

# An Introduction to Data Analysis using SPSS

## Introduction and Aims

Given some data from an experiment or survey of some kind, an important first step is to explore some of the basic features of the data using simple statistics and plots. Suppose, for example, that we conduct a survey of people and ask them how often they smoke cigarettes, as well as some demographic information such as age and sex. We might like to get an idea of the distribution of ages of people who responded to the survey, or the overall proportions of people who smoke regularly, occasionally and not at all, before we proceed to a more complex analysis to determine factors which are associated with increased/decreased levels of smoking.

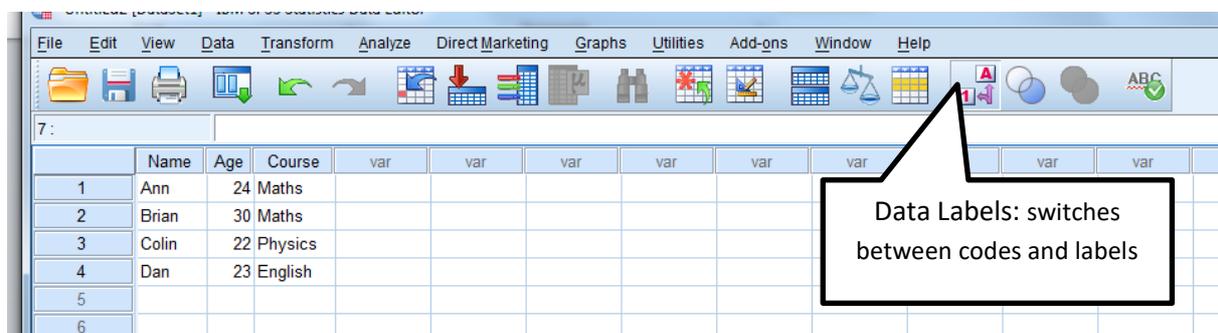
This introduction concentrates on using SPSS for the exploratory phase of data analysis, then briefly discusses some commonly used statistical techniques, as follows:

	<i>Page</i>
1. How data is input and stored in SPSS (including import from On-Line Survey and Excel)	1
2. Summary statistics and plots (for categorical data and for scale data)	4
3. Editing variables (recoding a variable and calculating a new variable)	7
4. Managing, Saving and Exporting SPSS Output	10
5. Testing for Normality (checking data prior to doing statistical tests)	11
6. Relationships between two variables (Cross-tabulation and Chi-Squared test, boxplots, scatter diagrams, correlation coefficient)	14
7. Data sorting, grouping, transformation and selection	17
8. Comparing means (comparison and t-tests)	17
9. Resources	18
APPENDIX 1.: Summary of Useful Commands in SPSS	19
APPENDIX 2.: What Statistical Test do I need?	25

## 1 Data input

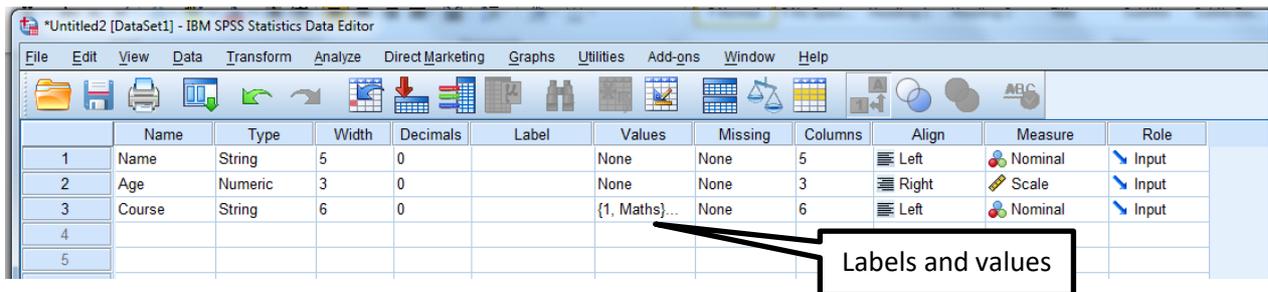
SPSS presents the data in two views: data and variable

**Data view** Looks like Excel. Each row is a “case”, e.g. person. Each column is an “attribute” or variable, e.g. height, age, gender, place of birth.



**Variable view** describes each attribute:

- variable name (which must not contain spaces),
- type (string/numeric),
- label (a meaningful name for the variable for use in the output),
- values e.g. 1="Maths", 2="English", 3="Physics", (use Data Label button to see the codes)
- missing values (code(s) to denote missing data, e.g. 999 to represent missing data on age)
- measure (nominal, ordinal, scale)



**How to enter Data:** In Data view, type in the data (just as you would in Excel)  
 Copy and paste data e.g. from Excel, or from a table in Word  
 Import data from On-Line Survey (section 1.1)  
 Import an Excel file using File > Open > Data (section 1.2)

### 1.1 Importing data from On-line survey (formerly BOS)

#### Export response data

Any filtering that you have applied will be reflected in the data that is exported.  
 If you want to export all your data, please clear any existing filters first.  
 For information on the options for exporting data please see: [Exporting response data](#).

##### Customise your export:

- Include unique response number for each respondent
- Include date of response submission
- Include date and time of response submission
- Include date and time response was started
- Include section headings
- Exclude free text responses
- Use alternative question text (if provided)

##### Options for coded exports:

- Code responses (for import into statistical software)
  - Zero index all multiple choice, multiple answer, selection list and scale questions
  - Reverse index values for all multiple choice, multiple answer, selection list and scale questions
  - Combine scale/rank values into a single column where possible

##### Select the format required:

- File format: **Microsoft Excel 2007 and later (.xlsx)**  
 Microsoft Excel 2003 and older (.xls)  
 Comma Separated Values (.csv)  
 SPSS (.sav)  
 Tab Separated Values (.tsv)

Choose the type of file you want to be created.

Please note, some spreadsheet and statistical software impose limits on the number of columns of data they can display (before truncating the data). Please refer to the [Online Surveys knowledgebase](#) for the limits in some common software tools and how to reduce the number of columns in your exported data.

**a) Downloading into SPSS**

Under “Options for coded exports” tick the box for “Combine scale/rank values into a single column where possible”. The purpose of this is to provide a single variable coded 1 to *n*, where *n* is the number of options on the Likert Scale, e.g.

4. Select the answer which best describes your attitude.

Please don't select more than 1 answer(s) per row.

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
Parking at work is adequate	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Commuting by car is the only feasible method of transport	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

The answers will be coded as 1=Strongly Agree, 2=Agree, ....5=Strongly disagree. This option is useful to deal with the scale/rank questions. It does not affect the other types of question.

**b) Downloading into Excel**

If you simply download, then the data will look like this:

	1	2	3	4	5	6	7	8	9	10	11	12	13
1.	How do you tr	2. I participate in the following sports/i	3. How do you c	4. Select the	4.1. Parking i	4.2. Commut	4.3. Car use i	4.4. Public trans	5. Enter the	5.1.a. Car - D	5.1.b. Car - F	5.2.a.	5.2.b.
2	Car	horse sports,dance,outdoor leisure (e.g.)	Car		Agree	Disagree	Agree	Strongly disagree	02/04/2019	Usually	Neve		
3	Other	team sport (e.g. football, hockey),horse	Car		Strongly agr	Strongly agr	Strongly agr	Strongly disagree	02/04/2019	Usually	Neve		
4	Passenger in car	individual sport (e.g. tennis, cycling),sv	Bus		Agree	Neither agre	Neither agre	Neither agree nor disagree	02/04/2019	Often	Neve		

However, if, under “Options for coded exports”, you tick the box for “Code responses” the data will be coded with 1’s and 0’s, or with codes e.g. 1=Car, 2=Passenger in car, 3=Motorcycle....8=Other, so it becomes easier to analyse.

	Q1	Q2_1	Q2_2	Q2_3	Q2_4	Q2_5	Q2_6	Q2_7	Q2_8	Q2_9	Q2_10	Question2	Q3
2	1	0	0	0	0	1	0	0	1	1	0		0
3	8	0	1	0	0	1	0	0	1	1	1		5
4	2	0	0	1	1	0	0	0	0	0	0		2

Question 1, multiple choice single answer and question 3, selection list question, have both been coded as required with the options given a code from 1 to 8 or from 1 to 7

Column L inserted to calculate the total for columns C to K, i.e. sum of Q2\_2 to Q2\_10. In row 2 I used SUM and it doesn't work, in rows 3 and 4 I used a formula, which does work.

### **Possible problems and issues**

- Be aware that choosing the “multiple choice (multiple answers) question” can cause problems with analyses. Only use this question after careful consideration of how you will deal with the data.
- If SPSS will not calculate a new variable it may be because that variable is counting 0 as the indicator of a missing value. Change the “Missing” attribute to “none” and it should work.
- Choosing the Scale/rank question or the Grid question will lead to variables in SPSS or Excel labelled as sub-questions, e.g. Q5\_1, Q5\_2 etc.
- Consider how to download the data:
  - For SPSS, under “Options for coded exports” tick the box for “Combine scale/rank values into a single column where possible”. This affects the format of the Scale/Rank questions, giving a single score to represent the choice made; it does not affect other question types.
  - For Excel, under “Options for coded exports”, tick the box for “Code responses”

## **1.2 Importing data from Excel**

We can open an Excel data file in SPSS using File > Open > Data

Change the ‘Files of type:’ option to “Excel”. A dialogue box will ask about what part of the spreadsheet it should import; if the first row of your Excel file is the field names then select the option, ‘Read variable names from the first row of data’. Click ‘OK’ and the data should appear in a new Data Editor window. Now change to the Variable View and make the necessary changes and additions to the attributes of the variables. As a minimum, make sure that you check/change the attributes ‘type’, ‘label’, ‘values’ and ‘measure’ for all the variables.

