

Analysing data using SPSS

(A practical guide for those unfortunate enough to have to actually do it.)

Andrew Garth, Sheffield Hallam University, 2008

Contents:

What this document covers...	2
Types of Data.	3
Structuring your data for use in SPSS	6
Part 1 - Creating descriptive statistics and graphs.	11
SPSS versions	11
Entering and saving Data.	11
Saving Your Work	12
Looking at the Data	15
Exploring the data.	16
More on drawing Boxplots	16
Using Descriptive Statistics	17
More on different types of data	19
The difference between Mean and Median	19
Standard Deviation (S.D.) what is it?	21
Histograms and the Normal Distribution	25
Bar charts.	30
Using Scatterplots to look for correlation	34
Line graphs.	36
Pie charts	40
Part 2 - Inferential Statistics.	43
From Sample to Population...	43
A Parametric test example	44
Using a Non-parametric Test	47
Observed Significance Level	49
Asymptotic significance (asyp. Sig.)	50
Exact significance (exact sig.)	50
Testing Paired Data	51
Correlation	53
Significance in perspective.	56
Looking for correlation is different from looking for increases or decreases	57
Correlation: descriptive and inferential statistics	57
What have we learned so far?	58
Test decision chart.	59
The Chi-Square Test.	62
Cross-tabulation	62
Some examples to get your teeth into.	68
Analysis of Variance - one-way ANOVA	71
Repeated measures ANOVA.	77
Making sense of the repeated measures ANOVA output.	78
Inter-Rater Agreement using the Intraclass Correlation Coefficient	82
Cronbach's Alpha	83
Inter rater agreement using Kappa	84
Calculating the sensitivity and specificity of a diagnostic test	86
Copying information from SPSS to other programs	87
More about parametric and nonparametric tests	89
Creating a new variable in SPSS based on an existing variable	91

Acknowledgements.

Thanks are due to Jo Tomalin whose original statistical resources using the Minitab software were invaluable in developing this resource. Thanks also go to the numerous students and colleagues who have allowed the use of their research data in the examples.



What this document covers...

This document is intended to help you draw conclusions from your data by statistically analysing it using SPSS (Statistical Package for the Social Sciences). The contents are set out in what seems a logical order to me however if you are in a rush, or you don't conform to my old fashioned linear learning model then feel free to jump in at the middle and work your way out! Most researchers will be working to a protocol that they set out way before gathering their data, if this is the case then theoretically all you need to do is flip to the pages with the procedures you need on and apply them. It is however my experience that many researchers gather data and then are at a loss for a sensible method of analysis, so I'll start by outlining the things that should guide the researcher to the appropriate analysis.

Q. How should I analyse my data?

A. It depends how you gathered them and what you are looking for.

There are four areas that will influence your choice of analysis;

- 1 *The type of data you have gathered, (i.e. Nominal/Ordinal/Interval/Ratio)*
- 2 *Are the data paired?*
- 3 *Are they parametric?*
- 4 *What are you looking for? differences, correlation etc?*

These terms will be defined as we go along, but also remember there is a glossary as well as an index at the end of this document.

This may at first seem rather complex, however as we go through some examples it should be clearer.

I'll quickly go through these four to help start you thinking about your own data.

The type of data you gather is very important in letting you know what a sensible method of analysis would be and of course if you don't use an appropriate method of analysis your conclusions are unlikely to be valid. Consider a very simple example, if you want to find out the average age of cars in the car park how would you do this, what form of average might you use? The three obvious ways of getting the average are to use the mean, median or mode. Hopefully for the average of car you would use the mean or median. How might we though find the average colour of car in the car park? It would be rather hard to find the mean! for this analysis we might be better using the mode, if you aren't sure why consult the glossary. You can see then even in this simple example that different types of data can lend themselves to different types of analysis.

In the example above we had two variables, *car age* and *car colour*, the data types were different, the age of car was *ratio* data, we know this because it would be sensible to say "one car is twice as old as another". The colour however isn't ratio data, it is categorical (often called *nominal* by stats folk) data.



