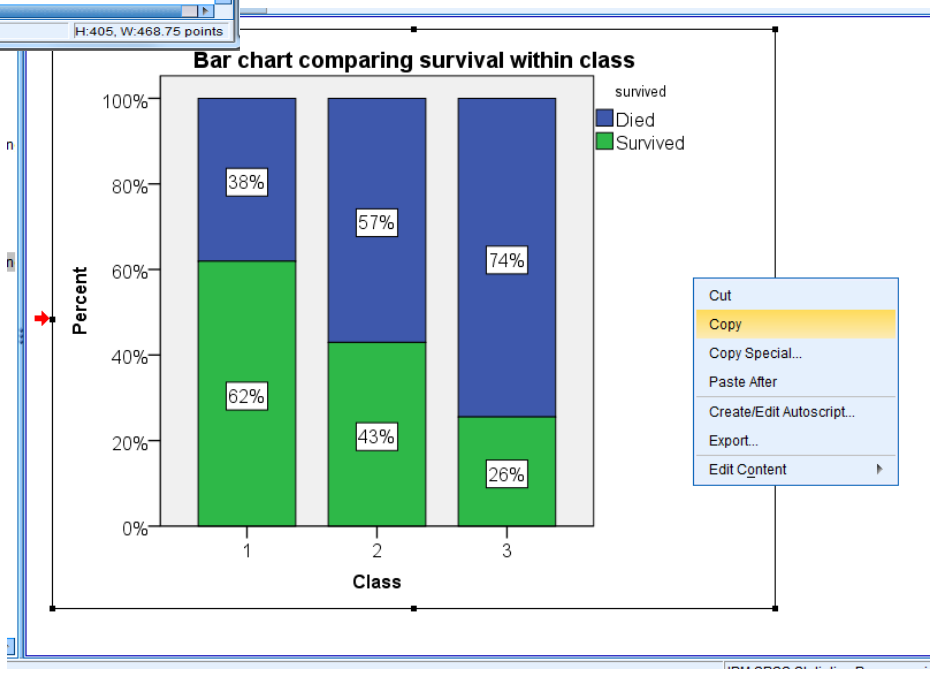
The font in graphs is usually small so adjust the axes titles etc. Select each axis and change the font size to 12. The axis titles and percentages displayed on the bars can also be changed in this way.

Reviewer: Jean Russell, University of Sheffield



Finally, give the chart a title and change the label on the y axis from 'Count' to 'Percentage'.

When finished, close the chart editor to return to the main output window. Right click on the chart in the output window, copy and paste into word. Sometimes you may need to select 'Copy Special' to move charts.



Pasting as a picture enables easy resizing of graphs/ output in Word.

It is clear from the bar chart that the percentage of those dying increased as class lowered. 38% of passengers in 1st class died compared to 74% in 3rd class.

Tips to give students on reporting data summary

Do not include every possible chart and frequency.

Think back to the key question of interest and answer this question.

Briefly talk about every chart and table you include.

Percentages should be rounded to whole numbers unless you are dealing with very small numbers e.g. 0.01%.

Reviewer: Jean Russell, University of Sheffield

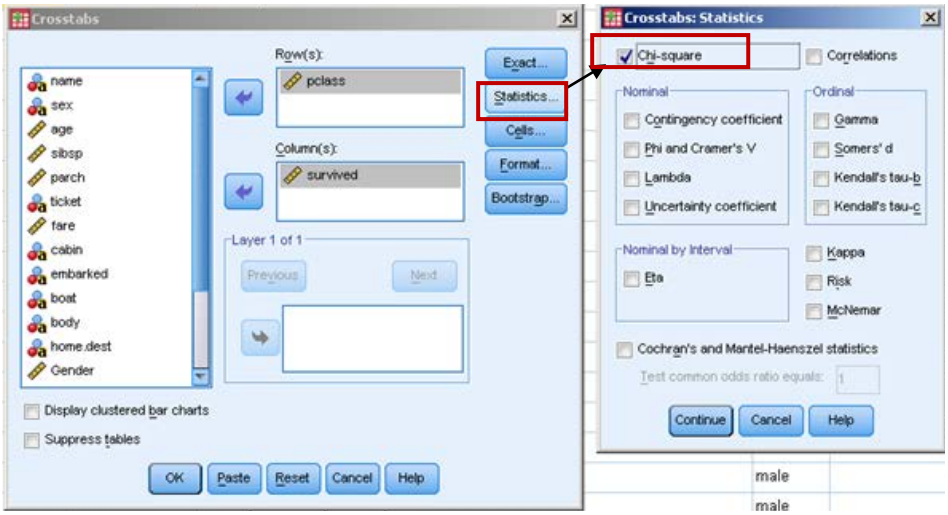
Chi-squared test

Dependent variable: Categorical
Independent variable: Categorical

Uses: Tests the hypothesis that there is no relationship between two categorical variables.
A chi-squared test compares the observed frequencies from the data with frequencies which would be expected if there was no relationship between the variables. A test statistic is calculated based on differences between the observed and expected frequencies.

The Chi-squared test is found within the crosstabs menu.

Analyse → Descriptive statistics → Crosstabs



From the statistics menu, select the Chi-squared test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	127.859[a]	2	.000
Likelihood Ratio	127.765	2	.000
Linear-by-Linear Association	127.709	1	.000
N of Valid Cases	1309		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 105.81.

In all SPSS tests, the p-value is contained in a column containing 'sig'.

Never state the p-value as being 0. Here the p-value is $p < 0.001$

All expected frequencies are above 5.

Reviewer: Jean Russell, University of Sheffield

Exercise 2:Chi-squared test

a) **Interpret the results of the Chi-squared test. If there is evidence of a relationship, what is the relationship?**

b) **What would you do if you had a 2x2 contingency table?**

Assumptions for the Chi-squared test

One of the assumptions for the chi-squared test is that at least 80% of the expected frequencies must be at least 5. SPSS tells you if this is not the case under the Chi-squared output. If it's possible, merge categories with small frequencies and re-run the analysis if SPSS says that over 20% of cells are less than 5. Alternatively, the Exact test could be used if the expected values are small for some categories. This needs to be requested via the Exact menu within Crosstabs.

Adjusting variables

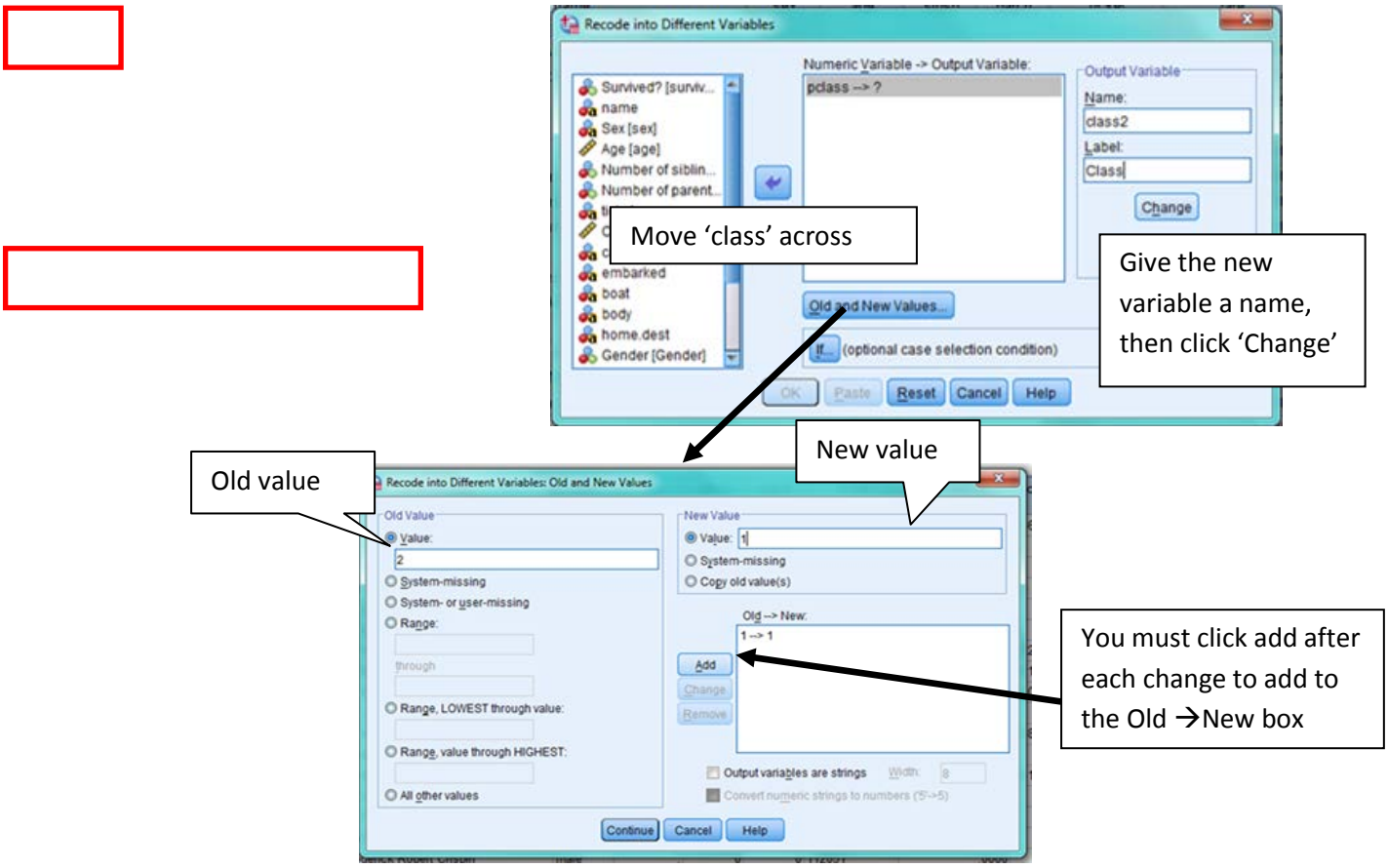
Reducing the number of categories

Sometimes categories can be merged if not all the information is needed. For example, a common summary is to calculate the percentage who agreed from a Likert scale i.e. % agree or strongly agree compared to everything else.

- Use 're-code to different variables' rather than 'Re-code into same variables' so that the re-coding can be checked.
- If there are numerous variables to be recoded in the same way, transfer several variables at the same time. Each variable needs an individual name though. Click change after each new name.

Here a new variable is created where 0 = 3rd class and 1 = 1st or 2nd class.

Transform →Recode into different variables



Select 'Continue' and then 'OK' to produce the new variable. Then label 0 = 3rd class and 1 = 1st or 2nd class in the value label box in variable view. Finally do a cross-tabulation of the old and new variables to check the re-coding is correct.

New variable

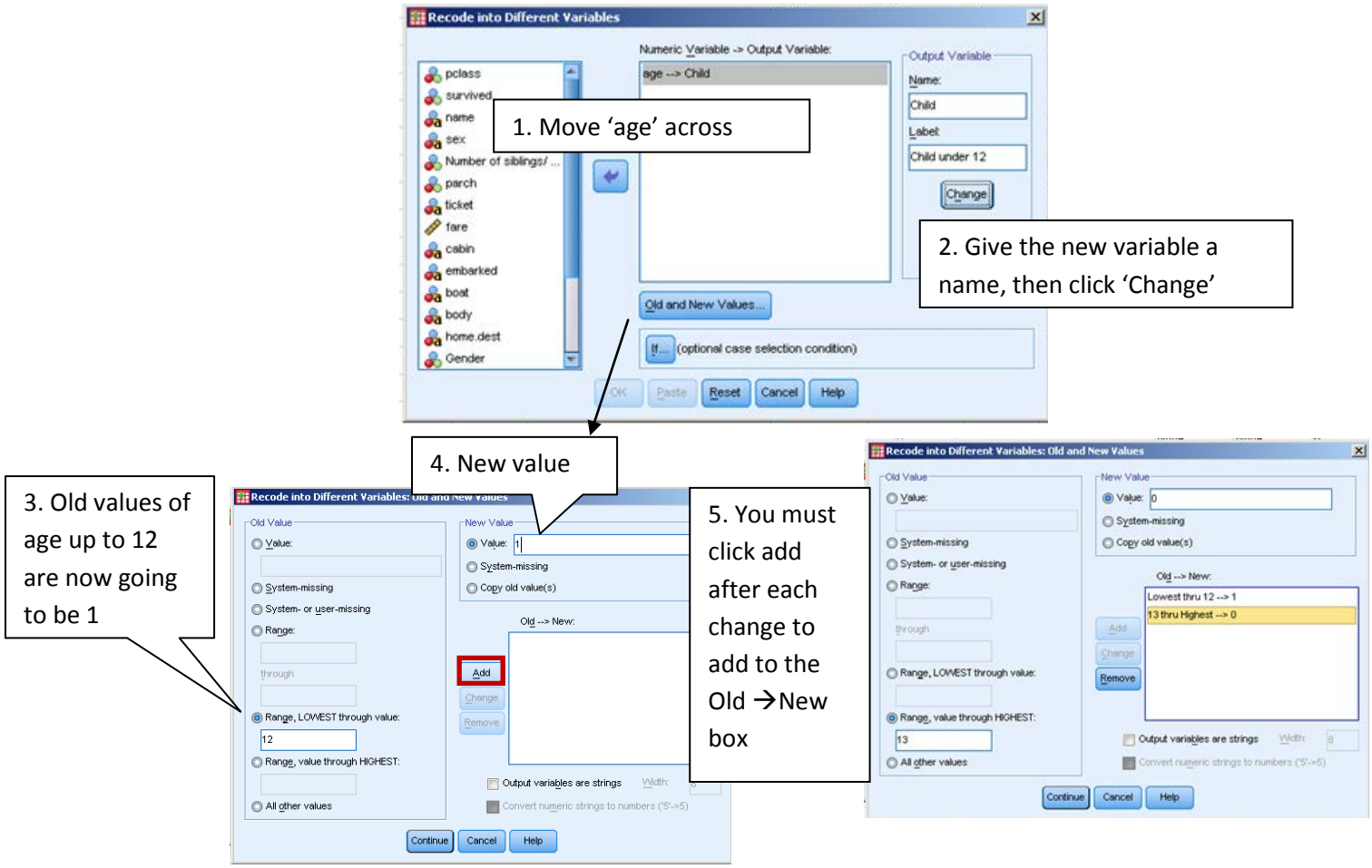
Old variable

		Class		
		3rd class	1st or 2nd class	Total
Class	1st	0	323	323
	2nd	0	277	277
	3rd	709	0	709
Total		709	600	1309

All 1st and 2nd class passengers have been correctly recoded as '1st or 2nd class.

Changing continuous to categorical variables

Although it is not recommended as information is lost, continuous (scale) variables can be categorised. Here we will create a new variable identifying children of 12 and under within the Titanic data set.



Go to variable view and label 0 as 'Adult' and 1 as 'Child'.

Use 'Crosstabs' for the old and new variable to check the re-coding is correct i.e. age vs Child to see all those of 12 and under are classified as a child.

Exercise 3: Nationality and survival

Were Americans more likely to survive than the British? Produce suitable summary statistics/ charts to investigate this and carry out a Chi-squared test.