An alternative method is to copy the data from Excel and paste it into SPSS, making sure that you edit the variable names, types etc.

## **2 Summary statistics and plots**

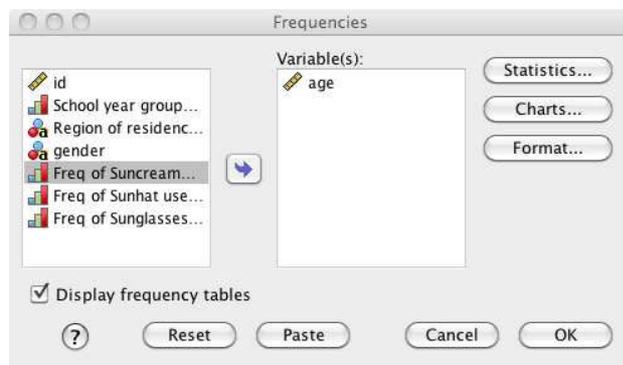
The way we summarise data, and indeed the way we choose to analyse all data, depends strongly on the nature of the data we are dealing with (in SPSS language, the “measure” of the variables involved): scale, ordinal or nominal.

### **2.1 Categorical variables**

We will use some examples based on data from a survey of secondary school children’s attitudes to and awareness of the dangers of the sun and skin cancer. First, we consider the summary of data associated with categorical (either nominal or ordinal) and sometimes discrete numerical variables.

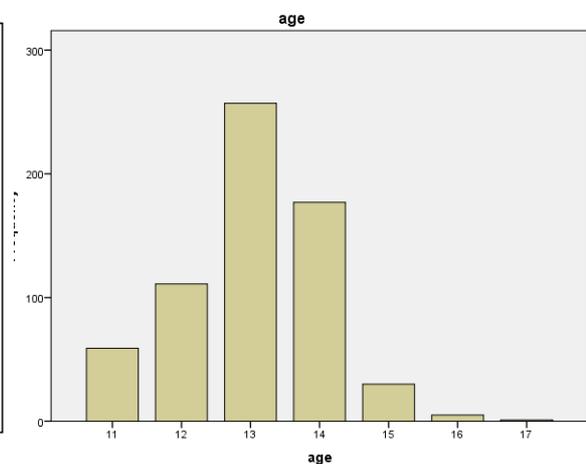
Use **Analyse > Descriptive Statistics > Frequencies**

Move the required variable (e.g. 'age') from the list of variables in the left-hand panel to the central panel labelled 'Variable(s):' Click on the Charts button, ensure that Bar Charts and Frequencies are selected, then click Continue and then OK.



The output window now shows the frequency table and bar chart that we requested.

age				
	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	11	59	9.1	9.2
	12	111	17.2	26.6
	13	257	39.8	66.7
	14	177	27.4	94.4
	15	30	4.6	99.1
	16	5	.8	99.8
	17	1	.2	100.0
	Total	640	99.1	100.0
Missing	System	6	.9	
Total		646	100.0	



From the bar chart we can immediately make observations like, 'most respondents are aged 13, with quite a lot aged 14 and 12 also'. Using the frequency table we can quantify these observations using the numbers or percentages SPSS has calculated.

Note that if a discrete variable takes lots of different values then then the frequency table will become large and unwieldy and the bar chart will have lots of bars, rendering both of them much less useful. It can be useful to recode the variable of interest e.g. into classes like 'high', 'medium' and 'low' (see Section 3.1). An alternative, if appropriate, is to treat the variable as a scale variable and use the techniques described in the next section.

## 2.2 Scale variables

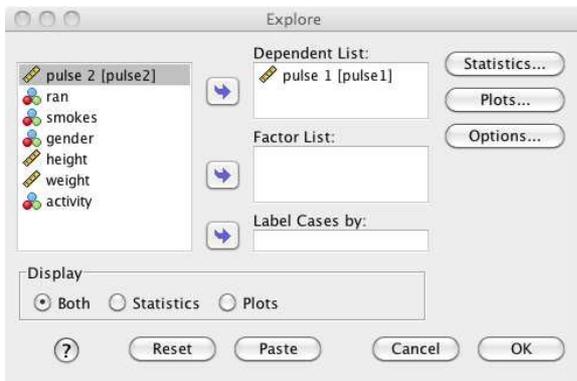
We now consider how to summarise numerical or scale variables, such as a person's height or weight or the number of questions they got correct in a test.

Simple summary statistics (minimum, maximum, mean, standard deviation) can be obtained from **Analyse > Descriptive Statistics > Descriptives**

We will probably want more than this. Our example is derived from an experiment where volunteers had their pulse rate measured, were randomly assigned to either run on the spot or sit still for 3 minutes and had their pulse rate measured again. Some other lifestyle and personal information was also collected. We can summarise the scale variable “pulse1” using

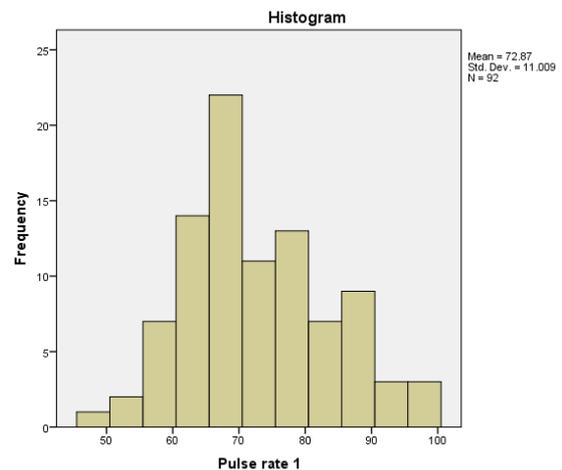
**Analyse > Descriptive Statistics > Explore**

Choose this menu item and move the variable pulse1 into the ‘Dependent List:’ panel. Click on the ‘Plots’ button and deselect ‘Stem-and-leaf’ and select ‘Histogram’, then click Continue and then OK.



The table headed Descriptives in the Output window shows a huge number of summary statistics, some of which might be incorporated into a report as part of a brief summary of this data. Next are a histogram and box-plot of the data.

Descriptives		Statistic	Std. Error
Pulse rate 1	Mean	72.87	1.148
	95% Confidence Interval for Mean	Lower Bound	70.59
		Upper Bound	75.15
	5% Trimmed Mean	72.63	
	Median	71.00	
	Variance	121.192	
	Std. Deviation	11.009	
	Minimum	48	
	Maximum	100	
	Range	52	
	Interquartile Range	16	
	Skewness	.397	.251
	Kurtosis	-.442	.498



The bin locations and widths for the histogram are chosen automatically, but can be changed through the Chart Editor which is obtained by double-clicking on the chart you wish to edit.

Note that a box-plot shows the minimum, maximum, median and the first and third quartiles of the data. An example is shown in Section 6.2.

It is possible to treat a discrete numerical variable (such as age) as though it were continuous and use the Explore menu item rather than Frequencies. This should be done with care, but can be appropriate when the number of values is very large and the output from frequencies would be unwieldy. In this situation the statistics given by Explore become more meaningful than the frequencies of all the individual values of the variable.

### 2.3 Discrete numerical variables

Numerical variables taking discrete values can be summarised using either “Explore” or “Frequencies”. If the range of values is quite small then both options give sensible output, e.g. a numerical representation of a Likert scale of 1, 2, 3, 4, 5. However, if the range of possible values is very large then the output from Frequencies will be much less meaningful (the tables will be enormously long), so the Explore option is preferred.

Think about what the output would be in each case for a variable recording the ages of 100 people randomly stopped on a high street to participate in a survey. With such a variable another option is to create your own groups using

Transform > Recode into Different Variables, (explained in 3.1 below), then analyse the new variable using Frequencies.

## 3 Editing Variables

### 3.1 Recoding a Variable

Sometimes we want to group items together into new groups. For example, suppose we have data on children in the different years but are particularly interested in comparing students in school years 7 and 8 with the other children in the survey. (Perhaps they have been the target of a recent campaign and we want to investigate whether it has been effective.)