Types of Data.

Nominal Data: These are data which classify or categorise some attribute they may be coded as numbers but the numbers has no real meaning, its just a label they have no default or natural order. *Examples:, town of residence, colour of car, male or female (this lat one is an example of a dichotomous variable, it can take two mutually exclusive values).*

Ordinal Data: These are data that can be put in an order, but don't have a numerical meaning beyond the order. So for instance, the difference between 2 and 4 in the example of a Lickert scale below might no be the same as the difference between 2 and 5. *Examples: Questionnaire responses coded: 1 = strongly disagree, 2 = disagree, 3 = indifferent, 4 = agree, 5 = strongly agree. Level of pain felt in joint rated on a scale from 0 (comfortable) to 10 (extremely painful).*

Interval Data: These are numerical data where the distances between numbers have meaning, but the zero has no real meaning. With interval data it is not meaningful to say than one measurement is twice another, and might not still be true if the units were changed. *Example: Temperature measured in Centigrade, a cup of coffee at 80°C isn't twice as hot a one at 40°C.*

Ratio Data: These are numerical data where the distances between data and the zero point have real meaning. With such data it is meaningful to say that one value is twice as much as another, and this would still be true if the units were changed. *Examples: Heights, Weights, Salaries, Ages. If someone is twice as heavy as someone else in pounds, this will still be true in kilograms.*

More restricted
in how they can
be analysed

Less restricted
in how they can
be analysed

Typically only data from the last two types might be suitable for parametric methods, although as we'll see later it isn't always a completely straight forward decision and when documenting research it is reasonable to justify the choice of analysis to prevent the reader believing that the analysis that best supported the hypothesis was chosen rather than the one most appropriate to the data. The important thing in this decision, as I hope we'll see, is not to make unsupported assumptions about the data and apply methods assuming "better" data than you have.

Are your data paired?

Paired data are often the result of before and after situations, e.g. before and after treatment. In such a scenario each research subject would have a pair of measurements and it might be that you look for a difference in these measurements to show an improvement due to the treatment. In SPSS that data would be coded into two columns, each row would hold the before and the after measurement for the same individual.

We might for example measure the balance performance of 10 subjects with a Balance Performance Monitor (BPM) before and after taking a month long course of exercise



designed to improve balance. Each subject would have a pair of balance readings. This would be paired data. In this simple form we could do several things with the data; we could find average reading for the balance (Means or Medians), we could graph the data on a boxplot this would be useful to show both level and spread and let us get a feel for the data and see any outliers.

In the example as stated above the data are paired, each subject has a pair of numbers.

What if you made your subjects do another month of exercise and measured their balance again, each subject would have three numbers, the data would still be paired, but rather than stretch the English language by talking about a pair of three we call this repeated measures. This would be stored in three columns in SPSS.

A word of warning, sometimes you might gather paired data (as above, before we pretended there was a third column of data) but end up with independent groups. Say, for example, you decided that the design above was flawed (which it is) and doesn't take into account the fact that people might simply get better at balancing on the balance performance monitor due to having had their first go a month before. i.e. we might see an increase in balance due to using the balance monitor! to counter this possible effect we could recruit another group of similar subjects, these would be assessed on the BPM but not undertake the exercise sessions, consequently we could assess the effect of measurement without exercise on this control group. We then have a dilemma about how to treat the two sets of data. We could analyse them separately and hope to find a significant increase in balance in our treatment group but not in the non exercise group. A better method would be to calculate the change in balance for each individual and see if there is a significant difference in that change between the groups. This latter method ends with the analysis actually being carried out on non-paired data. (An alternative analysis would be to use a two factor mixed factorial ANOVA - but that sounds a bit too hard just now! - maybe later.)

If you are not sure whether two columns of data are paired or not, consider whether rearranging the order of one of the columns would affect your data. If it would, they are paired. Paired data often occur in 'before and after' situations. They are also known as 'related samples'. Non-paired data can also be referred to as 'independent samples'. Scatterplots (also called scattergrams) are only meaningful for paired data.