Null:

Test Statistic

p-value

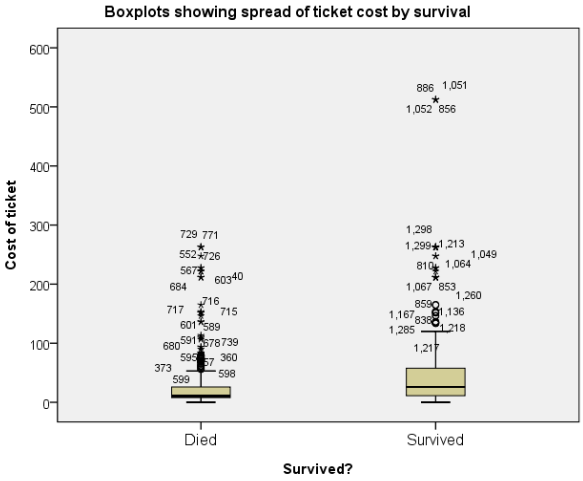
Reject/ do not reject null

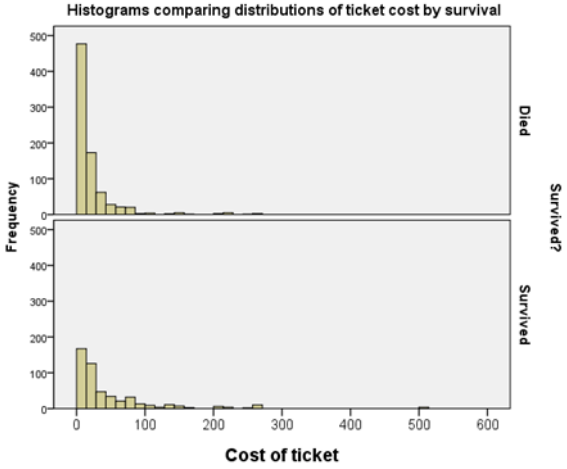
Conclusion

Summary statistics and graphs for continuous data

Did the cost of a ticket affect chances of survival on the Titanic?

Cost of ticket	Survived?	
	Died	Survived
Mean	23.35	49.36
Median	10.50	26.00
Standard Deviation	34.15	68.65
Interquartile range	18.15	46.56
Minimum	0.00	0.00
Maximum	263.00	512.33





Exercise 4: Comparison of continuous data by group

a) Is there a big difference in average ticket price by group?

b) Which group has data which is more spread out?

c) Is the data skewed? How can you tell if the data is skewed?

d) Is the mean or median a better summary measure?

Reviewer: Jean Russell, University of Sheffield

Diet data: The data set 'diet' contains information on 78 people who undertook 1 of three diets. There is background information such as age and gender as well as weights before and after the diet.

Females = 0

Diet 1, 2 or 3

Person	gender	Age	Height	preweight	Diet	weight10weeks
1	0	22	159	58	1	54.2
2	0	46	192	60	1	54.0
3	0	55	170	64	1	63.3
4	0	33	171	64	1	61.1
5	0	50	170	65	1	62.2
6	0	50	201	66	1	64.0
7	0	37	174	67	1	65.0
8	0	28	176	69	1	60.5

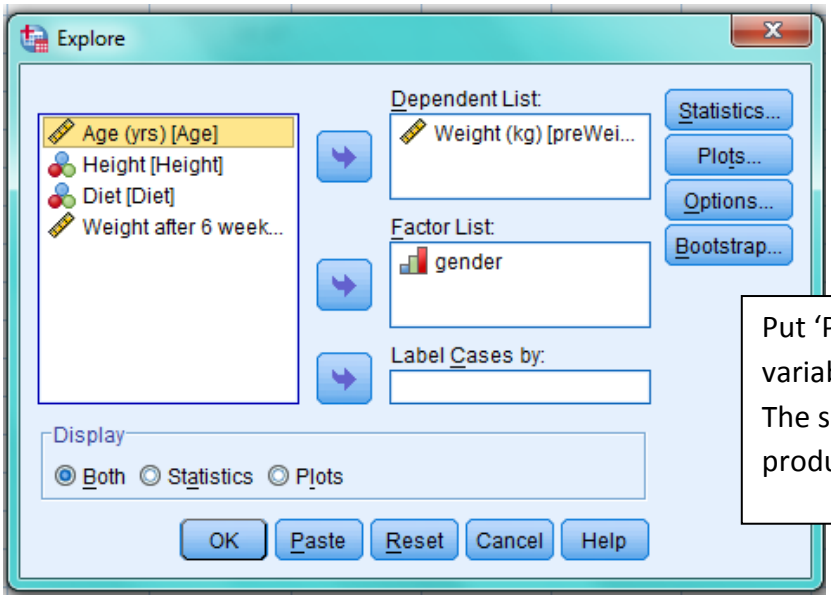
Weight before

Weight after

Open the data set from Excel. Go into the Variable View and make sure that each variable is correctly categorised e.g. nominal. Note: continuous is called 'Scale' in SPSS. It is important that variables are correctly categorised as SPSS will only carry out some analysis on certain variable types.

There are several ways to produce summary statistics and charts. This option uses 'Explore' which contains the most summary statistics to compare weight before the diet for males and females.

Analyse → Descriptive statistics → Explore



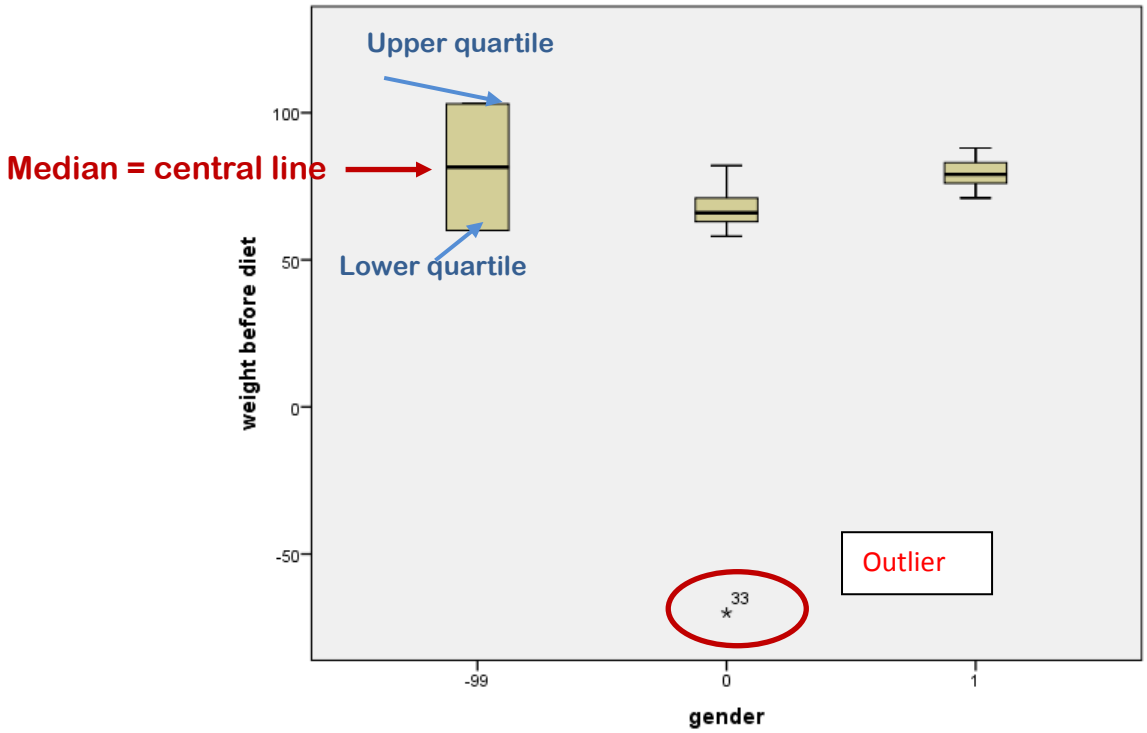
Put 'Pre-weight' as the dependent variable and 'Gender' in the factor list. The summary statistics will be produced for each gender separately.

Exercise 5: Weight before the diet by gender

a) **Fill in the following table using the summary statistics table in the output.**

	Female = 0	Male = 1
Minimum	-70	
Maximum	82	
Mean	64	
Median	66	
Standard Deviation	21.6	

b) **Interpret the summary statistics by gender. Which group has the higher mean and which group is more spread out?**



A box-plot shows the spread of a distribution of values. The box contains the middle 50% of values. SPSS labels outliers with a circle and extreme values with a star.

c) **How could the chart be improved and is there anything odd?**

Reviewer: Jean Russell, University of Sheffield

Research question 2: Which of three diets was best?

Before the next section, change the error of -70 to 70. Outliers should not normally be changed unless they are clearly data entry errors as in this case.

gender	Age	Height	weightbeforediet	Diet
0	45	155	69	3
0	28	176	69	
0	28	165	70	
0	45	165	70	
0	58	141	-70	

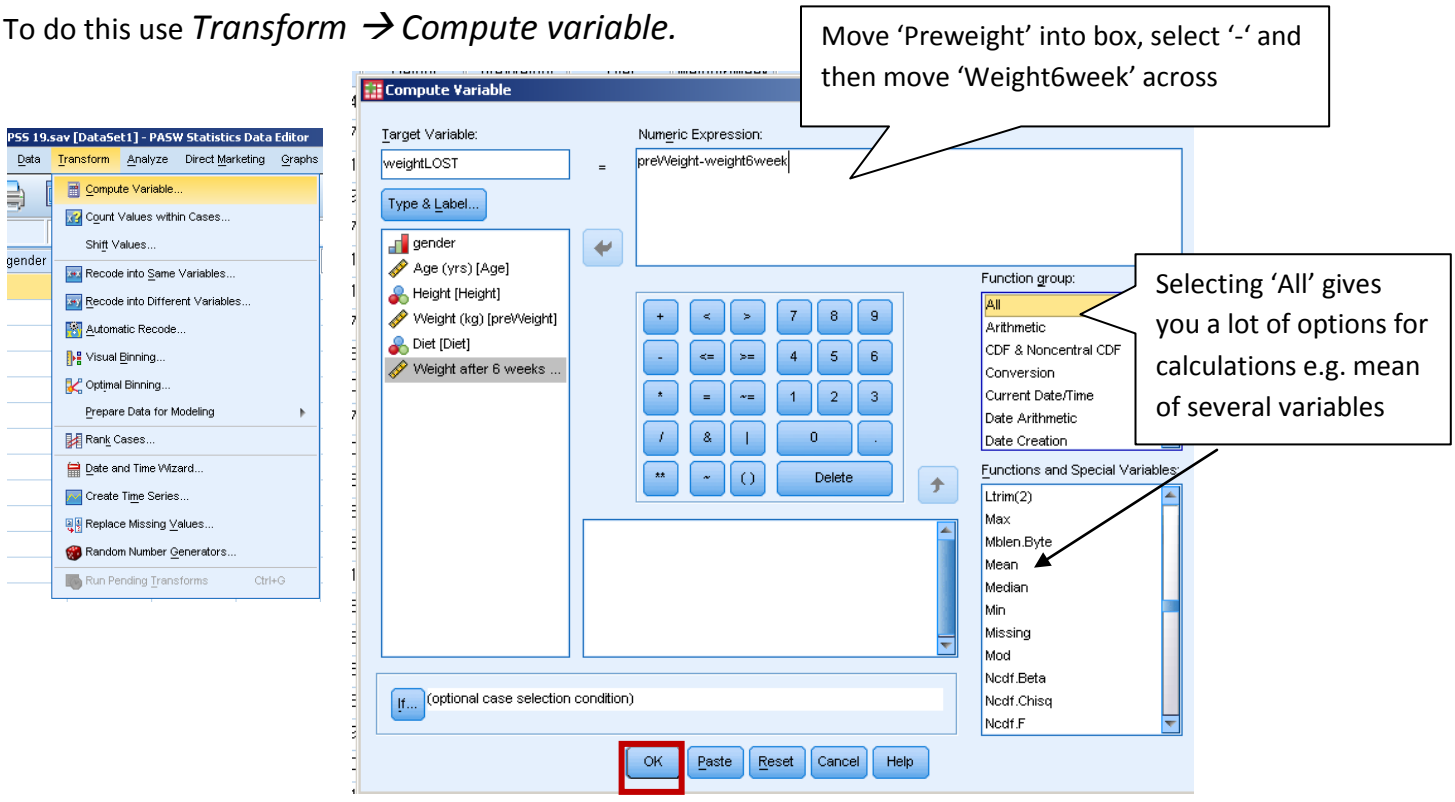
Change -70 to 70kg

Give the variables sensible labels and label gender with 0 = Female and 1 = Male. Tell SPSS that -99 is a missing value.

Calculations using variables

Producing the charts for gender and weight before the diet was useful for demonstrating SPSS but the main question of interest is **'Which diet led to greater weight loss?'**. How could this be assessed?
To answer this, a new variable 'weight lost' (weight before – weight after) would be useful. As spaces are not allowed in variable names, use weightLOST as a name and give a better name in the label section in variable view.

To do this use *Transform → Compute variable.*



After putting the calculation into the 'Numeric Expression' box, select 'OK' and the new variable will appear last in the Data and variable view sheets. Before carrying out the official test of a difference, use summary statistics and charts to look at the differences.

Producing tables in SPSS

SPSS has a table function which can produce more complicated tables although it is a little temperamental and frustrating at times!

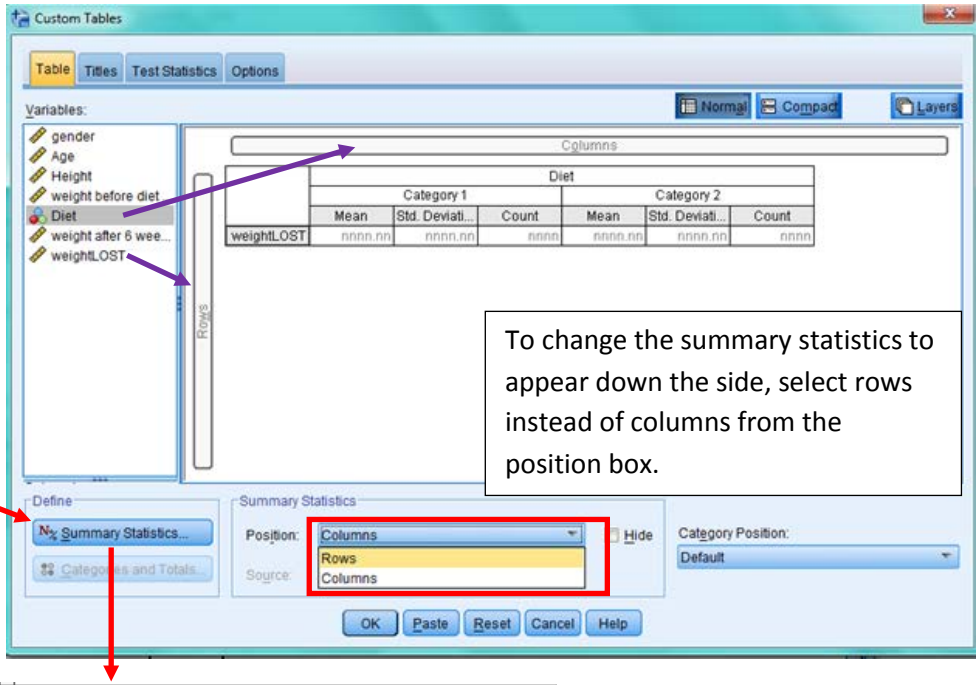
To open the table window: *Analyse → Tables → Custom Tables*

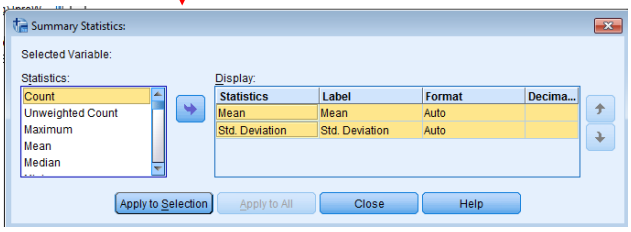
Drag variables to either the row or column bars to include them in the table.
To create sub categories, drag the categorical variable to the front of the variable already in the table.
By default, SPSS will choose means to summarise continuous (scale) variables and counts to summarise categorical variables. It is vital that variables are correctly defined as scale or categorical.