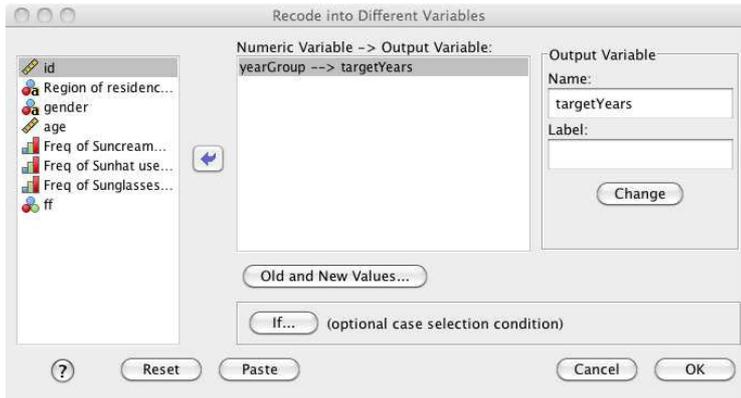
SPSS can create a new variable for us that indicates whether the respondents are younger than the target group, in the target group, or older than our target school years. We want to recode the ‘yearGroup’ variable to a new variable, say ‘targetYears’, as in the table:

yearGroup	6 and below	7 and 8	9 and above
targetYears	0	1	2

We could do this manually, but SPSS can do the hard work for us through the menu command

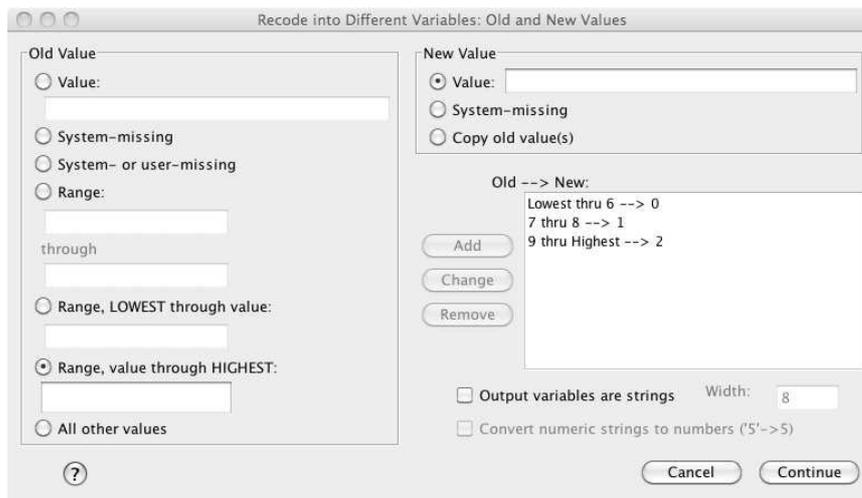
Transform > Recode into Different Variables.

Choose this menu item, select the yearGroup variable from the list on the left hand side and give the Output Variable the name ‘targetYears’ and a sensible label and click ‘Change’.



Click on 'Old and New Values. . . '.

In the Old Value section of the dialogue box, click on 'Range, LOWEST through value:' and enter '6' in the input box; then enter '0' (zero) under New Value and click the 'Add' button. Repeat this for the other values (1 and 2) of the new variable, using the 'Range:' and 'Range: value through HIGHEST:' options in the left-hand panel of the dialogue box. (If you make a mistake use the 'Change' and/or 'Remove' buttons.) The dialogue box should look like this:



Check the entries in the box labelled 'Old --> New', then click 'Continue' and 'OK'; and the if new variable with its appropriate values, will be added to the data, ready for further analysis. In Variable View of the Data Editor give the new variable an appropriate Label and Values and check that the Measure attribute is appropriate.

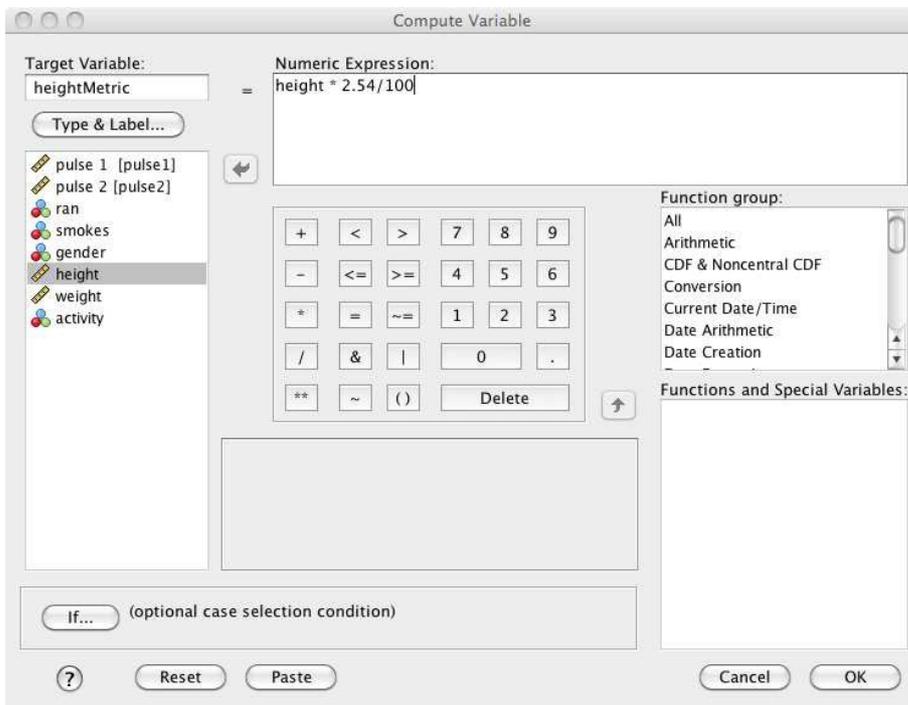
There is also a menu item 'Recode into Same Variables' which overwrites old values with recoded values. Use this option carefully. It is usually wise to recode into a new variable and keep your original data. If you are sure that won't want to directly analyse the original variable and have your data backed up elsewhere then, after checking that the recode has worked as you intend, you might delete the old variable.

### 3.2 Calculating a new variable

Suppose that we want to convert a height measured in inches to metres for further analysis and so our reporting will conform to standard practice. We could do the calculation ourselves, but we can also get SPSS to do it for us using the

Transform > Compute Variable

menu item. You need to name the new variable you want to create (e.g. heightMetric) and then enter the formula (there are 2.54 cm in an inch) to compute its values and click OK. (Also note the extensive list of available functions in the list in the lower-right section of the Compute Variable dialogue box.)



If you switch to Data View then you should see the new variable with all its values calculated. Note the difference between the Compute and Recode functions. They do similar things, but you should use Compute when there is a simple mathematical formula to describe the relationship between the existing variables and the one you want to calculate; otherwise use Recode. A little thought about which one will be simpler can save you a lot of time!

## 4 Managing, Saving and Exporting SPSS output

### 4.1 Exporting output

Having produced several tables and plots that are now in the SPSS output window, you will perhaps want to incorporate some of this output into a more formal document of some kind (a short report, paper or thesis, for example). To export tables to Word or Excel, simply select the relevant object in the SPSS output window and copy it, then paste it into the relevant application. The same copy-paste method works for exporting charts/plots to Word.

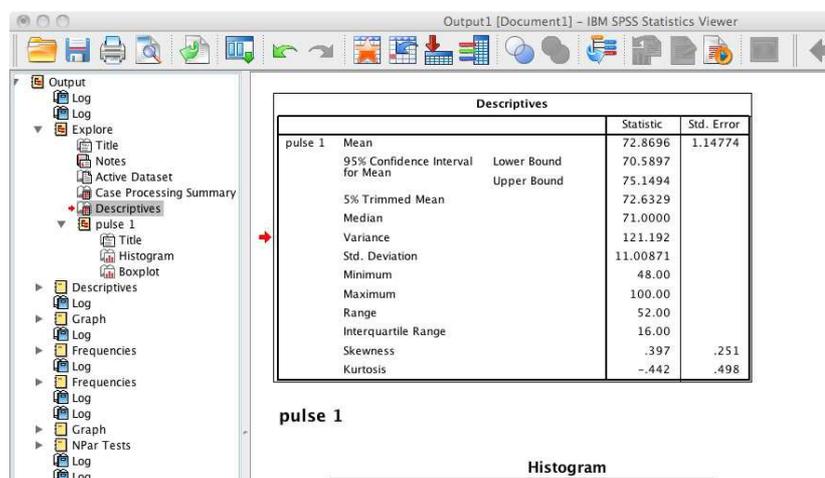
To export charts/plots for use in some other program (like LATEX, or just so that you have a copy of a plot that you can then import into various other documents without having to copy-paste from SPSS), right-click on the plot to export and select 'Export' from the menu that pops up. Make sure that the Document Type is set to 'None (Graphics only)' and the Graphics Type is set to an appropriate format. Before clicking OK, check under the heading 'Root File Name:' to see (and possibly change) where your file will be exported to.

To export a table, right-click on it and choose Export in the same way; this time choose 'Text - Plain' under Document Type and check the location your output will be saved to, which is shown under the heading 'File Name:'

### 4.2 Managing output

Once SPSS has produced a number of tables and charts in the output window, it can be a bit difficult to navigate around the output and find a particular chart or table. The separate pane on the left hand side of the output window is very useful for navigating around the output: the items in this pane are a sort of 'table of contents'.