Parametric or Nonparametric data

Before choosing a statistical test to apply to your data you should address the issue of whether your data are parametric or not. This is quite a subtle and convoluted decision but the guide line here should help start you thinking, remember the important rule is not to make unsupported assumptions about the data, don't just assume the data are parametric; you can use academic precedence to share the blame "Bloggs et. al. 2001 used a t-test so I will" or you might test the data for normality, we'll try this later, or you might decide that given a small sample it is sensible to opt for nonparametric methods to avoid making assumptions.

- Ranks, scores, or categories are generally non-parametric data.
- Measurements that come from a population that is normally distributed can usually be treated as parametric.



If in doubt treat your data as non-parametric especially if you have a relatively small sample.

Generally speaking, parametric data are assumed to be normally distributed – the normal distribution (approximated mathematically by the Gaussian distribution) is a data distribution with more data values near the mean, and gradually less far away, symmetrically. A lot of biological data fit this pattern closely. To sensibly justify applying parametric tests the data should be normally distributed.

If we you unsure about the distribution of the data in our target population then it is safest to assume the data are non-parametric. The cost of this is that the non parametric tests are generally less sensitive and so you would stand a greater chance of missing a small effect that does exist.

Tests that depend on an assumption about the distribution of the underlying population data, (e.g. t-tests) are parametric because they assume that the data being tested come from a normally distributed population (i.e. a population we know the parameters of). Tests for the significance of correlation involving Pearson's product moment correlation coefficient involve similar assumptions.

Tests that do not depend on many assumptions about the underlying distribution of the data are called non-parametric tests. These include the Wilcoxon signed rank test, and the Mann-Whitney test and Spearman's rank correlation coefficient. They are used widely to test small samples of ordinal data. There is more on this later.

Are you looking for differences or correlation?

- You can look for differences whenever you have two sets of data. (*It might not always be a sensible thing to do but you can do it!*)
- You can only look for correlation when you have a set of paired data, i.e. two sets of data where each data point in the first set has a partner in the second. If you aren't sure about whether your data are paired review the section on paired data.
- You might therefore look for the difference in some attribute before and after some intervention.

Ponder these two questions...

1. Does paracetamol lower temperature?
2. Does the number of exercises performed affect the amount of increase in muscle strength?

Which of these is about a difference and which is addressing correlation? - well they aren't all that well described but I recon the first on is about seeing a difference and the second is about correlation, i.e. does the amount of exercise correlate with muscle strength, whereas the first is about "does this drug make a difference".

A variant on this is when conducting a reliability study, in many respects the data structure is similar to a correlational experiment however the technique used to analyse the data is different.



Structuring your data for use in SPSS

The way you lay out your data in SPSS will depend upon the kind of data you have and the analysis you propose to carry out. However there are some basic principals that apply in all situations.

- 1 SPSS expects you to put each case on a row. Usually this means that each research subject will have a row to their self.
- 2 Categorical variables are best represented by numbers even if they are not ordered categories, they can then be ascribed a text label using the "Variable Labels" option.
- 3 The variable name that appears at the top of the column in SPSS is limited in length and the characters it will hold, the variable label can hold a more meaningful description of the variable and will be used on output (graphs etc.) if you fill it in.
- 4 If you have two (or more) groups of subjects each subject will still have a row to their self, however you will need to dedicate a variable (column) to let the system know which group each subject belongs to.

Examples of some typical data structures are below;

Two Independent Groups of data. (This structure would arise from what stats books might call a between groups experiment.) These data were gathered as part of an investigation into the effect of horse riding on balance. Swayarea is a measure of balance, or more correctly, unbalance, a small value indicates good balance. The variable called "rider" discriminates between riders and non-riders, it can be referred to as a discriminatory variable.