1) Move 'WeightLOST' to the row section and 'Diet' to the Columns section.
2) Select the summary statistics you require
3) Choose 'Columns' in the 'Position' options for a better display.

Selecting the 'Summary Statistics' button opens a window where options for statistics displayed can be chosen.

The summary statistics button will only highlight when a variable is selected in the main window. Here, make sure weightLOST is highlighted in yellow in the central window.





Select Standard deviation and count from the options and click 'Apply to all'.

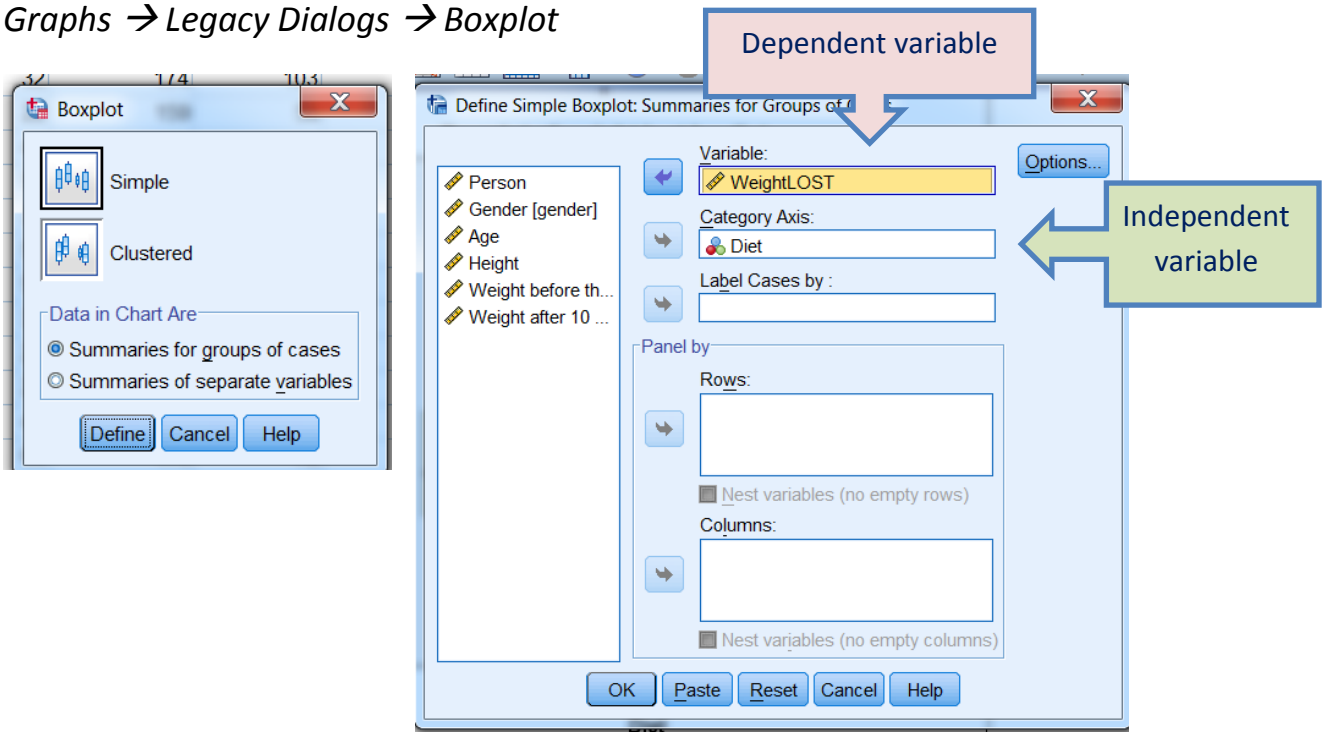
		Diet		
		1	2	3
Weight lost on diet (kg)	Mean	3.30	3.03	5.15
	Standard Deviation	2.24	2.52	2.40
	Count	24	27	27

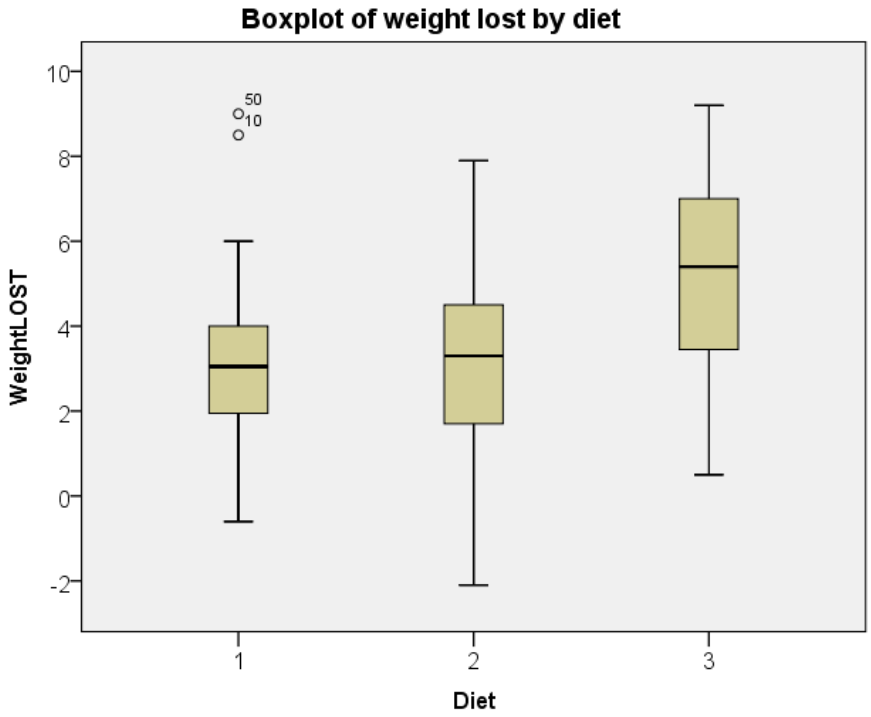
Which diet seems the best and which diet has the most variation in weight loss?

Summary statistics and charts can also be produced separately by group using the split file option in *Data → Split File* but remember to un-split the file when you have finished.

Box-plots

Graphs → Legacy Dialogs → Boxplot





Exercise 6: Confidence intervals

Use Explore to get the confidence intervals of the mean weight lost by diet.

Diet	Mean	95% Confidence interval for the mean
1	3.3	
2	3.03	
3	5.15	

What is the correct definition of a confidence interval for the mean?

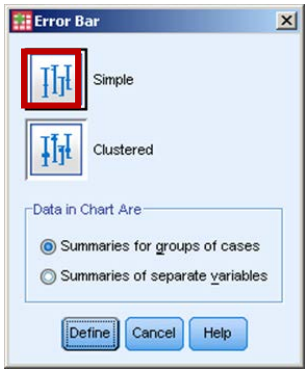
How would you explain a confidence interval to a student?

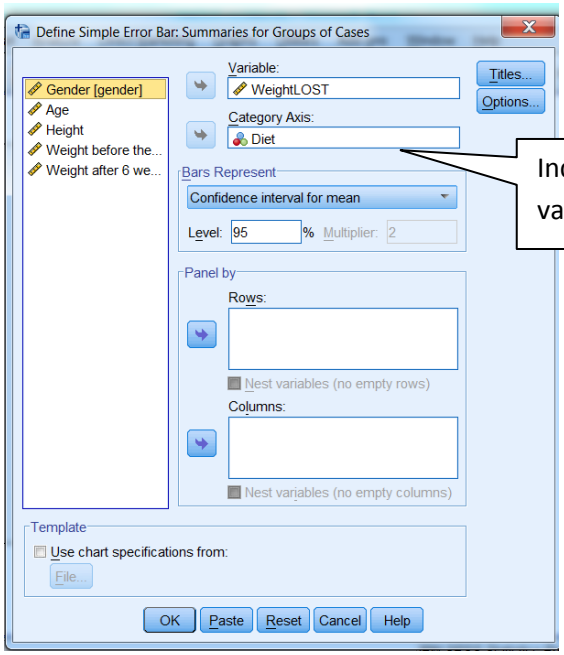
Reviewer: Jean Russell, University of Sheffield

Confidence Interval plot

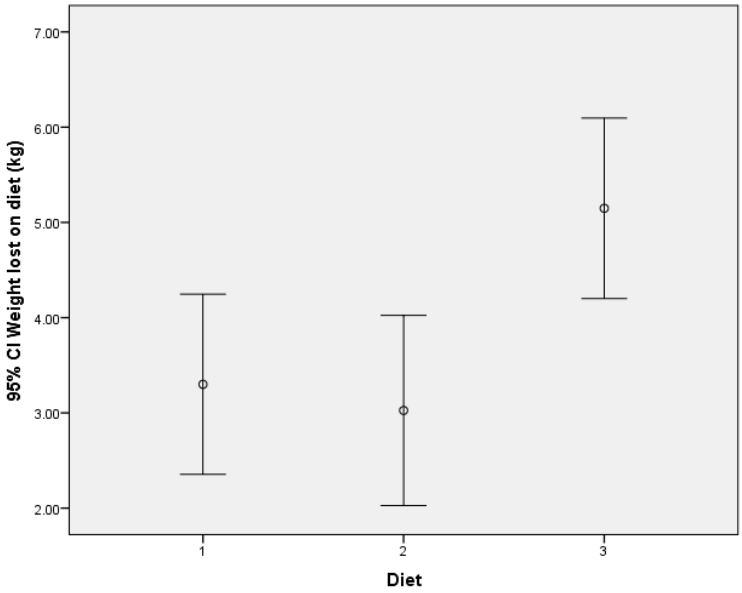
When comparing the means of several groups, a plot of confidence intervals by group is useful.

Graphs → Legacy Dialogs → Error Bar





Independent grouping variable



For more information on reading Confidence interval plots, see 'Inference by Eye'
http://www.apastyle.org/manual/related/cumming-and-finch.pdf

ANOVA (Analysis of variance)

Dependent variable: Scale

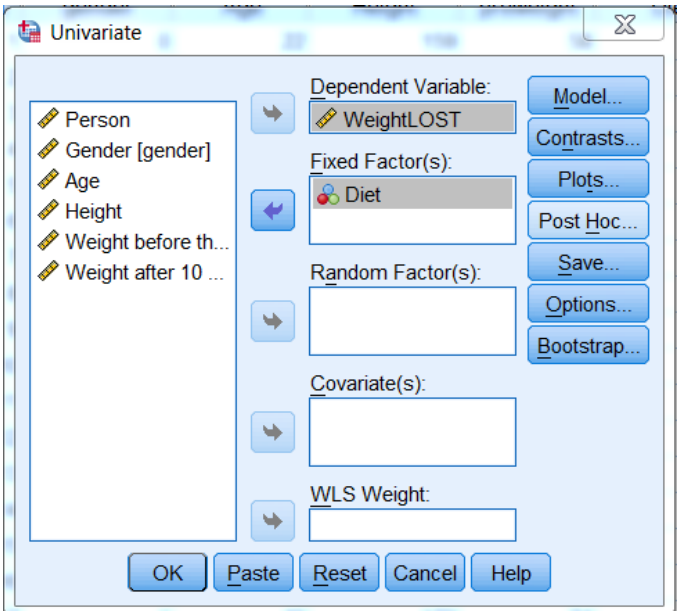
Independent variable: Categorical

Exercise 7: ANOVA and assumptions

a) **Explain briefly why ANOVA is called Analysis of variance instead of Analysis of means.**

b) **What are the assumptions for ANOVA and how can they be tested?**

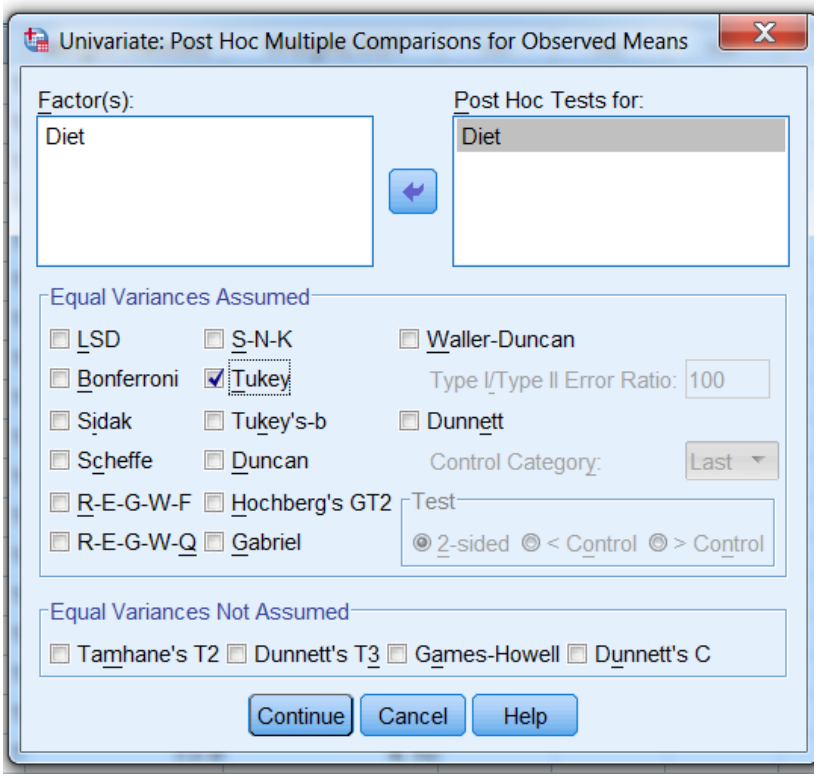
To carry out an ANOVA, select *ANALYZE → General Linear Model → Univariate*



Put the dependent variable (weight lost) in the dependent variable box and the independent variable (diet) in the 'Fixed Factors' box.

The post hoc window

Move the Factor of interest to the Post hoc box at the top, then choose from the available tests. Tukey's and Scheffe's tests are the most commonly used post hoc tests. Hochberg's GT2 is better where the sample sizes for the groups are very different. If the Levene's test concludes that there is a difference between group variances, use the Games-Howell test.