Clicking on the grey triangles near the start of the items in the left hand pane collapses their contents so you can more easily see the remaining items in the list. Clicking on an item in the list moves the output in the main part of the window so that you can see the element whose title you clicked on



Descriptives				Statistic	Std. Error
pulse 1	Mean			72.8696	1.14774
	95% Confidence Interval for Mean	Lower Bound		70.5897	
		Upper Bound		75.1494	
	5% Trimmed Mean			72.6329	
	Median			71.0000	
	Variance			121.192	
	Std. Deviation			11.00871	
	Minimum			48.00	
	Maximum			100.00	
	Range			52.00	
	Interquartile Range			16.00	
	Skewness			.397	.251
	Kurtosis			-.442	.498

pulse 1

Histogram

(notice the red arrows that appear in both the 'contents pane' and the main output window showing which entries in the left pane correspond to which parts of the output file). Output that is not needed can be deleted by right-clicking on the appropriate entry in the left hand pane and choosing 'Cut' from the menu that pops up. You can also drag and drop items in this contents pane to move them around and change the order in which they appear in the output window.

It is good practice to delete any unnecessary output promptly to reduce the amount of clutter in the output window and make it easier to find what you are looking for!

To further help you keep a track of what you have in the output window, you can use

Insert > New Text to write short notes in the output window about what analyses you have done and why.

### 4.3 Saving Output

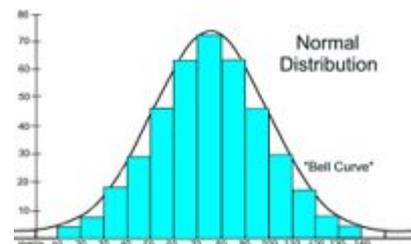
With the output window as the front window on your computer the usual

File > Save or File > Save As will allow you to save the output.

Note that this generates a .spv file, whereas data is saved as a .sav file.

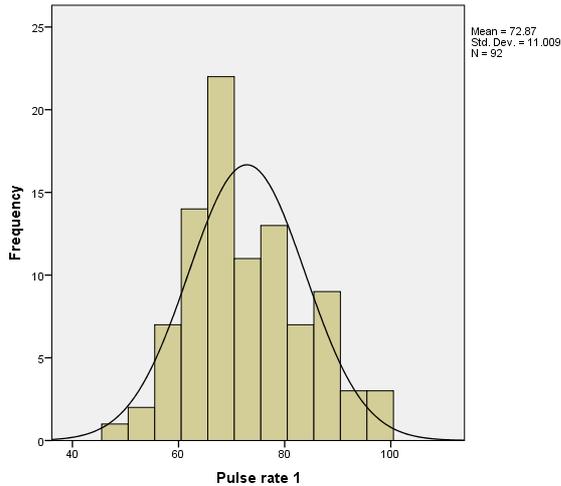
## 5 Testing for Normality

An important part of many statistical procedures which you will most likely come across in the future involves checking assumptions about data which are necessary for the procedures to be useful. One of the most common checks you will have to do will be a variation of the question 'are the data (approximately) Normally distributed?' The Normal (or Gaussian) distribution is perhaps familiar to you through its famous 'bell curve'.



Parametric test	What to check for normality
Independent t-test	Dependent variable by group
Paired t-test	Paired differences
One-way ANOVA	Residuals
Repeated measures ANOVA	Residuals at each time point
Pearson's correlation coefficient	Both variables are normally distributed
Simple linear regression	Residuals

## 5.1 Histograms



There are several ways to assess whether a set of data is well-described by the Normal distribution. One way is to draw a histogram, with the 'best-fitting' Normal distribution superimposed over it and make a subjective assessment of whether the histogram and Normal distribution seem to agree. To get a histogram with a superimposed normal curve you can use the Graphs > Legacy Dialogs > Histogram menu item.

Choose the appropriate variable(s) from the list on the left-hand side (in this case, pulse1) and make sure that the 'Display Normal Curve' option is checked before clicking OK. Remember that the appearance of a histogram can be very strongly affected by the number and size of the bins, so treat this method with caution unless your histogram is based on a very large sample (a couple of hundred or more).

## 5.2 Skew and Kurtosis

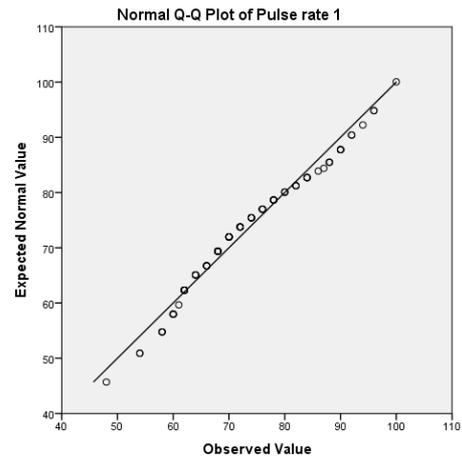
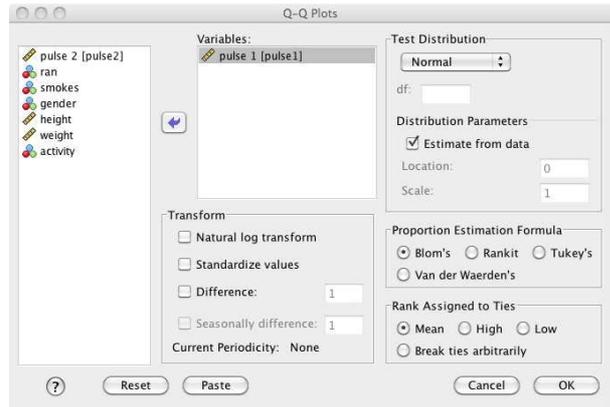
Skew measures the whether the data is symmetrical. A Normal distribution has a skew of zero. A positive value for skewness indicates more low scores; a negative value indicates more high scores. Kurtosis measures how scores are distributed between the tails, "shoulders" and centre of the distribution. A "flat" distribution, e.g. the uniform distribution, has kurtosis<3 (negative excess kurtosis) and is called platykurtic; a distribution with less in the "shoulders" and more in the centre and tails, e.g. the t-distribution, has kurtosis>3 (positive excess kurtosis) and is called leptokurtic. The excess kurtosis of a Normal distribution is zero (the kurtosis of a Normal distribution is 3, but SPSS calculates the "excess kurtosis", i.e. the difference from 3). Small samples from a Normal distribution may vary considerably in their values of skew and kurtosis, so look at the histogram, Q-Q plots or tests (Section 5.4).

## 5.3 Q-Q plots

A better graphical method (and probably the best and most-used method overall) to assess whether some data are well-described by a Normal distribution is to use a Q-Q plot. The details of how this is produced are too complicated to discuss here, but the key point is that the data fits a Normal distribution if the points on a Q-Q plot closely follow a straight line.

Analyze > Descriptive Statistics > Q-Q Plots.

Select the variable(s) to be analysed and make sure that the Test Distribution is Normal and the "Estimate From Data" option is checked under the heading Distribution Parameters.



The definite ‘waviness’ of the Normal Q-Q plot for the “pulse1” data indicates that the data is not exactly described by a Normal distribution; however the fact that it does not deviate very much from the straight line indicates that a Normal distribution does provide a fairly close approximation. In almost all applications this is good enough!

#### 5.4 Statistical tests

Another, more objective, way of assessing whether data might be well-described by a Normal distribution is to do a formal statistical test, a hypothesis test. The simplest of these is the Kolmogorov-Smirnov test, which can be accessed through

Analyze > Nonparametric Tests > Legacy Dialogs > 1-Sample K-S

Other commonly used tests are the Shapiro-Wilk and Anderson-Darling tests. The important output from such tests is the p-value (labelled ‘Sig.’ in SPSS output tables), which is a measure of how likely the observed data are, under the assumption that they are perfectly described by a Normal distribution.

If the p-value is small then, assuming the data is described by a Normal distribution, the observed data is quite unlikely; we thus conclude that the observed data is most likely not perfectly described by a Normal distribution. Conversely, if the p-value is large then, if the data is described by a Normal distribution, the observed data is quite likely, so the data we have observed is entirely consistent with it being described by a Normal distribution. As a guideline, a p-value larger than 0.1 would usually be considered as giving no evidence against the Normal distribution being appropriate, between 0.1 and 0.05 is weak/some evidence against Normality, 0.05 to 0.01 is good evidence and less than 0.01 is strong evidence that the data is not Normally distributed.

However, caution must be exercised when using such tests for Normality! If the sample size is even moderately large (50–100 and larger), then such tests are very sensitive even to quite small differences between the distribution of the data and a Normal distribution. If the p-value is small but the sample size is large then we need to interpret the results of this test with a little caution and common sense. If the data do not exactly fit a Normal distribution, they may still be very close and

for all practical purposes the Normal distribution is a good enough approximation. In this situation, there might be a statistically significant difference between the distribution of the data and the Normal distribution that best fits it, but no practically significant difference. A Q-Q plot is very useful to make this case, as in the example above.

Lastly, note that Q-Q plots and the Kolmogorov-Smirnov test can also be used to test if data is well described by other specified probability distributions. You may have heard of the Poisson, Exponential, Gamma, Pareto and Laplace distributions; SPSS allows you to compare data to these distributions (and a few others too). The fit of data to these other distributions is assessed using the same methods described above for Q-Q plots and Statistical tests, selecting the appropriate distribution instead of the Normal that we used here.

## 6 Relationships between two variables

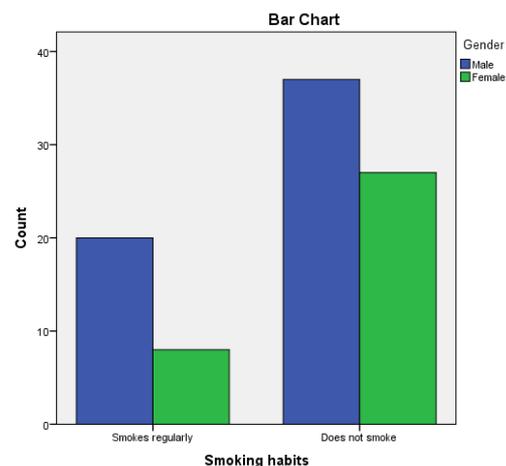
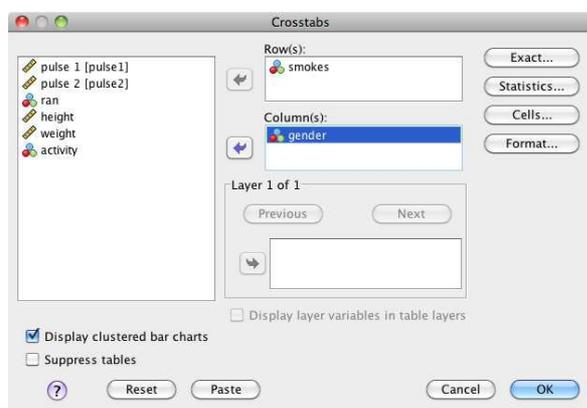
We now look at exploring the relationship between 2 (or more) variables; asking questions like ‘Does changing the value of one variable affect the value(s) of the other in some way?’

### 6.1 Two categorical variables

Firstly, we look at the case where both variables are categorical (ordinal or nominal). The data can most easily be explored using a table giving the frequencies of occurrence of particular combinations of the two variables of interest. To do this we use

Analyse > Descriptive Statistics > Crosstabs

We might wish to explore a possible relationship between gender and whether someone smokes. Enter “smokes” as the row variable and “gender” as the column. The Statistics button allows you to get SPSS to calculate various statistics and perform some statistical tests on the data e.g. the Chi-squared test to check for association. The Cells button allows you to alter the information displayed in the output tables; in particular, the Percentages section can be useful. There is also an option to display clustered bar charts.



Choose the options carefully: can you determine what proportion of females are smokers? And what proportion of smokers are female? These are not the same! Other options could be to use “smokes” as the Row variable, “activity” as the Column and “gender” as a Layer.

Also note that the Crosstabs procedure is only sensible for categorical variables, there is a separate row/column for every distinct value the Row/Column variable takes, so the output will be meaningless if we put a scale variable anywhere in the Crosstabs dialogue box.

## 6.2 One categorical and one scale variable

Now we shall see how to explore how a categorical variable might influence a continuous variable. An example might be to see if typical levels of activity seem to affect people’s normal resting heart rate, (measured by pulse1). To do this we use

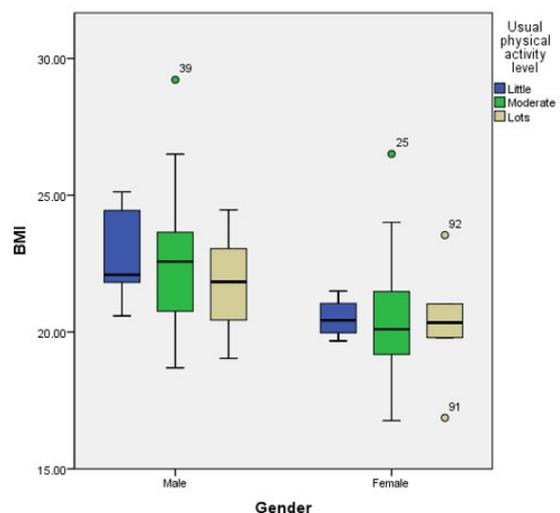
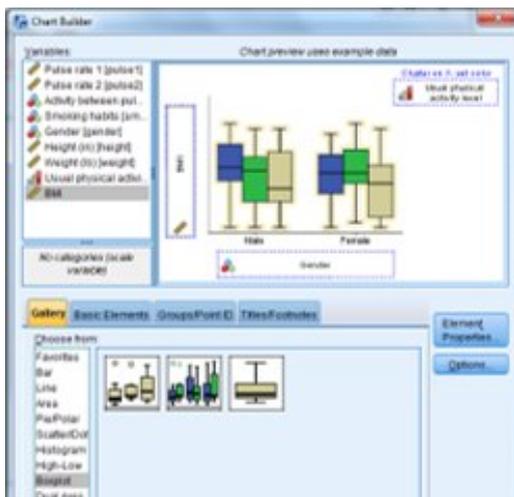
Analyse > Descriptive Statistics > Explore

as before. Now, however, we want to separate out the pulse1 data according to their activity levels; to do this put activity in the Factor List.

(You might want to go into the Plots options and make sure that Histograms and Stem and Leaf plots will not be produced.) A box plot is produced which allows some comparison between the individuals grouped according to activity level. This can also be done by comparing some of the statistics produced in the table above the box plot.

For a more complex box plot, use

Graphs > Legacy Dialogs > Boxplot or Graphs > Chartbuilder



For example, choose a Clustered box plot with pulse1 as the Variable, activity as the Category Axis and Define Clusters by gender. Or, using a ‘simple’ boxplot, put gender as a variable to “Panel by” instead of using it to cluster.

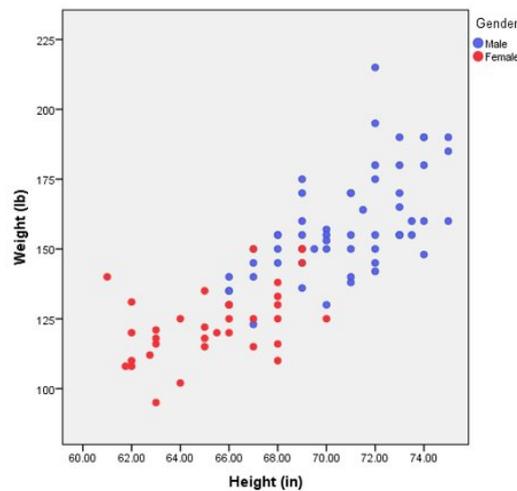
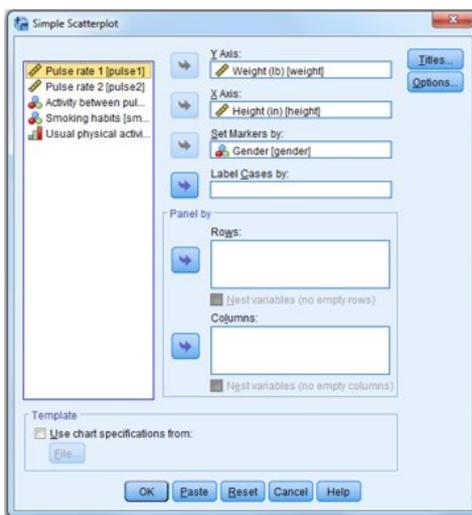
Note that this analysis is suited to a situation where we suspect that the discrete variable influences the scale variable in some way; the plots make less sense if we suspect that the influence is the other way around, in which case other, more complex, methods are required.

### 6.3 Two continuous variables

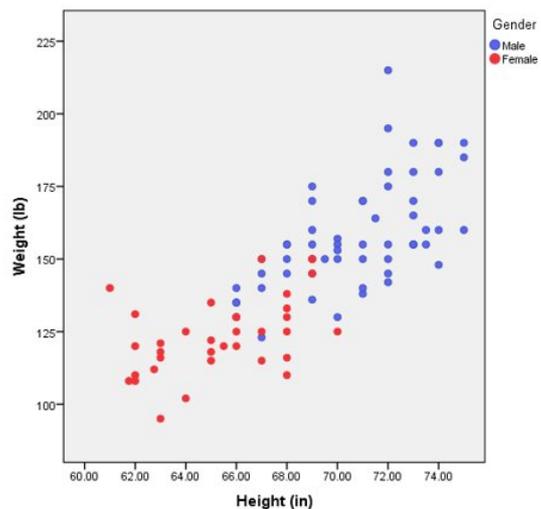
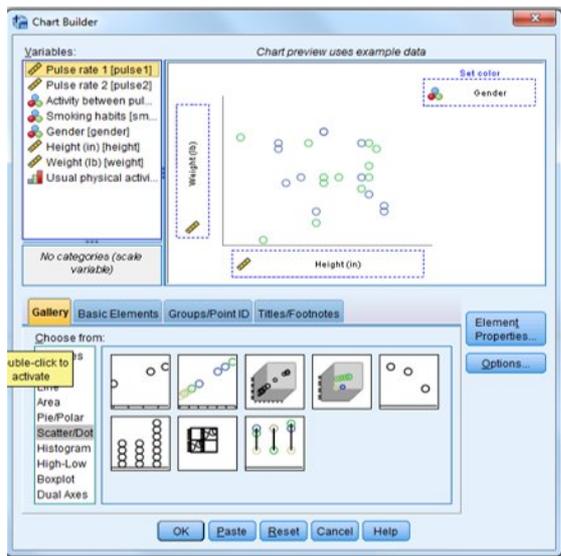
Lastly, we look at a way of exploring the relationship between two continuous (scale) variables. The simplest method is to do a scatterplot, using the menu command

Graphs > Legacy Dialogs > Scatter/Dot.