Reviewer: Jean Russell, University of Sheffield

The ANOVA table:

Tests of Between-Subjects Effects

Dependent Variable: Weight lost on diet (kg)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	71.094[a]	2	35.547	6.197	.003
Intercept	1137.494	1	1137.494	198.317	.000
Diet	71.094	2	35.547	6.197	.003
Error	430.179	75	5.736		
Total	1654.350	78			
Corrected Total	501.273	77			

a. R Squared = .142 (Adjusted R Squared = .119)

The post hoc tests:

Multiple Comparisons

Dependent Variable: Weight lost on diet (kg)

	(I) Diet	(J) Diet	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	1	2	.2741	.67188	.912	-1.3325	1.8806
		3	-1.8481*	.67188	.020	-3.4547	-.2416
	2	1	-.2741	.67188	.912	-1.8806	1.3325
		3	-2.1222*	.65182	.005	-3.6808	-.5636
	3	1	1.8481*	.67188	.020	.2416	3.4547
		2	2.1222*	.65182	.005	.5636	3.6808

Exercise 8: Interpret the ANOVA and post hoc output

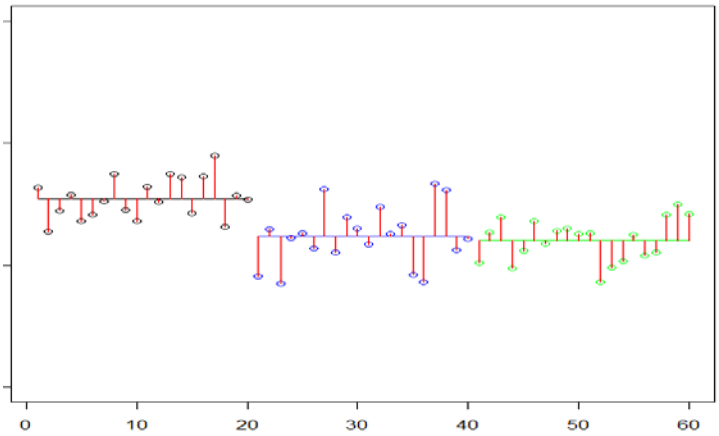
Assumptions for ANOVA:

Assumption	How to check	What to do if assumption not met
Normality: The residuals (difference between observed and expected values) should be normally distributed	Histograms/ QQ plots/ normality tests of residuals	Do a Kruskall-Wallis test which is non-parametric (does not assume normality)
Homogeneity of variance (each group should have a similar standard deviation)	Levene's test	a) Welch test instead of ANOVA and Games-Howell for post hoc or b) Kruskall-Wallis

There are two ways of carrying out a one-way ANOVA (One-way ANOVA and GLM Univariate) but both have something missing. The One-way ANOVA does not produced the residuals needed to check normality and the GLM does not carry out the Welch test. Use the GLM method unless the Welch test is needed.

Normality of residuals

The residuals are the differences between the weight lost by subject and their group mean:



Check they are normally distributed by plotting a histogram. Histograms should peak roughly in the middle and be approximately symmetrical.

There are official tests for normality such as the Shapiro-Wilk and Kolmogorov-Smirnoff tests
If $p > 0.05$, normality can be assumed
Use them with caution:

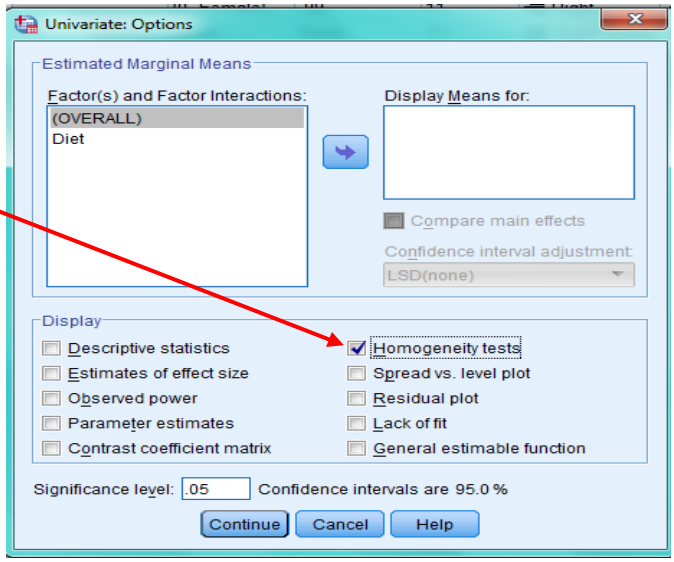
- For small sample sizes ($n < 20$), the tests are unlikely to detect non-normality
- For larger sample sizes ($n > 50$), the tests can be too sensitive
- Very sensitive to outliers

Re-run the ANOVA with the following adjustments

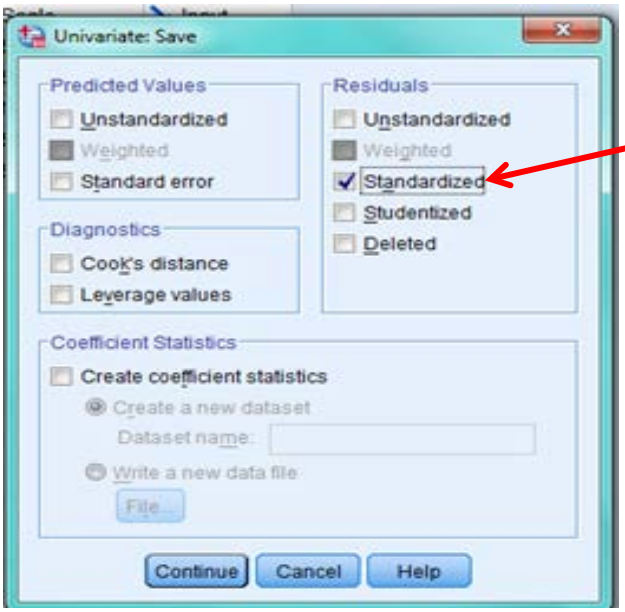
The options window

Testing the assumption of homogeneity:

One of the assumptions for ANOVA is that the group variances should be similar. The Levene's test is carried out if the 'Homogeneity of variance test' option is selected. If the assumption is violated, the Welch test is more appropriate. This can be accessed via ANALYSE → Compare Means → One-way ANOVA.



The 'Save' window



Testing the assumption of normality:

One of the assumptions for ANOVA is that the residuals should be normally distributed. To do this a residual for each observation needs to be produced (individual score – group mean). There are several types of residuals but the standardised residuals will be used here. Outliers are outside ± 3. A new column containing the residuals will be added to the data set.

Exercise 9: Checking ANOVA assumptions

Check the assumptions of normality and Homogeneity of variance using the output

The output

Testing the assumption of homogeneity:

Levene's Test of Equality of Error Variances[a]

Dependent Variable: Weight lost on diet (kg)

F	df1	df2	Sig.
.659	2	75	.520

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + Diet

If sig (the p-value) < 0.05, the assumption of equal variances has not been met. If this is the case, use the Welch test instead of the ANOVA (only available in the Analyse → Compare means → One-way ANOVA method) and Games Howell post hoc tests or a non-parametric test (Kruskall-Wallis)

Can equal variances be assumed?

Null:

Alternative:

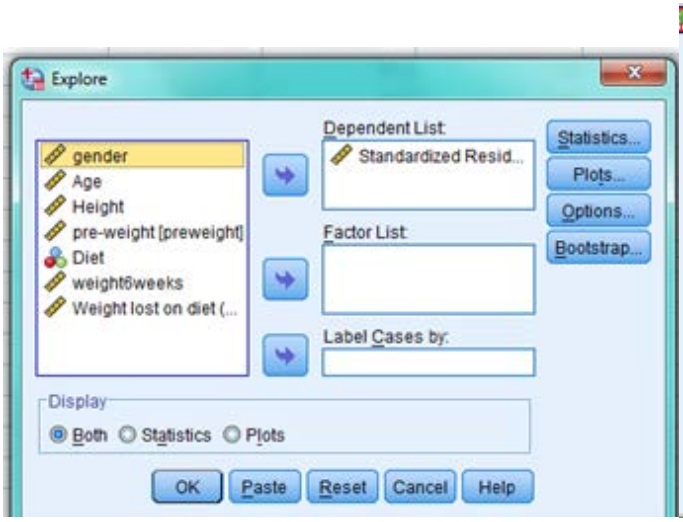
P – value:

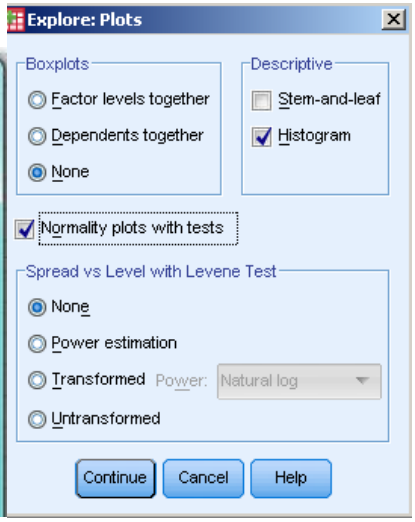
Reject / do not reject null

Checking the assumption of normality

Produce a histogram for the residuals using Explore. Generally, to check the assumption of normality use

ANALYZE → DESCRIPTIVE STATISTICS → EXPLORE





Select the options for Histogram and normality plots with tests from the plots option.

Are the residuals normally distributed?

Reporting ANOVA

When writing up the results of an ANOVA, check papers from the students' discipline as reporting can vary. Generally, it is common to present certain figures from the main ANOVA table.

$F(df_{between}, df_{error})$= Test Statistic, p =

F(2, 75)= 6.197, p =0.03

Tests of Between-Subjects Effects

Dependent Variable: Weight lost on diet (kg)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	71.094[a]	2	35.547	6.197	.003
Intercept	1137.494	1	1137.494	198.317	.000
Diet	71.094	2	35.547	6.197	.003
Error	430.179	75	5.736		
Total	1654.350	78			
Corrected Total	501.273	77			

a. R Squared = .142 (Adjusted R Squared = .119)

A one-way ANOVA was conducted to compare the effectiveness of three diets. Normality checks and Levene's test were carried out and the assumptions met.

There was a significant difference in mean weight lost [F(2,75)=6.197, p = 0.003] between the diets.

Post hoc comparisons using the Tukey HSD test were carried out. There was a significant difference (p = 0.02) between diet 1 (M = 3.3, SD = 2.24) and diet 3 (M = 5.15, SD = 2.4) and a significant difference (p = 0.005) between diet 2 (M = 3, SD = 2.52) and diet 3 but not between diets 1 and 2.

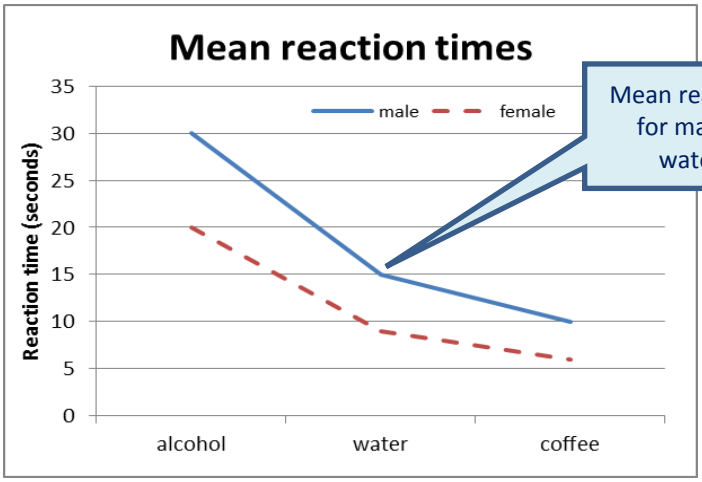
Summarising the effect of two categorical variables on one independent variable

A line chart can be used to compare the means of combinations of two independent variables. It is particularly useful for looking at interaction effects and can also be called an interaction plot or means plot. The lines connect means of each combination.

Example: An experiment was carried out to investigate the effect of drink on reaction times in a driving simulator. Participants were given alcohol, water or coffee. The mean reaction times by group are contained in the table below:

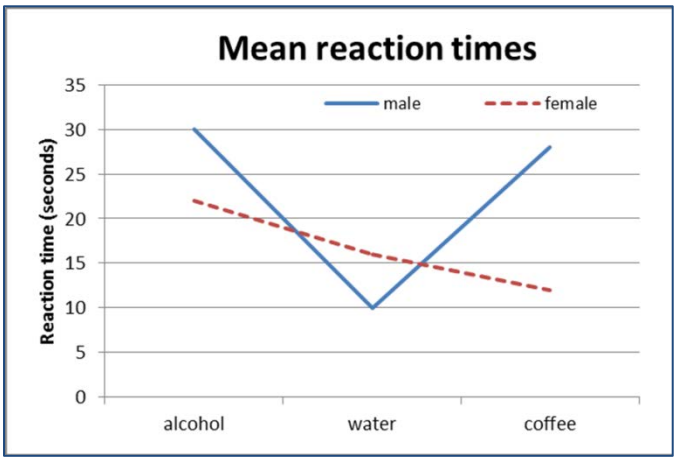
Mean Reaction Times	Male	Female
Alcohol	30	20
Water	15	9
Coffee	10	6

The six means can be displayed in a line/ means plot:



For both males and females, the fastest (i.e. lowest) reactions times are after coffee, followed by water then alcohol. Females are faster than males after all three drinks. There is no interaction between gender and drink as the lines are reasonably parallel.

What does an interaction look like?



An interaction occurs when the lines are not quite so parallel; such the means of one group do not follow the same pattern as the other group. Here males have their fastest reaction after water, but females have their fastest reaction after coffee. Males are faster than females after water but females are faster after coffee and alcohol.

Reviewer: Jean Russell, University of Sheffield

Two way ANOVA

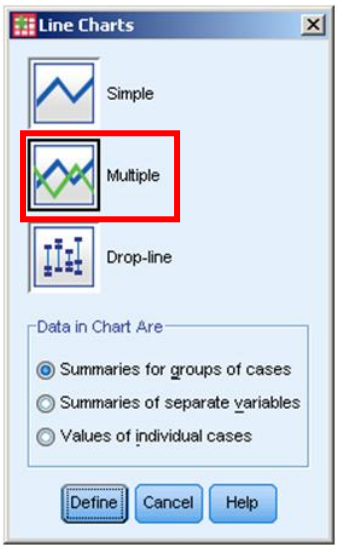
Two way ANOVA has two categorical independent variables and tests three hypotheses.

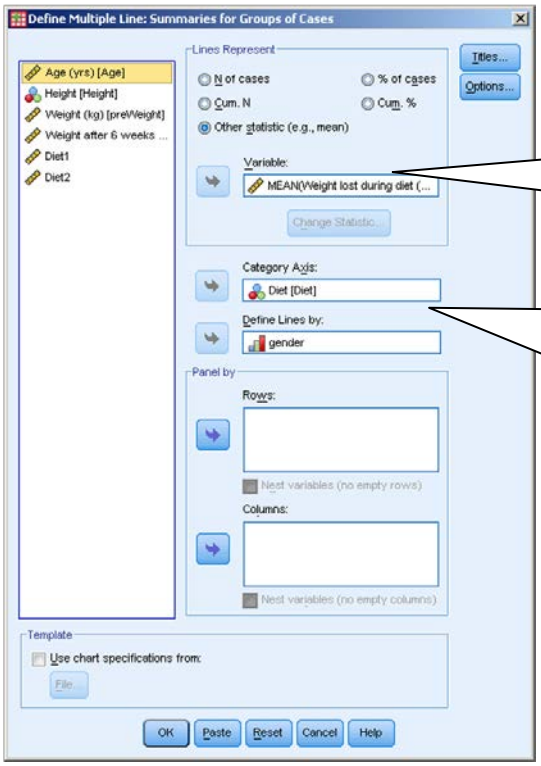
It tests the two main effects of each independent variable separately and also whether there is an interaction between them.

A means plot should be used to check for an interaction between the two independent variables.

This example compares weight lost by diet and gender.