Note that 2 different colours are used here, for male and female, but are difficult to distinguish in black & white!



Or use Graphs > Chartbuilder



SPSS will also calculate the correlation coefficient and the  $R^2$  value.

## 7 Data Grouping/Selection

Sometimes it is useful to analyse only a certain subset of the data you have collected, or alternatively to compare different subsets of your data.

The Data > Split File menu item allows you to break the data into disjoint subsets and do analyses independently on each subset. Choosing to 'Compare groups' or 'Organise output by groups' affects the way tables in the output are presented, but they will give the same information. Returning to the Split File menu option and selecting 'Analyse all cases, do not create groups' cancels this option.

The menu item Data > Select Cases allows you to temporarily ignore some of the data and analyse one specific subset of the data, e.g. just males. You can restrict which observations (rows of data) SPSS uses in any analysis by changing the contents of the 'Select' box; the 'Output' box controls what happens to the observations that are not selected. Use the SPSS Help to assist you if you're unsure of what the options do.

Data > Sort will sort the data into order, e.g. to order records by region.

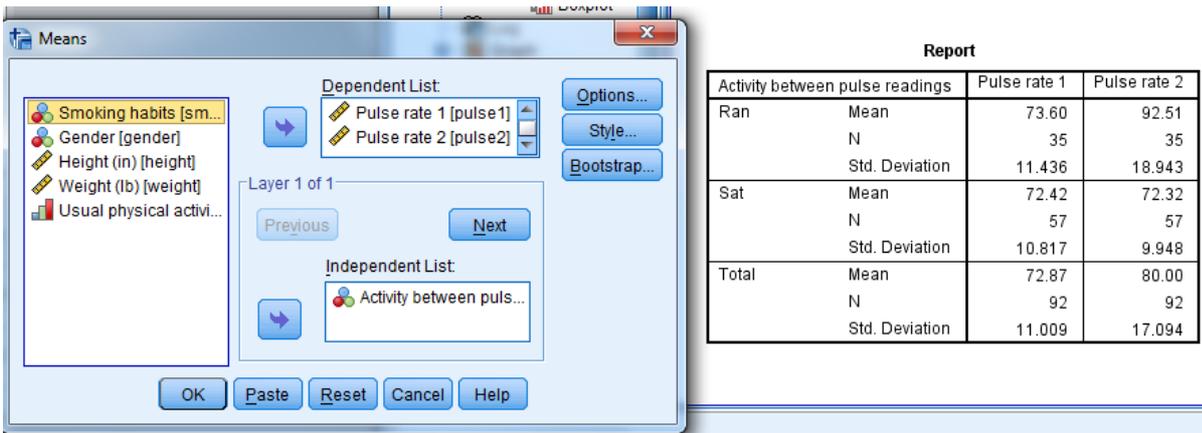
## 8 Comparison of Means

### 8.1 Calculation of means for different groups

Example: You may want to see whether exercise increases the pulse rate.

Choose Analyze > Compare Means

then drag variables into the boxes as shown.



The screenshot shows the SPSS Means dialog box on the left and a report table on the right. The dialog box has a list of variables on the left: Smoking habits [sm...], Gender [gender], Height (in) [height], Weight (lb) [weight], and Usual physical activi... The 'Dependent List' contains 'Pulse rate 1 [pulse1]' and 'Pulse rate 2 [pulse2]'. The 'Independent List' contains 'Activity between puls...'. There are buttons for 'Options...', 'Style...', 'Bootstrap...', 'Previous', 'Next', 'OK', 'Paste', 'Reset', 'Cancel', and 'Help'. The report table on the right is titled 'Report' and shows the following data:

Activity between pulse readings		Pulse rate 1	Pulse rate 2
Ran	Mean	73.60	92.51
	N	35	35
	Std. Deviation	11.436	18.943
Sat	Mean	72.42	72.32
	N	57	57
	Std. Deviation	10.817	9.948
Total	Mean	72.87	80.00
	N	92	92
	Std. Deviation	11.009	17.094

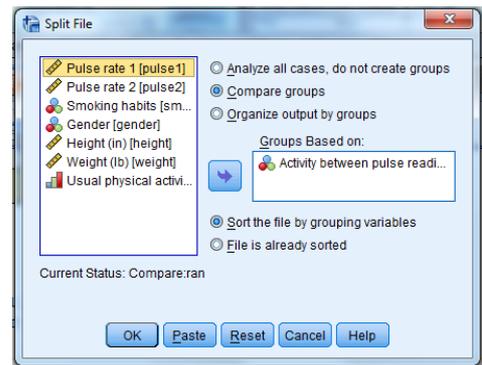
The mean pulse rate appears to be higher after running, but unchanged after sitting. However, the standard deviation is large so we may want to check that this increase is not just part of the inherent variability of the data.

## 8.2 T-tests

These are used to compare the means of 2 groups of Normal scale data.

You could split the data into those who ran and those who sat (use Data > Split File), then test the hypothesis that there is no difference between the pulse rate before and the pulse rate after the activity for these two groups.

Use Analyze > Compare Means > Paired t test



**Paired Samples Statistics**

Activity between pulse readings			Mean	N	Std. Deviation	Std. Error Mean
Ran	Pair 1	Pulse rate 1	73.60	35	11.436	1.933
		Pulse rate 2	92.51	35	18.943	3.202
Sat	Pair 1	Pulse rate 1	72.42	57	10.817	1.433
		Pulse rate 2	72.32	57	9.948	1.318

**Paired Samples Correlations**

Activity between pulse readings			N	Correlation	Sig.
Ran	Pair 1	Pulse rate 1 & Pulse rate 2	35	.607	.000
Sat	Pair 1	Pulse rate 1 & Pulse rate 2	57	.923	.000

**Paired Samples Test**

Activity between pulse readings			Paired Differences				t	df	Sig. (2-tailed)	
			Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
						Lower	Upper			
Ran	Pair 1	Pulse rate 1 - Pulse rate 2	-18.914	15.050	2.544	-24.084	-13.745	-7.435	34	.000
Sat	Pair 1	Pulse rate 1 - Pulse rate 2	.105	4.161	.551	-.999	1.209	.191	56	.849

The data shows that the pulse rate increases after running (  $t = -7.435$ ,  $df = 34$ ,  $p < 0.001$ ). Also, the 95% confidence interval for the change in pulse rate is -24.084 to -13.745, indicating that the pulse rate before running was less than that after running.

## 9 Resources

- The SPSS Help menu: Topics search, Tutorials and Statistics Coach can be very useful.
- Books: ask colleagues for recommendations or just look in the library for something that suits you. Two suggestions are: "SPSS for Psychologists" by Brace, Kemp & Snelgar and "Discovering Statistics using IBM SPSS Statistics" by Andy Field.
- The internet: the web has many examples, tutorials, guides, etc. that various people have put up with the intention of helping people in a similar position to you. They can be very useful if you're not sure about something related to either SPSS or statistics in general. Try
  - Statstutor, <http://www.statstutor.ac.uk/>, which has worksheets and guides on many commonly used statistical techniques
  - Laerd Statistics: e.g. information on data input is found at <https://statistics.laerd.com/spss-tutorials/entering-data-in-spss-statistics.php>

## APPENDIX 1: Useful Commands in SPSS

1. Menu: Main options	page	19
2. Data Summaries (Analyze > Descriptive Statistics.)		20
3. Comparison of means, Confidence Intervals		20
4. Some commonly used statistics including Non-parametric tests		20
5. Graphs		22
6. Other useful information and output		24

### 1. Menu Main Options and some useful commands

**File** the usual things such as Open (including opening Excel files), Save, Print

**Edit** Undo/Redo, Copy/Paste, Insert Cases, Find, Go to case (to find an item)

**Data** Sort Cases, Select cases, Transpose variables, Split File

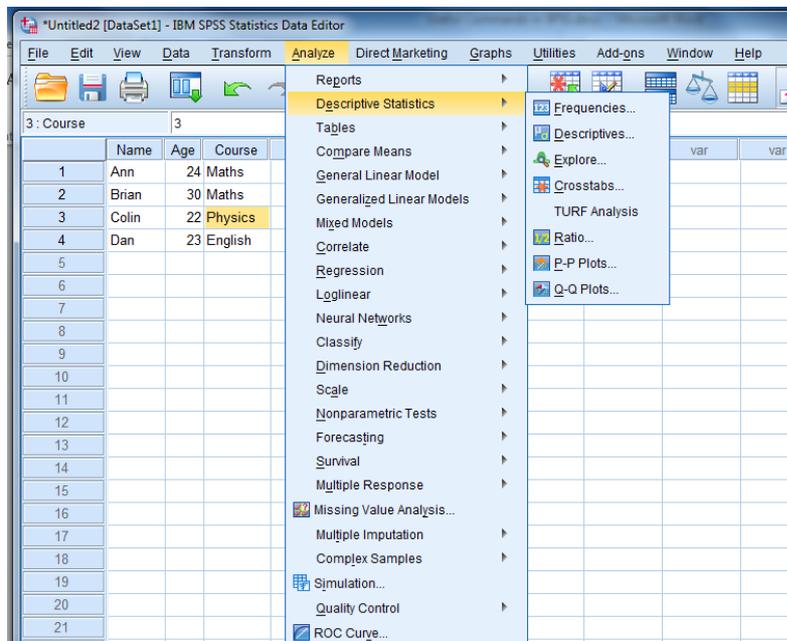
**Transform** Recode (to group data) into the same or a different variable,  
Compute (to calculate another value based on other variables, e.g. calculating BMI from height and weight)

**Analyze** Used a lot. See sections below.

**Graphs** Used a lot. See below.

View, Direct Marketing, Utilities, Add-ons, Window can be ignored for now

### Analyze



#### Analyze Main options:

- Descriptive Statistics,
- Compare means,
- General Linear Model,
- Correlate,
- Regression,
- Non-parametric tests.

#### Other options include

- Dimension Reduction (cluster analysis, factor analysis)
- Scale (Reliability analysis,
- Etc.

## 2. Data Summaries

Analyze > Descriptive Statistics > .....

**Frequencies** *ordinal or nominal* data only. Used to count the numbers in each category and provide a frequency table (not appropriate for continuous data)

Statistics tab: allows calculation of means, quartiles, range, sd etc.

Charts tab: allows selection of barchart, piechart or histogram.

A Normal curve can be added to a histogram by double clicking on the chart and choosing Add Distribution.

**Descriptives** *scale* data to calculate minimum, maximum, mean, standard deviation

**Explore** *scale or ordinal data, discrete data with a large range of values*

Used to get statistics for different groups e.g. summaries for people on different courses. Options include descriptive statistics (mean, sd, median, IQ range, confidence interval for mean etc) and box plots.

**Crosstabs** (*categorical data*) to produce contingency tables to investigate relationships between two variables). Use this option to choose a Chi-squared test for association.

**Q-Q Plot** To produce a plot which lets you test for Normality. It compares the observed values with the expected Normal values. (Note that we could also get a P-P plot, but Q-Q plots are usually preferred as they compare the quantiles, not the proportions, so get better resolution at the tails).

## 3. Comparison of Means

Analyze > Compare Means

**Means** Calculates the means and s.d. of specified groups (e.g. Weight at start for the two treatment groups). Other statistics can also be chosen, e.g. skewness, range, standard error etc.

**t-tests** Choice of one sample, independent samples and paired samples. Options include confidence intervals

**One-way ANOVA** To compare *several* means (see below).

## 4. Some Commonly-used statistical techniques

### 4.1 ANOVA

ANOVA tables are obtained from:

**Analyze > Compare means > One way ANOVA,**

**Analyze > Regression > Linear regression,**

**Analyze > General Linear Model >**



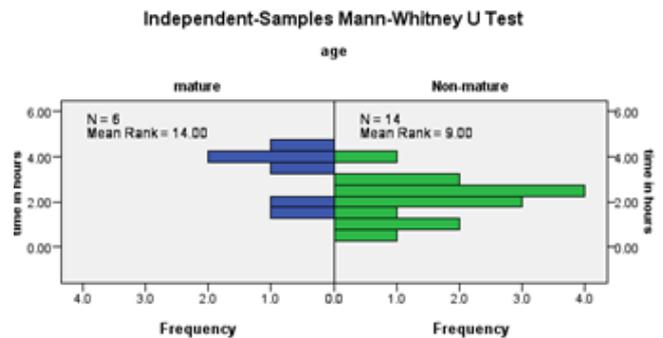
Double-clicking on the output table will produce additional charts etc. which can be very helpful.

**Example** Hours spent in library by mature and non-mature students

Hypothesis Test Summary			
Null Hypothesis	Test	Sig.	Decision
1 The distribution of time in hours is the same across categories of age.	Independent-Samples Mann-Whitney U Test	.091 <sup>1</sup>	Retain the null hypothesis.

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

<sup>1</sup>Exact significance is displayed for this test.



Total N	20
Mann-Whitney U	21.000
Wilcoxon W	126.000
Test Statistic	21.000
Standard Error	12.001
Standardized Test Statistic	-1.750
Asymptotic Sig. (2-sided test)	.080
Exact Sig. (2-sided test)	.091

#### 4.4 Factor Analysis and Reliability

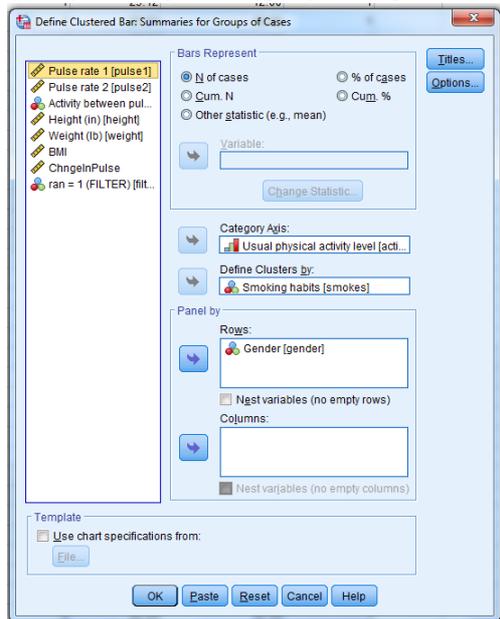
**Analyze > Dimension Reduction > Factor** for factor analysis

**Analyze > Scale > Reliability Analysis** for Cronbach's Alpha, Split-half reliability, Inter-item correlations and covariances, Intra-class correlation coefficient etc.

### 5. Graphs

**Graphs > Legacy Dialogs > Bar** groups shown in adjacent panels.

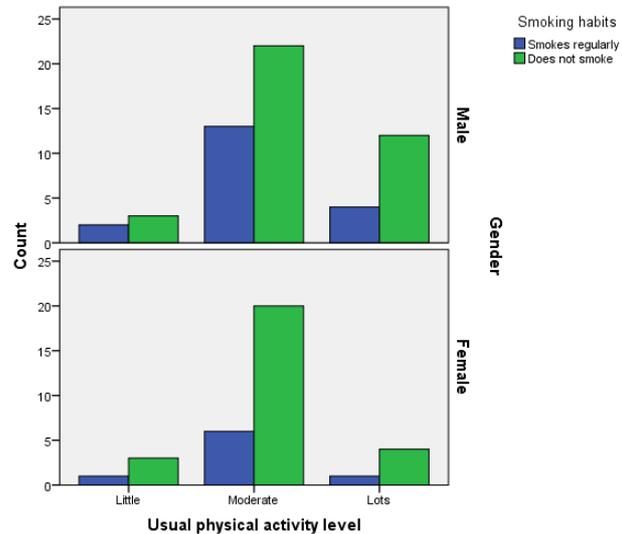
Example: **Graphs > Legacy Dialogs > Bar**



This includes clustered and stacked charts, with different

**Define**

Produces a clustered bar chart to compare smoking habits of the different physical activity groups, split by gender.



**Graphs > Legacy Dialogs >**

options for

**Boxplot**

**Scatter/Dot**

**Histogram**

**To edit charts**

**Change interval widths**

**Add a line of best fit**

**Include a reference line**

**Add a distribution curve**

double click on the chart

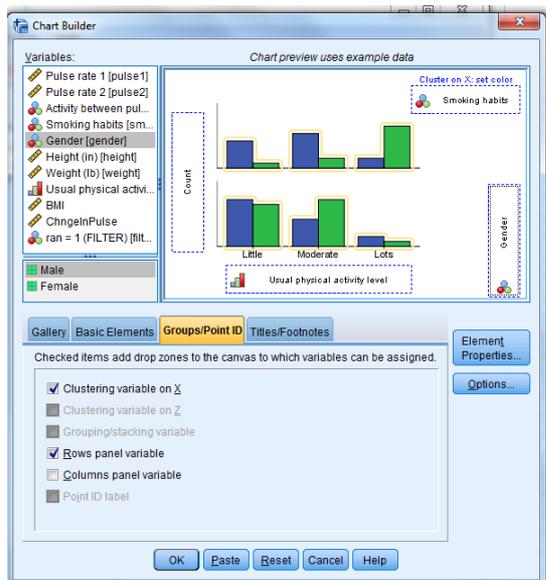
in Properties > Binning or Properties > Scale

Elements > Fit Line

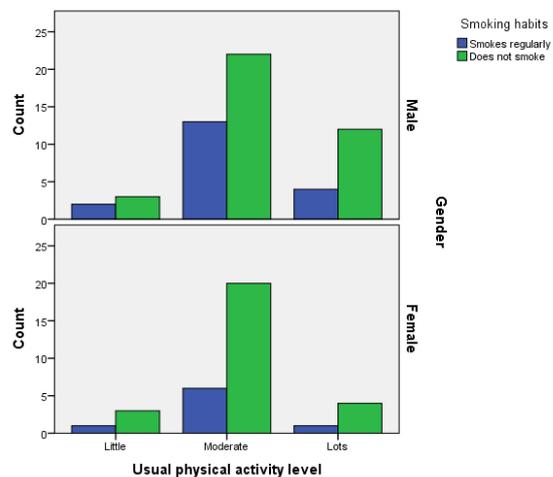
Options > reference line

Elements > Show distribution curve

Graphs can also be obtained from Graphs > Chart Builder (drag items into the boxes)



Produces this:



(cf. the one from the legacy dialogues above)

## 6. Other useful information

### 6.1 Input

Value Labels Button (use when in data view to see the meaning of the coded data)

In Variable view, in the “Values” column, input values and labels for variables (click on the cell, then on the blue box at the right-hand side)

File > Open > Data (enables you to open and input an Excel file, or data from appropriately formatted text files)

### 6.2 Recoding a Variable

Transform > Recode into Different Variables (usually better than “into Same variables)

Transform > Compute variable (e.g. to change from Imperial to Metric units)

### 6.3 Initial Checks

Data > Sort Cases (to sort the data by one or more variables)

Analyse > Descriptive Statistics > Descriptives (to check the amount of data, its location, range and spread)

Analyse > Descriptive Statistics > Frequencies (to check the distribution)

Click on Charts to get a barchart or pie chart or histogram

Analyse > Descriptive Statistics > Explore

Click on Plots to get a histogram and stem & leaf. (tick the box to get a Normal curve).