Graphs → Legacy Dialogs → Line

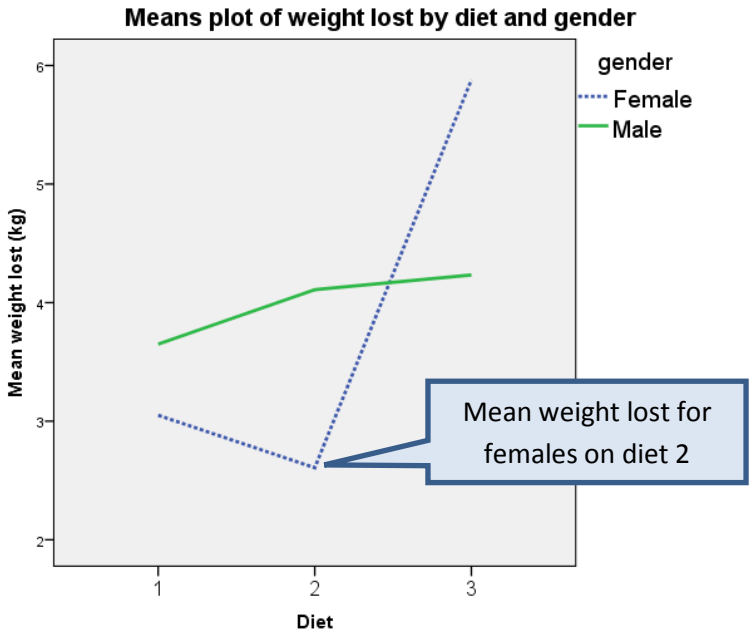




Select the lines represent 'other' category, choose 'mean' and move the dependent variable across

Move the two categorical independent variables to the 'Category axis' and 'Define lines by' boxes. The x-axis will be the category axis option. Think carefully about which way round to have the two variables

Quick question: Is there an interaction between gender and diet when it comes to weight lost?



What if there is a significant interaction?

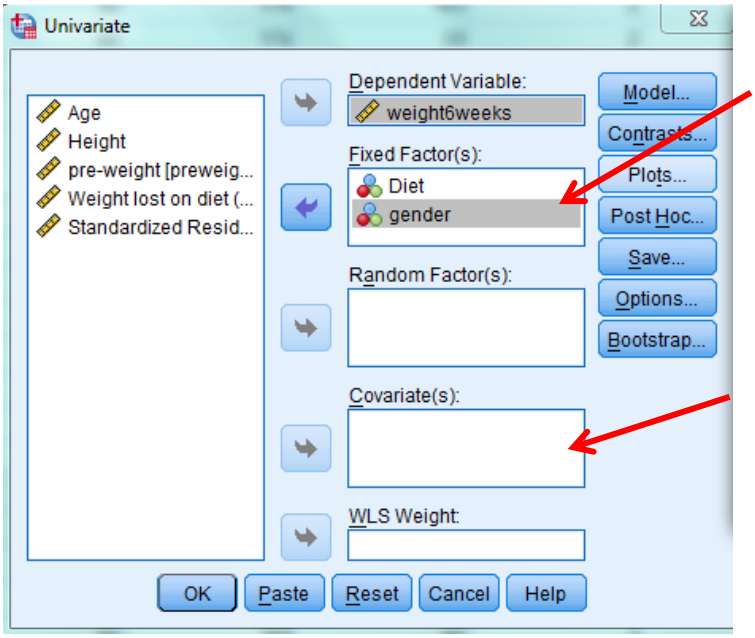
The main effects need to be discussed by group e.g. for males/ females separately. The interaction can be described using the means plot but separate ANOVA's can be carried out by group e.g. testing diet by gender.

Exercise 10: Two-way ANOVA

Carry out a two-way ANOVA. Report on the results of the following tests:

1. **Is there a gender effect?**
2. **Is there diet effect**
3. **Is there an interaction between gender and diet?**

If there is a significant interaction, split the data file and carry out an ANOVA of diet for each gender as SPSS does not automatically do this for significant interactions.

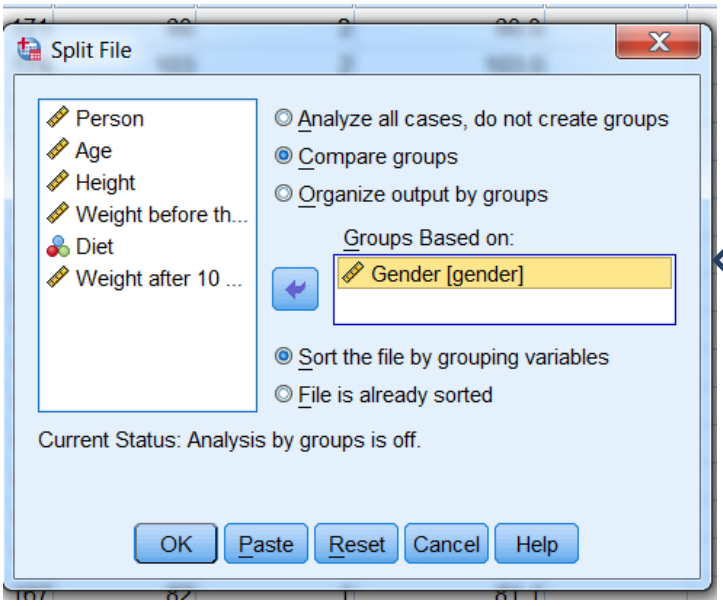


Categorical independent variables are added in the 'Fixed Factors' box.

ANCOVA is used when there is one or more continuous independent variables which are added in the 'Covariates' box.

Splitting a file

To carry out separate analysis by category: *Data → Split file*

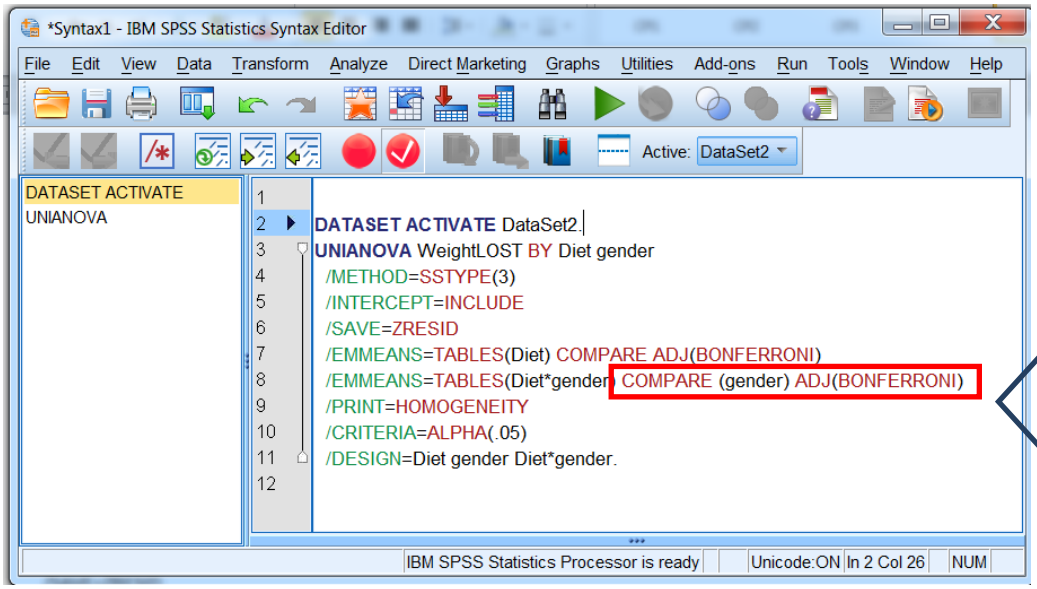


Select compare groups and then move the factor to the 'Groups Based on' box

Note: You will need to go back to split file after the analysis and select 'Analyse all cases, do not create groups' or all further analysis will be carried out separately by group.

Syntax

Note: There is an adjustment that can be made to the syntax of a two way ANOVA so that post hoc tests are carried out on the interactions but only demonstrate this if the student can cope with programming. The syntax is the program SPSS uses to run the analysis and can be requested for any procedure by selecting 'Paste' instead of 'Ok' after selecting all the required options. It is very useful for students doing a lot of recoding as it is a record of what has be done and can be used to repeat analysis at a future date. Once in the syntax window, click on the green arrow to run the program. Below is the syntax for a two way ANOVA. The adjustment made to the syntax is highlighted. The EMMEANS line comes from selecting post hoc tests from the options menu within the ANOVA procedure.



The 'COMPARE (gender) ADJ(BONFERRONI) will carry out post hoc tests for gender for each diet separately. Putting (Diet) after compare will produce diet comparisons for each gender

Pairwise Comparisons

Dependent Variable: WeightLOST

Diet	(I) Gender	(J) Gender	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
1	Female	Male	-.600	.960	.534	-2.515	1.315
	Male	Female	.600	.960	.534	-1.315	2.515
2	Female	Male	-1.502	.934	.112	-3.365	.361
	Male	Female	1.502	.934	.112	-.361	3.365
3	Female	Male	1.647	.898	.071	-.144	3.438
	Male	Female	-1.647	.898	.071	-3.438	.144

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Reviewer: Jean Russell, University of Sheffield

Non-parametric tests

Exercise 11: Non-parametric tests

How would you explain what non-parametric means to a student?

What should be checked for normality for the following tests and what is the equivalent non-parametric test:

Key non-parametric tests

Parametric test	What to check for normality	Non-parametric test
Independent t-test		
Paired t-test		
One-way ANOVA		
Repeated measures ANOVA		
Pearson's Correlation Co-efficient		
Simple Linear Regression		

Reviewer: Jean Russell, University of Sheffield

Kruskal-Wallis

(Non-parametric equivalent to ANOVA)

Research question type: Differences between several groups of measurements

Dependent variable: Continuous/ scale/ ordinal but not normally distributed
Independent variable: Categorical
Common Applications: Comparing the mean rank of three or more different groups in scientific or medical experiments when the dependent variable is not normally distributed.

Descriptive statistics: Median for each group and box-plot

Example: Alcohol, coffee and reaction times

Reaction time		
Water	Coffee	Alcohol
0.37	0.98	1.69
0.38	1.11	1.71
0.61	1.27	1.75
0.78	1.32	1.83
0.83	1.44	1.97
0.86	1.45	2.53
0.9	1.46	2.66
0.95	1.76	2.91
1.63	2.56	3.28
1.97	3.07	3.47

An experiment was carried out to see if alcohol or coffee affects driving reaction times. There were three groups of participants; 10 drinking water, 10 drinking beer containing two units of alcohol and 10 drinking coffee. The reaction time on a driving simulation was measured.

Exercise 12: Kruskal-Wallis

Enter the data into a new SPSS sheet in a suitable way to be analysed using ANOVA/ Kruskall-Wallis. Carry out a one-way ANOVA and check the assumptions. Have the assumptions been met?

Produce suitable summary statistics and follow the instructions below to perform the Kruskall-Wallis test. Hint: One row per person

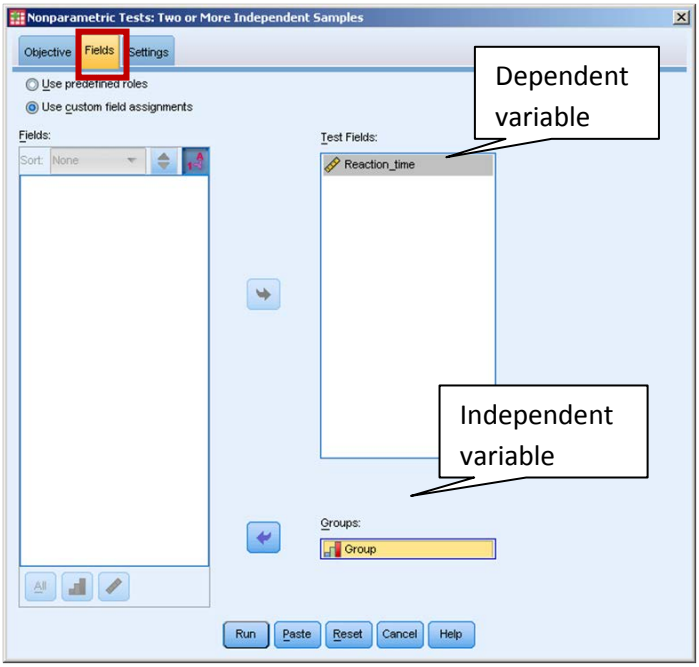
The Kruskal-Wallis test ranks the scores for the whole sample and then compares the mean rank for each group.

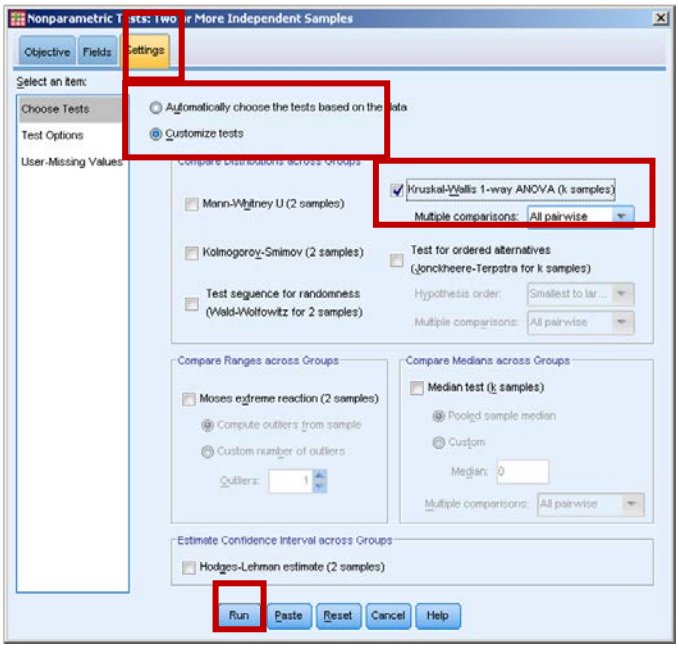
To carry out the Kruskal-Wallis test:

Analyse → Nonparametric Tests → Independent samples

In the *Fields* tab, move the dependent variable to the 'Test Field' box and the grouping factor to the 'Groups' box. **Note: The dependent variable has to be classified as Scale to perform the analysis even if it's actually ordinal data.**

In the *Settings* tab choose to customise the tests and then select the Kruskal-Wallis test. Leave the multiple comparisons as 'All pairwise' and 'Run'.





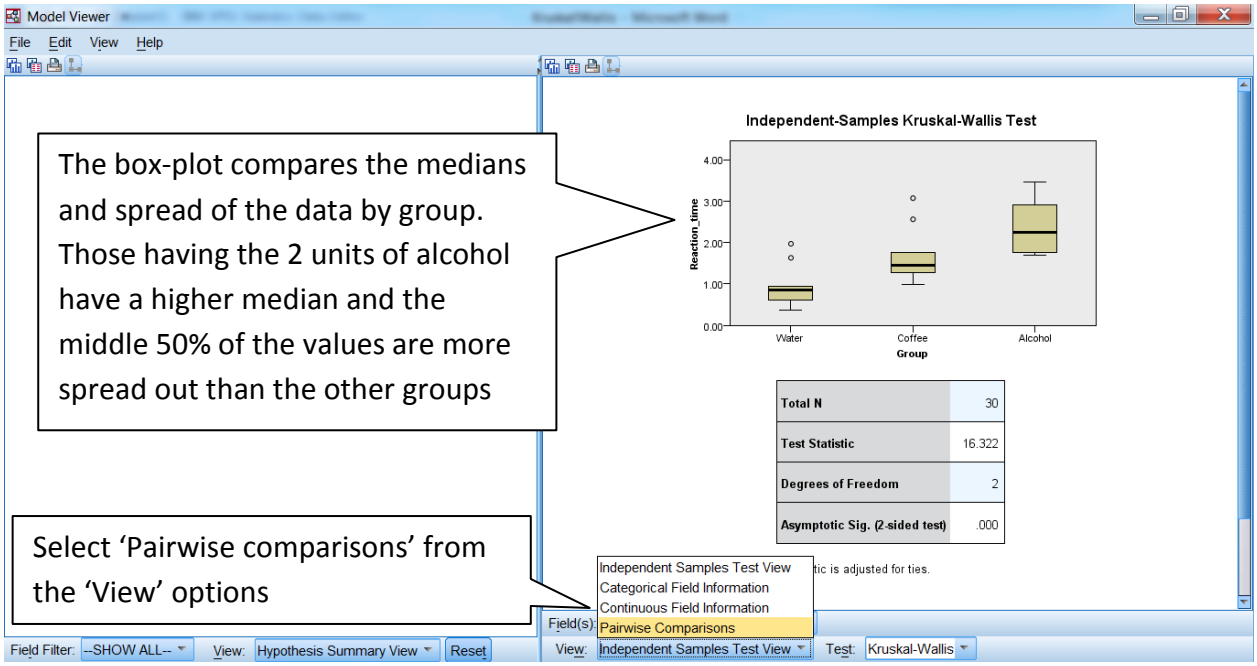
This gives the following basic output for the Kruskal-Wallis test:

Hypothesis Test Summary

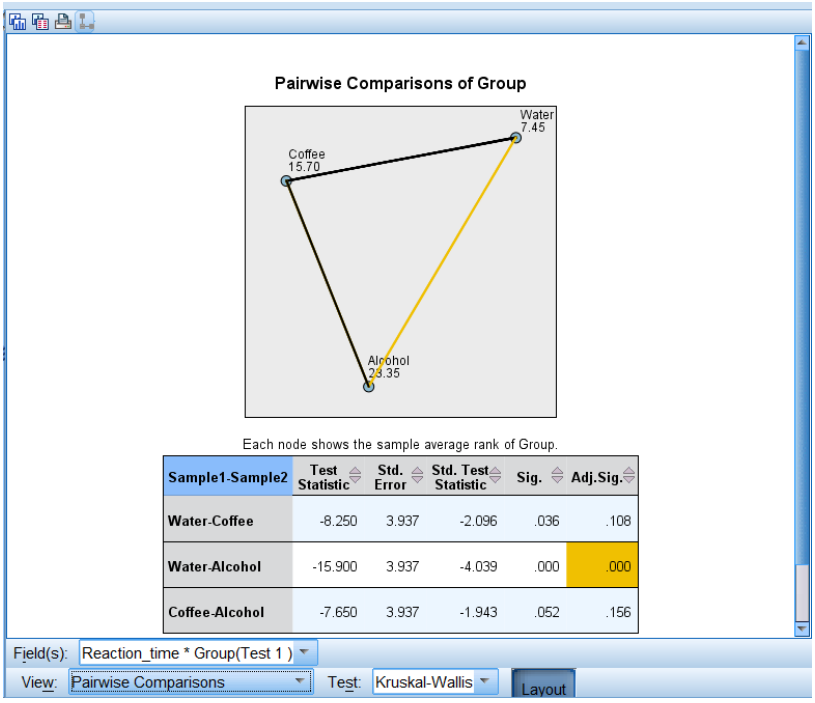
	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Reaction_time is the same across categories of Group.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

As $p < 0.001$, there is very strong evidence to suggest a difference between at least one pair of groups but which pairs? To find out, **double click** on the output to open this additional screen. Change the 'Independent Samples Test View' to 'Pairwise comparisons' in the bottom right corner. Note: The pairwise comparisons are only available when the initial test result is significant.



Mann-Whitney tests are carried out on each pair of groups. As multiple tests are being carried out, SPSS makes an adjustment to the p-value. The adjustment is to multiply each Mann-Whitney p-value by the total number of Mann-Whitney tests being carried out (Bonferroni correction).



Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
Water-Coffee	-8.250	3.937	-2.096	.036	.108
Water-Alcohol	-15.900	3.937	-4.039	.000	.000
Coffee-Alcohol	-7.650	3.937	-1.943	.052	.156

The pairwise comparisons page shows the results of the Mann-Whitney tests on each pair of groups. The Adj. Sig column makes the adjustments for multiple testing. Only the p-value for the test comparing the placebo and alcohol groups is significant ($p < 0.001$).

Significant differences are also highlighted using an orange line to join the two different groups in the diagram which shows the mean rank for each group.

Reporting the results

A Kruskal-Wallis test provided strong evidence of a significant difference ($p<0.001$) between the mean ranks of at least one pair of groups. Mann-Whitney pairwise tests were carried out for the three pairs of groups. There was a significant difference between the group who had the water and those who had the beer with two units of alcohol. The median reaction time for the group having water was 0.85 seconds compared to 2.25 seconds in the group consuming two units of alcohol.

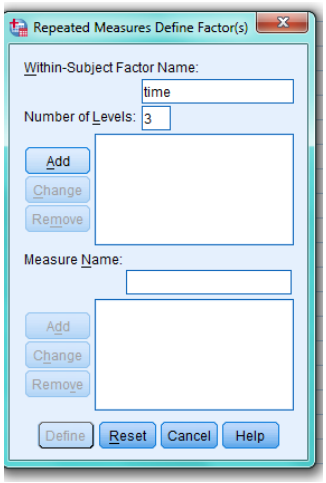
Research question 3: Does Clora margarine reduce cholesterol?

Cholesterol data: Participants used Clora margarine for 8 weeks. Their cholesterol was measured before the special diet, after 4 weeks and after 8 weeks. Open the Excel sheet 'Cholesterol' and follow the instructions to see if the using the margarine has changed the mean cholesterol.

Repeated measures ANOVA

To carry out a repeated measures ANOVA, use

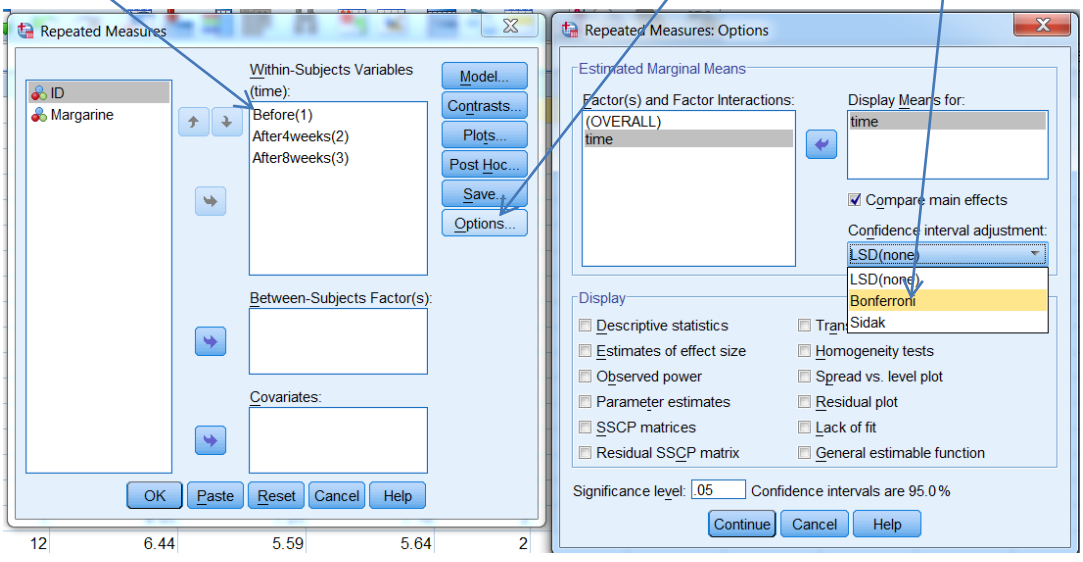
Analyse → General Linear Model → Repeated measures.



This screen comes up first. This is where we define the levels of our repeated measures factor which in our case is time. We need to name it using whatever name we like (we have used "time" in this case) and then state how many time points there are (which here is 3; before the experiment, after 4 weeks and after 8 weeks). Make sure in your data set there is one row per person and a separate column for each of the three time points.

Make sure you click on the Add button and then click on the Define button.

The next screen you see should like that below. Move the three variables across into the within-subjects box. Post hoc tests for repeated measures are in 'Options'. Choose Bonferroni from the three options.



In the 'Save' menu, ask for the standardised residuals. A set of residuals will be produced for each time point which should then be checked using 'Explore'.

Two way repeated measures ANOVA is also possible as well as 'Mixed ANOVA' with some between-subject and within-subject measures. The 'Post hoc' section is for between-subject factors when running a 'Mixed Model' with between-subject and within-subject factors.

Reviewer: Jean Russell, University of Sheffield

The output

One of the assumptions for repeated measures ANOVA is the assumption of Sphericity which is similar to the assumption of equal variances in standard ANOVA. The assumption is that the variances of the differences between all combinations of the related conditions/ time points are equal, although it is a little more than this and relates to the variance-covariance matrix but we won't go into that here!

Mauchly's Test of Sphericity[a]

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
time	.381	15.440	2	.000	.618	.642	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept
Within Subjects Design: time

b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

The test above is significant so the assumption of Sphericity has not been met. If Sphericity can be assumed, use the top row of the 'Tests of Within-Subjects Effects' below. If it cannot be assumed, use the Greenhouse-Geisser row (as shown below) which makes an adjustment to the degrees of freedom.

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
time	Sphericity Assumed	4.320	2	2.160	212.321	.000
	Greenhouse-Geisser	4.320	1.235	3.497	212.321	.000
	Huynh-Feldt	4.320	1.284	3.365	212.321	.000
	Lower-bound	4.320	1.000	4.320	212.321	.000
Error(time)	Sphericity Assumed	.346	34	.010		
	Greenhouse-Geisser	.346	21.001	.016		
	Huynh-Feldt	.346	21.822	.016		
	Lower-bound	.346	17.000	.020		

As $p < 0.001$, there's a difference in cholesterol between at least 2 time points

The post hoc tests

Pairwise Comparisons

Measure: MEASURE_1

(I) time	(J) time	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
1	2	.566*	.037	.000	.469	.663
	3	.629*	.042	.000	.517	.741
2	1	-.566*	.037	.000	-.663	-.469
	3	.063*	.017	.004	.019	.107
3	1	-.629*	.042	.000	-.741	-.517
	2	-.063*	.017	.004	-.107	-.019

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Exercise 13: Repeated measures example

Interpret the post hoc tests and check the assumption of normality. Does the change in mean cholesterol look meaningful?

Do the residuals at each time point look normally distributed?

What test would you use instead if the assumption of normality has not been met?

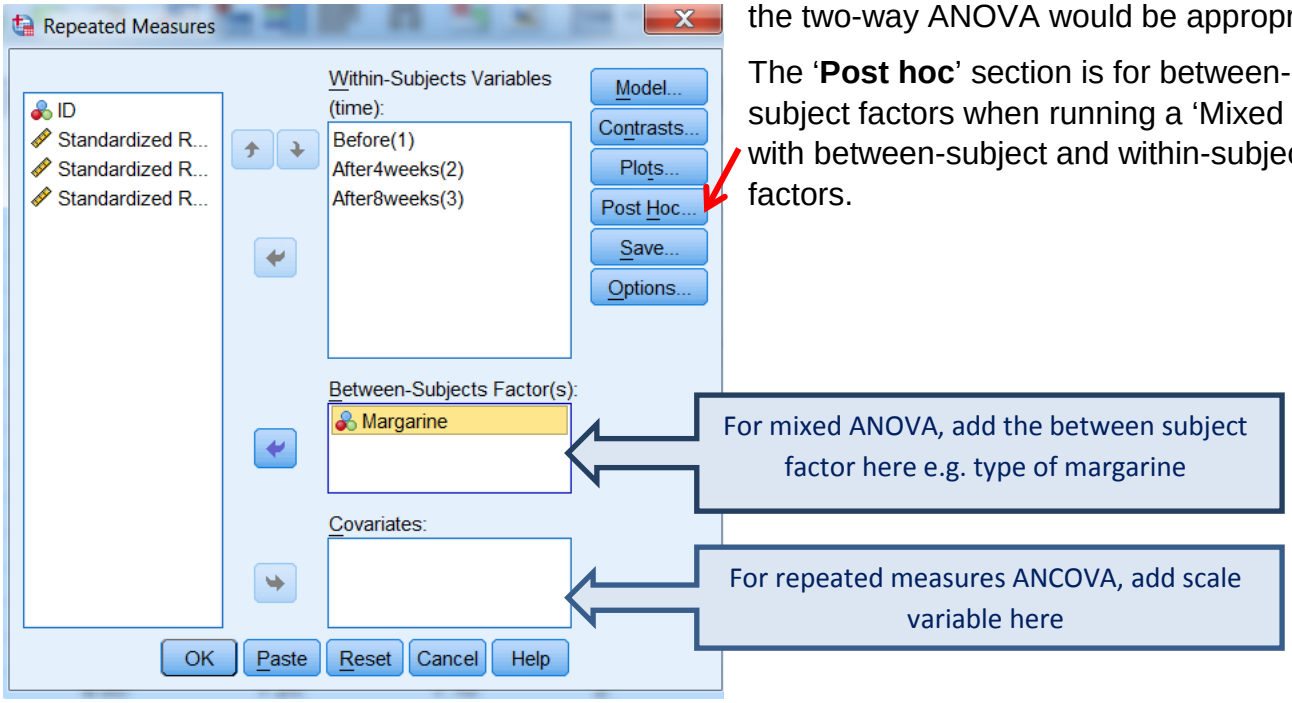
One scale dependent and several independent variables

This table shows a summary of which tests to use for a scale dependent variable and two independent variables.

1st independent	2nd independent	Test
Scale	Scale/ binary	Multiple regression
Nominal (Independent groups)	Nominal (Independent groups)	2 way ANOVA
Nominal (repeated measures)	Nominal (repeated measures)	2 way repeated measures ANOVA
Nominal (Independent groups)	Nominal (repeated measures)	Mixed ANOVA
Nominal	Scale	ANCOVA

Two way repeated measures ANOVA is also possible as well as 'Mixed ANOVA' with some between-subject and within-subject measures. For example, if participants were given either Margarine A or Margarine B, Margarine type would be a 'between groups' factor so a 'Mixed ANOVA' would be used. If all participants had Margarine A for 8 weeks and Margarine B for a different 8 weeks (giving 6 columns of data, the two-way ANOVA would be appropriate.

The '**Post hoc**' section is for between-subject factors when running a 'Mixed Model' with between-subject and within-subject factors.



Reviewer: Jean Russell, University of Sheffield

Friedman test

If the assumption of normality has not been met or the data is ordinal, the Friedman test can be used instead of repeated measures ANOVA.

The Friedman test ranks each person's score and bases the test on the sum of ranks for each column. This example uses data from a taste test where each participant tries three cola's and gives each a score out of 10. For each participant, their three scores are ranked. e.g. Participant 1 rated Cola 1 the highest, then Cola 2 and Cola 3 last, so their ranks are 1, 2 and 3 for Cola 1, Cola 2 and Cola 3 respectively.