(Use Explore rather than Frequencies when summarising discrete items where the range of values is large: Frequencies gives too many groups in this case).

Chart Editor double click on the chart

Graphs > Legacy Dialogs > Histogram (tick the box if you want to “Display Normal Curve”)

### 6.4 Data Grouping and selection

Data > Split File

Data > Select Cases

### 6.5 Testing for Normality

1. Draw a histogram and tick “Display Normal Curve”
2. Q-Q plot Analyse > Descriptive Statistics > Q-Q Plots
3. Analyze > Nonparametric Tests > Legacy Dialogs > 1-Sample K-S (for a Kolmogorov-Smirnov Test).

### 6.6 Output

Copy and Paste items to Word or Excel

Or right click and use Export to set up a new file.

**Deciding on appropriate statistical methods for your research:**

- What is your research question?
- Which variables will help you answer your research question and which is the dependent variable?
- What type of variables are they?
- Which statistical test is most appropriate? Should a parametric or non-parametric test be used?

**KEYWORDS:**

**VARIABLE:** Characteristic which varies between independent subjects.

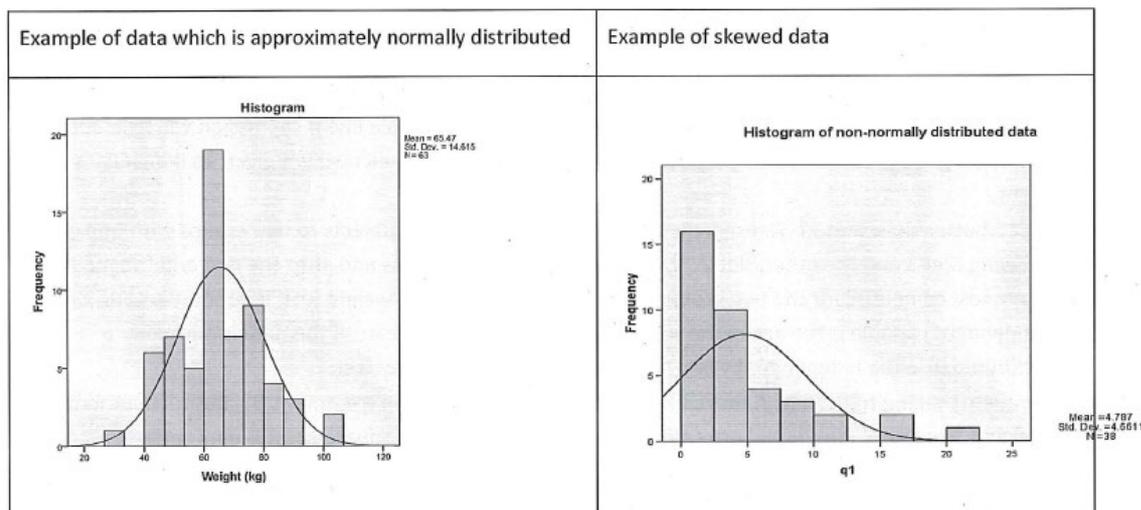
**CATEGORICAL VARIABLES:** variables such as gender with limited values. They can be further categorised into **NOMINAL** (naming variables where one category is no better than another e.g. hair-colour) and **ORDINAL**, (where there is some order to the categories e.g. 1<sup>st</sup>, 2<sup>nd</sup> 3<sup>rd</sup> etc).

**CONTINUOUS (SCALE) VARIABLES:** Measurements on a proper scale such as age, height etc.

**INDEPENDENT VARIABLE:** The variable we think has an effect on the dependent variable.

**DEPENDENT VARIABLE:** The variable of interest which could be influenced by independent variables.

**PARAMETRIC TESTS:** there are various assumptions for parametric tests including the assumption that continuous dependent variables are normally distributed. There are specific tests for this within packages such as SPSS but plotting a histogram is also a good guide. As long as the histogram of the dependent variable peaks in the middle and is roughly symmetrical about the mean, we can assume the data is normally distributed (see examples below).



<b>Table of tests</b>	Dependent (outcome) variable	Independent (explanatory) variable	Parametric test	Non-parametric alternative
<b>Comparing means</b>				
The averages of two INDEPENDENT groups	Scale	Nominal/ binary	Independent t-test	Mann-Whitney test (Wilcoxon rank sum)
The averages of 3+ independent groups	Scale	Nominal	One-way ANOVA	Kruskal-Wallis test
The averages of 2 paired (matched) samples e.g. weight before and after a diet	Scale	Nominal Time/condition variable	Paired t-test	Wilcoxon signed rank test
The 3+ measurements on the same subject	Scale	Nominal	Repeated measures ANOVA	Friedman test
<b>Investigating relationships</b>				
Relationship between 2 continuous variables	Scale	Scale	Pearson's Correlation Coefficient.	Spearman's Correlation Coefficient.
Predicting the value of one variable from the value of a predictor variable	Scale	Any number of scale or binary	Simple Linear Regression	Transform the data
	Binary	Any number of scale or binary	Logistic regression	
Assessing the relationship between 2 Nominal variables	Nominal	Nominal		Chi-squared test

*Note: The table only shows the most common tests for simple analysis of data.*

### Examples:

**Are height and weight related?** Both are continuous variables so Pearson's Correlation Co-efficient would be appropriate if the variables are both normally distributed.

**Can height predict weight?** You cannot determine height from weight, but you could estimate weight given height, so height is the continuous independent variable. Simple linear regression will help decide if weight is a good predictor of height and produce an equation to predict weight given an individual's height.

**Is Diet 1 better than Diet 2?** A researcher would randomly allocate subjects to two groups with one group following Diet 1 and the other Diet 2. Weight would be taken before and after the diet and the mean weight lost compared for the two groups. The dependent variable 'weight lost' is continuous and the independent variable is the group the subject is in, which is categorical. If the data is normally distributed, use the independent t-test, if not use the Mann-Whitney test.

**Are patients taking treatment A more likely to recover than those on treatment B?** Both 'Treatment' (A or B) and 'Recovery' (Yes or No) are categorical variables so the Chi-squared test is appropriate.

## One scale dependent and several independent variables

1 <sup>st</sup> independent	2 <sup>nd</sup> 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

Regression or ANOVA? Use regression if you have only scale or binary independent variables. Categorical variables can be recoded to dummy binary variables but if there are a lot of categories, ANOVA is preferable.

### Index

ANOVA  
 APPENDIX 1: Useful commands in SPSS  
 APPENDIX 2: What statistical test do I need?  
 Bar charts  
 Boxplots  
 Calculating a new variable  
 Categorical variables  
 Chi-squared test  
 Cross-tabulation  
 Correlation  
 Data input  
 Data Select Cases / Split File / Sort  
 Descriptive statistics  
 Discrete variables  
 Downloading data  
 Exporting output  
 Graphs  
 Histogram  
 Kurtosis  
 Means  
 Menu: main options  
 Menu: Analyze  
 Non-parametric tests  
 Normality, testing for  
 On-line survey (BOS)  
 Output  
 Q-Q plots  
 Recoding a variable  
 Regression  
 Relationship between two variables  
 Resources  
 Scatter diagrams  
 Scale variables  
 Skew  
 Statistical tests, Appendix 2  
 Summaries for categorical, scale discrete variables  
 Summary statistics  
 Testing for normality  
 Transform > Compute variable  
 Transform > Recode  
 T- tests

### Page

20, 21  
 19  
 25  
 5, 22  
 15, 21  
 9  
 4, 14  
 14, 20, 21  
 14, 20, 21  
 21  
 1, 24  
 17  
 4 to 6, 14, 20  
 7  
 3  
 10  
 15, 16, 22  
 12  
 12  
 17, 20  
 19  
 4, 6, 17 to 22  
 13, 21  
 11  
 2  
 10  
 12, 20  
 7  
 21  
 14  
 18  
 16, 21  
 5  
 12  
 25 to 28  
 4, 5, 7  
 4, 20  
 11 to 14  
 9  
 7  
 18, 20

Mary Lorimer, 19/06/19

# Flow Chart for statistical tests