Participant	Taste score (out of 10)		
	Cola 1	Cola 2	Cola 3
1	8	6	2
2	7	8	6
3	8	6	7
4	6	4	7
5	5	4	1

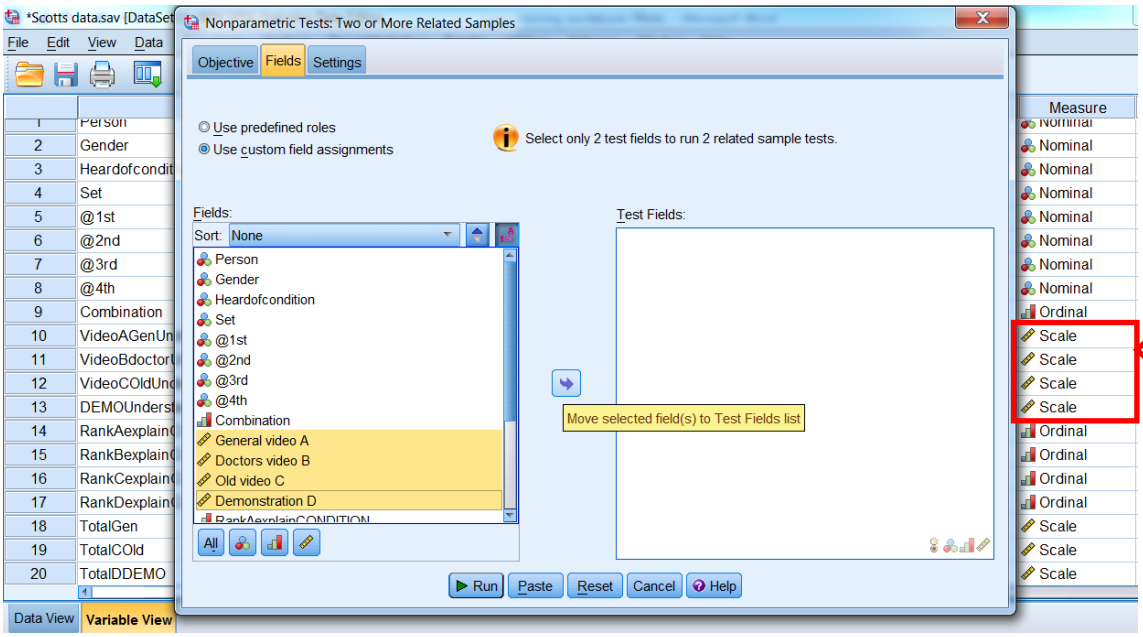
	Taste score rank		
	Cola 1	Cola 2	Cola 3
1	1	2	3
2	2	1	3
3	1	3	2
4	2	3	1
5	3	2	1
Sum of ranks	9	11	10

As the raw data is ranked to carry out the test, the Friedman test can also be used for data which is already ranked. So if the participants had not scored the cola's but just ranked them 1 – 3, the Friedman test can also be used.

Research question 4: Rating different methods of explaining a medical condition

Video data: The video dataset contained in the Excel file contains some of the results from a study comparing videos made to aid understanding of a particular medical condition. Participants watched three videos and one product demonstration and were asked several Likert style questions about each. The order in which they watched the videos was randomised. Here we will compare the scores for understanding the condition.

Analyse → Nonparametric tests → Related samples



Make sure the ordinal variables are classified as scale or the test won't run

Exercise 14: Friedman example

Carry out the Friedman test and interpret the output including the post hoc tests

Additional notes about ordinal data

Some students may want to carry out parametric statistics on ordinal data, since that may be what is expected from their department or supervisor. You can tell them that it is not considered appropriate by statisticians but accept that it is considered reasonable in other disciplines. If there are 7 or more categories and the data is approximately normally distributed, using a parametric test is reasonable although 5 categories are considered reasonable within certain disciplines. Check that the range of numbers has been used as it's common for people not to opt for the extremes. If less than 5 categories have been used, strongly advise against using a parametric test.

Where there are several questions on a questionnaire measuring one underlying latent variable, the ordinal scores can be summed/averaged and the sum/average treated as a continuous measure. For the video questionnaire, there were 5 Likert questions for each product. The 'Total' variables contain the sum.

Research question 5: Factors affecting birth weight of babies

Birth weight data: Information about 42 babies is contained in the 'reduced Birth weight' data set.

Birthweight	Gestation	smoker	motherage	mnocig	mheight	mppwt
5.8	33	0	24	0	58	99
4.2	33	1	20	7	63	109
6.4	34	0	26	0	65	140
4.5	35	1	41	7	65	125

Open the dataset ' reduced birthweight' from Excel and give the variables the labels in the following table.

Variable	Label
id	Baby ID
headcir	Head Circumference (cm)
leng	Length of baby (inches)
weight	Baby's birthweight
gest	Gestational age
mage	Maternal age
mnocig	No. cigarettes smoked per day by mother
mheight	Maternal height
mppwt	Mothers pre-pregnancy weight
fage	Fathers age
fedyrs	Years father was in education
fnocig	No. cigarettes smoked per day by father
fheight	Fathers height
lowbwt	Low birth weight baby 1 = under 5lbs
smoker	1 = smoker

Exercise 15: Assumptions for regression

Correlation and regression will be carried out using this data but what are the assumptions for multiple regression?

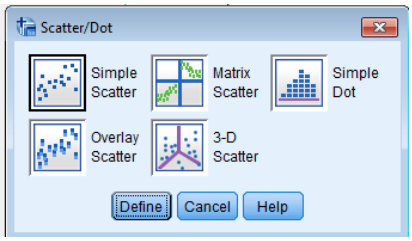
Reviewer: Jean Russell, University of Sheffield

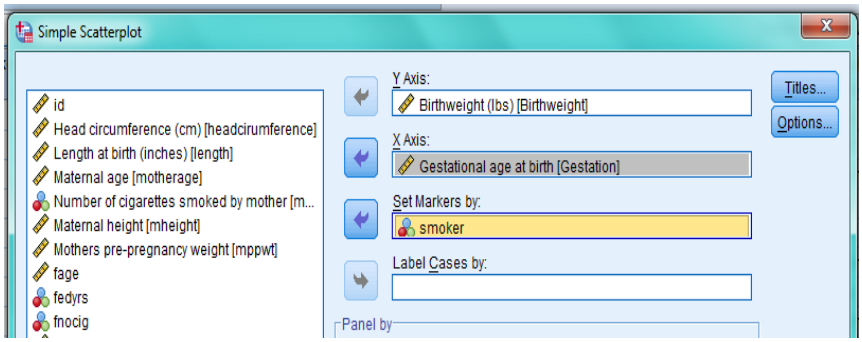
Scatterplots

A scatterplot helps assess a relationship between two continuous (scale) variables by plotting a different point for each individual based on their scores on two variables.

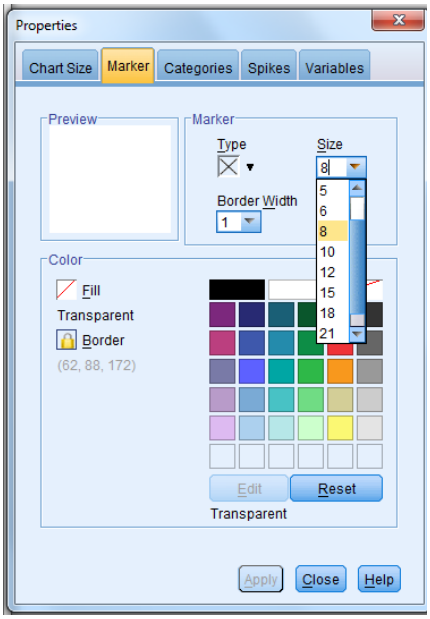
A scatterplot can be colour coded by a third categorical variable using the 'Set marker by' option within the *Graphs → Legacy Dialogs → scatterplot* menu.

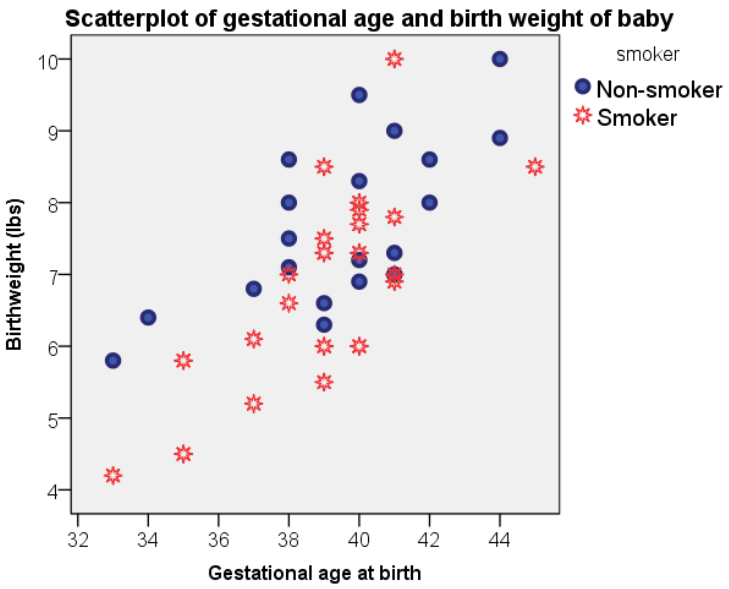
Here, we will look at the relationship between gestational age and birth weight with different shapes for mothers who smoke/ do not smoke.





Double click on the chart to open the edit window. To change the shape of the scatter, click on the scatter, then again on just one of the smokers to open the properties window. Change the marker type and size.





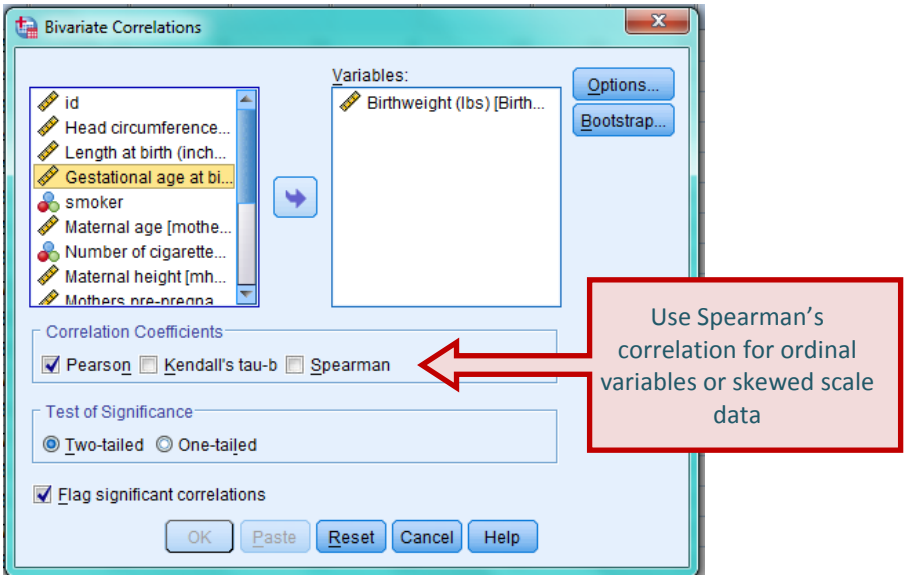
Exercise 16: Scatterplots

Describe the relationship between gestational age, smoking and birth weight. Does it look like there is an interaction between smoking and gestational age?

Correlation

To calculate correlation coefficients in SPSS:

Analyse → Correlate → Bivariate



Interpreting coefficients

Although it is possible to test correlation coefficients, this does not confirm that there is a strong relationship (only that the correlation coefficient is not 0). The test is highly influenced by sample size. For a sample size of 150, a correlation coefficient of 0.16 is significant! The best way to interpret the correlation is by using the classification proposed by Cohen.

An interpretation of the size of the coefficient has been described by Cohen (1992) as:
r = –0.3 to +0.3 (weak relationship)
r = 0.3 to 0.5 or -0.5 to -0.3 (moderate relationship)
r = 0.5 to 0.9 or -0.9 to -0.5 (strong relationship)
r = 0.9 to 1.0 or -1 to -0.9 (very strong relationship)
Source: Cohen, L. (1992). Power Primer. Psychological Bulletin, 112(1) 155-159.

Exercise 17: Correlation

Produce correlations between birth weight, gestational age, height and weight of mother. Interpret the coefficients.

Correlations

		Birthweight (lbs)	Gestational age at birth	Maternal height	Mothers pre-pregnancy weight
Birthweight (lbs)	Pearson Correlation	1	.706**	.368*	.390*
	Sig. (2-tailed)		.000	.017	.011
	N	42	42	42	42
Gestational age at birth	Pearson Correlation	.706**	1	.231	.251
	Sig. (2-tailed)	.000		.141	.110
	N	42	42	42	42
Maternal height	Pearson Correlation	.368*	.231	1	.671**
	Sig. (2-tailed)	.017	.141		.000
	N	42	42	42	42
Mothers pre-pregnancy weight	Pearson Correlation	.390*	.251	.671**	1
	Sig. (2-tailed)	.011	.110	.000	
	N	42	42	42	42

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Regression

Dependent variable: Continuous/ scale

Independent variables: Continuous/ scale or binary (coded as 0/1). Note: Categorical variables with 3+ categories can be used if recoded as several binary variables.

Uses: Assessing the effect of independent variables on the dependent variable and producing an equation to predict values of the dependent variable.

The maths:

For multiple regression, the model can be used to predict the value of a response or dependent variable y using the values of a number of explanatory or independent variables x_1 to :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_q x_q + \varepsilon$$

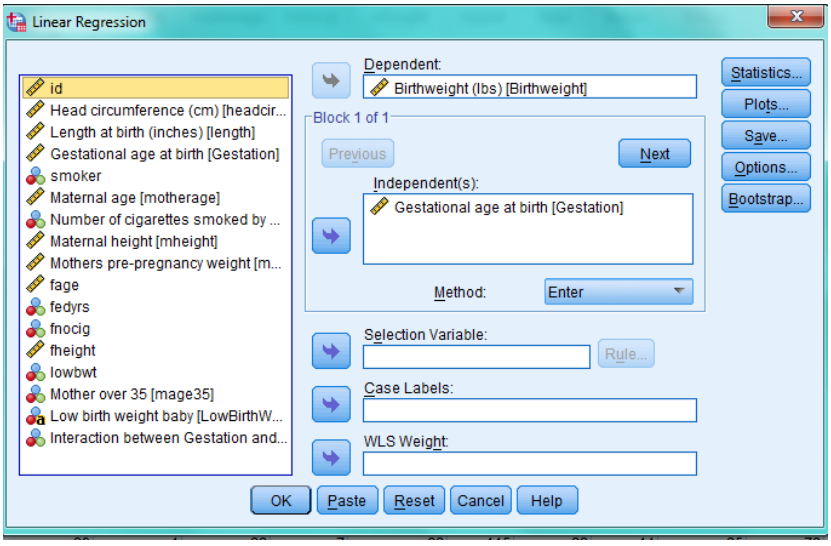
$\beta_0 =$ Constant/ intercept , $\beta_1 \rightarrow \beta_q$ are the coefficients for q independent variables $x_1 \rightarrow x_q$

The regression process finds the coefficients which minimises the squared differences between the observed and expected values of y (the residuals).

Important note: Students are not usually interested in finding the best model or using the model to predict. They are just looking for significant relationships. Model selection is not normally needed but if a student asks about it, then show them how to do it.

Quick question: What is being tested in regression?

To carry out regression *Analyze → Regression → Linear*



Independent variables:
Gestational age
Weight of mother (ppwt)
Whether the mother smokes or not

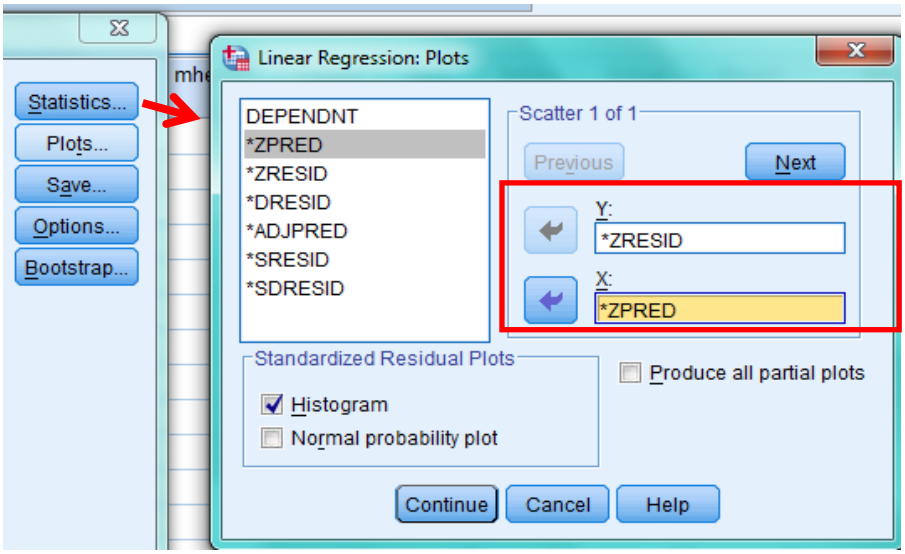
Reviewer: Jean Russell, University of Sheffield

Checking the assumptions:

Assumption	Plot to check
The relationship between the independent and dependent variables is linear.	Original scatter plot of the independent and dependent variables
Homoscedasticity: The variance of the residuals about predicted responses should be the same for all predicted responses.	Scatterplot of standardised predicted values and residuals
The residuals are normally distributed	Plot the residuals in a histogram
The residuals are independent. Are adjacent observations related? Example: Weather by day	If you suspect that the data may be auto correlated you can use the Durbin Watson statistic. Note: Time series is beyond the scope of most students

If the residuals are not normally distributed, the data needs to be transformed. The most common transformation is to take the log of the dependent variable and re-run the analysis again as there is no non-parametric test for regression. The interpretation of the coefficients is different so check how to interpret the output correctly.

The options for assumption checking are in the 'plots' window.



Check the assumption of homogeneity by asking for the graph of the standardised predicted values and residuals

Exercise 18: Regression

Interpret the output from the regression including answering the following questions:

a) ***Which independent variables are significant and what is their relationship with the dependent variable?***

b) ***What is the equation of the model?***

c) ***How good is the model at predicting birth weight?***

Dummy variables and interactions

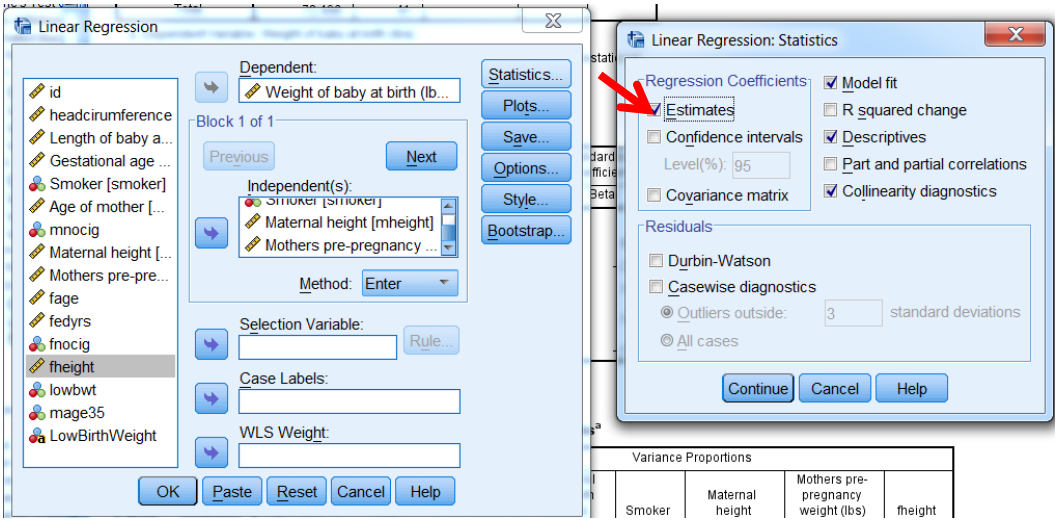
Dummy variables are binary variables created from a categorical variable. For example, if smoking status was classified as Non-smoker, light smoker or heavy smoker, 2 dummy variables are needed. The first would be 'Is the mother a non-smoker?' yes (1) or no (0) and the second would be 'Is the mother a light smoker' yes (1)or no (0). If the answer to both is No, they must be a heavy smoker.

An interaction between two independent variables means that the effect of one is different depending on the value of a second. For example, the effect of gestation may differ between smokers and non-smokers.

To test if this interaction is significant, an interaction term must be calculated using *Transform → Compute variable* to multiply the two variables together. Then add this new variable to the regression model and re-run.

Multicollinearity

For multiple regression, another issue is multicollinearity. This occurs when independent variables are too correlated with each other and causes problems with the calculations. Pairwise correlations can help asses if this is a problem. Correlations between independent variables of above 0.8 indicate a problem. There are also specific checks within SPSS to look at this problem. Request Collinearity diagnostics through the Statistics menu. Asking for the descriptives, gives a correlation matrix without the correlation p-values.



Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	-10.662	4.060		-2.626	.012		
Gestational age at birth (weeks)	.308	.053	.613	5.809	.000	.920	1.087
Smoker	-.679	.268	-.258	-2.534	.016	.987	1.013
Maternal height	.072	.072	.138	1.008	.320	.544	1.839
Mothers pre-pregnancy weight (lbs)	.012	.012	.146	1.062	.295	.540	1.853

a. Dependent Variable: Weight of baby at birth (lbs)

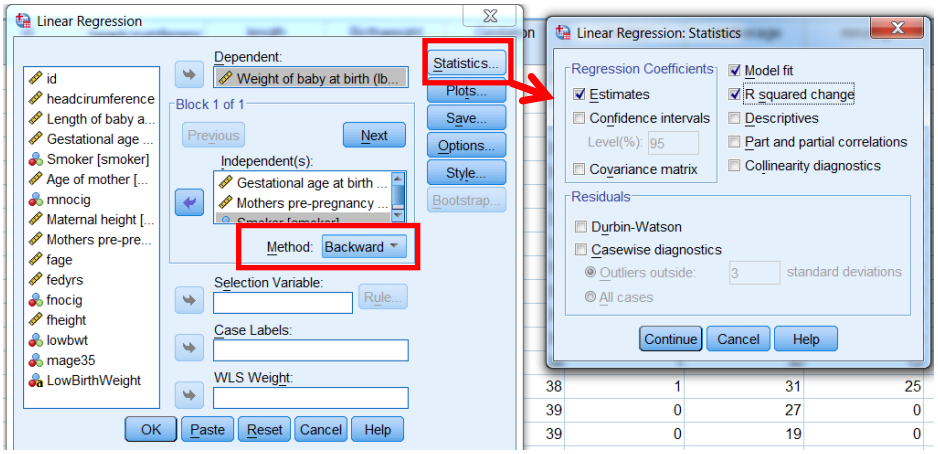
The VIF should be close to 1. Above 5 is a potential issue and above 10 indicates severe multicollinearity so the variable should be removed

Model selection

- If models are to be used for prediction, only significant predictors should be included unless they are being used as controls
- Methods include forward, backward and stepwise regression
- Backward means that the predictor with the highest p-value is removed and the model re-run. Keep going until only significant predictors are left

Don't let the student enter all variables into the model. They must think carefully about what to include and check for multicollinearity.

Add the height of the mother to the model and select Method = Backward underneath the independent variables box. Also ask for 'R squared change' from the statistics options.



The output will contain information for each step until all the variables are significant.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.788[a]	.621	.580	.8620	.621	15.143	4	37	.000
2	.781[b]	.610	.580	.8622	-.010	1.016	1	37	.320

a. Predictors: (Constant), Maternal height, Smoker, Gestational age at birth (weeks), Mothers pre-pregnancy weight (lbs)

b. Predictors: (Constant), Smoker, Gestational age at birth (weeks), Mothers pre-pregnancy weight (lbs)

The R squared change test tests whether removing the least significant variable has made a significant change to the R squared value. Removing height has not made a significant difference.

Logistic regression

Logistic regression is the same as standard regression but the outcome variable is binary and leads to a model which can be used to predict the probability of an event happening for an individual.

The maths:

Since the outcome of a logistic regression model is binary, the model is based on predicting the probability that an event will occur for an individual and is expressed in linear form as follows:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_q x_q,$$

where p = probabilty of event occuring e.g. person dies following heart attack.

In the above p /(1- p) represents the "odds" of the event occurring and so ln[p /(1- p)] is the log-odds of the event. The term ln[p /(1- p)] is often referred to as the logit hence the name logistic regression.

The model can also be expressed in terms of p in the following way which is equivalent:

$$p = \frac{\exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_q x_q\right)}{1 + \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_q x_q\right)}, \quad 0 < p < 1.$$

The key variables of interest are:

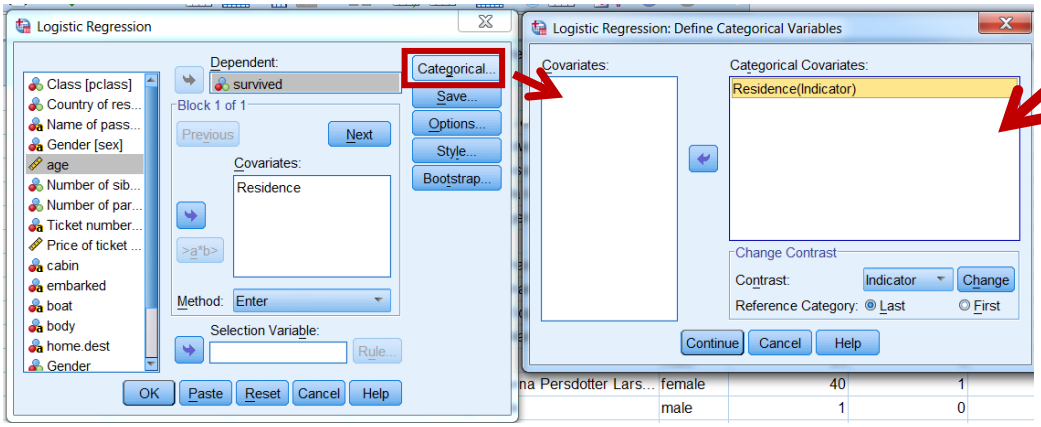
Dependent variable: Whether a passenger survived or not (survival is indicated by survived = 1).

Possible explanatory variables: Country of residence, age, gender (recode so that sex = 1 for females and 0 for males), class (pclass = 1, 2 or 3)

Firstly, a model with just country of residence as an independent and survival as the dependent will be run.

In SPSS, use *ANALYZE → Regression →Binary logistic*

When interpreting SPSS output for logistic regression, it is easier if binary variables are coded as 0 and 1. Also, categorical variables with three or more categories need to be recoded as dummy variables with 0/ 1 outcomes e.g. class needs to appear as two variables 1st/ not 1st with 1 = yes and 2nd/ not 2nd with 1 = yes. Luckily SPSS does this for you! When adding a categorical variable to the list of covariates, click on the **Categorical** button and move all categorical variables to the right hand box.



Where there are more than two categories, the last category is automatically the reference category. This means that all the other categories will be compared to the reference in the output e.g. 1st and 2nd class will be compared to 3rd class.

The following table in the output shows the coding of the categorical variables.

Categorical Variables Codings

		Frequency	Parameter coding (1)	Parameter coding (2)
Country of residence	America	258	1.000	.000
	Britain	302	.000	1.000
	Other	749	.000	.000

For country of residence, 'Other' is the reference category, America will be residence (1) and Britain will be residence (2).

Interpretation of the output

The output is split into two sections, block 0 and block 1. Block 0 assesses the usefulness of having a null model, which is a model with no explanatory variables. The 'variables in the equation' table only includes a constant so each person has the same chance of survival.

The null model is: $\ln\left(\frac{p}{1-p}\right) = \beta_0 = -0.481, \quad p = \text{probability of survival} = \frac{\exp(-0.481)}{1+\exp(-0.481)} = 0.382$

SPSS calculates the probability of survival for each individual using the block model. If the probability of survival is 0.5 or more it will predict survival (as survival = 1) and death if the probability is less than 0.5. As more people died than survived, the probability of survival is 0.382 and therefore everyone is predicted as dying (coded as 0). As 61.8% of people were correctly classified, classification from the null model is 61.8% accurate. The addition of explanatory variables should increase the percentage of correct classification significantly if the model is good.

Block 0: Beginning Block

Classification Table[a,b]

Observed			Predicted: survived: Died	Predicted: survived: Survived	Percentage Correct
Step 0	survived	Died	809	0	100.0
		Survived	500	0	.0
	Overall Percentage				61.8

a. Constant is included in the model.
b. The cut value is .500

Reviewer: Jean Russell, University of Sheffield

Block 1: Method = Enter

Block 1 shows the results after the addition of the explanatory variables selected.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	43.765	2	.000
	Block	43.765	2	.000
	Model	43.765	2	.000

The Omnibus Tests of Model Coefficients table gives the result of the Likelihood Ratio (LR) test which indicates whether the inclusion of this block of variables contributes significantly to model fit. A p-value (sig) of less than 0.05 for block means that the block 1 model is a significant improvement over the block 0 model. Adding Country of residence has therefore made a significant improvement to the model.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1697.260[a]	.033	.045

a. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

In standard regression, the coefficient of determination (R^2) value gives an indication of how much variation in *y* is explained by the model. This cannot be calculated for logistic regression but the 'Model Summary' table gives the values for two pseudo R^2 values which attempt to measure something similar. From the table above, using the Nagelkerke R^2 we can sort of conclude that about 4.5% of the "variation in survival can be explained by the model in block 1".

The classification table shows that the correct classification rate has increased from 61.8% to 64.2%.

Finally, the 'Variables in the Equation' table summarises the importance of the explanatory variables individually whilst controlling for the other explanatory variables.

Odds ratios

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Residence			43.250	2	.000	
	Residence(1)	.887	.147	36.338	1	.000	2.428
	Residence(2)	-.126	.146	.749	1	.387	.882
	Constant	-.638	.077	68.878	1	.000	.529

a. Variable(s) entered on step 1: Residence.

The Wald test is used to test the hypothesis that each $\beta_i = 0$. In the 'Sig' column, the p-value for Residence (1), which is America, is significant but the p-value for Residence (2) is not. When interpreting the differences, look at the $\exp(\beta)$ column which represents the odds ratio for the individual variable. The odds of an American surviving were 2.428 times higher than for those in the "other" (i.e. not America or Britain) group.

Note: If the student needs Confidence Intervals for the odds ratios, request them through the 'Options'.

Exercise 19: Logistic regression

Look at the relationship between nationality and survival but control for gender and class.

Which variables are significant? Interpret the odds ratio for those variables.