	Subject	swayarea	rider
	1	496	1.00
	2	791	1.00
	3	1,309	1.00
	4	1,746	1.00
	5	223	1.00
	6	940	0.00
	7	3,300	0.00
	8	1,571	0.00
	9	2,444	0.00
	10	970	0.00

Data View Variable View

To set up a "Value Label" to give meaning to this variable first, click on the "Variable View" tab at the bottom of the data screen, second, in the variable view screen, notice that each variable now occupies a row, and the columns represent the attributes of that variable, the rider variable is numeric, 11 characters wide with no decimal places. On graphs

*horse.sav [DataSet1] - SPSS Data Editor										
File Edit View Data Transform Analyze Graphs Utilities Window Help										
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	Subject	Numeric	11	0		None	None	8	Right	Scale
2	swayarea	Numeric	11	0		None	None	8	Right	Scale
3	rider	Numeric	11	0	Horse rider	{0, Non-rider}...	None	8	Right	Scale
4										

and other output the variable will be labeled as "Horse rider" rather than just "rider" and some test has been attached to the numeric values stored in the variable. This test gives meaning to the values 1 and 0, it was added by clicking into the grid on the variable view

Value Labels

Value:

Value Label:

Add Change Remove

.00 = "Non rider"
1.00 = "Rider"

OK Cancel Help



where you can now see the text “{0,Non-rider}” and typing the value and the label then clicking the “Add” button in the Value Labe dialog box. This is a really useful method of making the graphs more readable. If you are using Likert scales then the value labels can reflect the meaning of each ordinal category. Labeling variables is good practice regardless of the data structure.

1 = dislike strongly
2 = dislike
3 = ambivalent
4 = like
5 = like strongly

This type of design gets more complex if there are more than two groups, for example if we had Non-riders, Horse-riders and bike-riders. The data would still fit in two columns, one for the measurement and the other to let us know which group they are in. Things get more complex if we bring another grouping variable to the equation, maybe gender, this would need a new variable to sit in, we could though then see if gender affects balance. We could even look at both factors at once (rider and gender) and the effect of one on the other in a clever analysis called Univariate Analysis of Variance, but lets not for now.

Typical structure for simple paired data. (This structure would arise from what stats books might call a within subjects experiment.) Again these data were gathered in a study of balance, a large sway area indicates a more wobbly person. These subjects had to stand on their dominant leg while their balance performance was assessed they then had their leg

Participant Number	Sway Area Before Ice	Sway Area After Ice
1	42	51
2	158	336
3	67	125
4	557	3406
5	121	52
6	50	44
7	40	113
8	85	268
9	171	402
10	232	462

immersed in iced water for a period and were tested again. We have a measurement taken before and after a treatment. These data are paired. The research question was asking if the reduced temperature adversely affected balance so the researcher was looking for a difference in sway area before and after the treatment. We could also use these data to look for correlation since they are paired. We would, I think, expect to find positive correlation, i.e. people who are naturally wobbly before having their foot frozen will still be more wobbly afterwards. The before and after data appear in separate columns but each subjects data are adjacent.

It might of course be that case they the subjects have been subjected to more than two conditions, for example our intrepid researcher might have chosen to heat the subjects leg and see if this alters balance. In such a case there would be another adjacent column of data for each additional condition. In such a case the data are again paired but the term repeated measures might better describe the experiment.

Groups of paired data. Sometimes its hard to workout how to struacture the data for example when we have paired data for two or more groups...

In this example, about the effect of exercise on balance, the data are initially paired but we want to find out the effect of an exercise on balance. Group1 have done the exercise and Group 2 are the control – they didn’t do the exercises. We really are interested in the effect of the exercise on balance in each group. To find this out for each group we can calculate the "difference due to treatment" for each individual. One issue here though is that it is important to check that there was no initial difference between the groups, i.e. in the "Sway



Area Before Ice". The ideal way to analyse these data using an inferential technique would be to use a mixed model ANOVA on the before and after values, but this is a little complex for now.

group	Sway area before	Sway area after	Difference in sway area
1	55	46	9
1	343	161	182
1	134	74	60
1	55	124	-69
1	52	52	0
1	117	48	69
2	84	80	4
2	93	88	5
2	46	52	-6
2	233	242	-9
2	51	53	-2
2	123	121	2
2	165	165	0

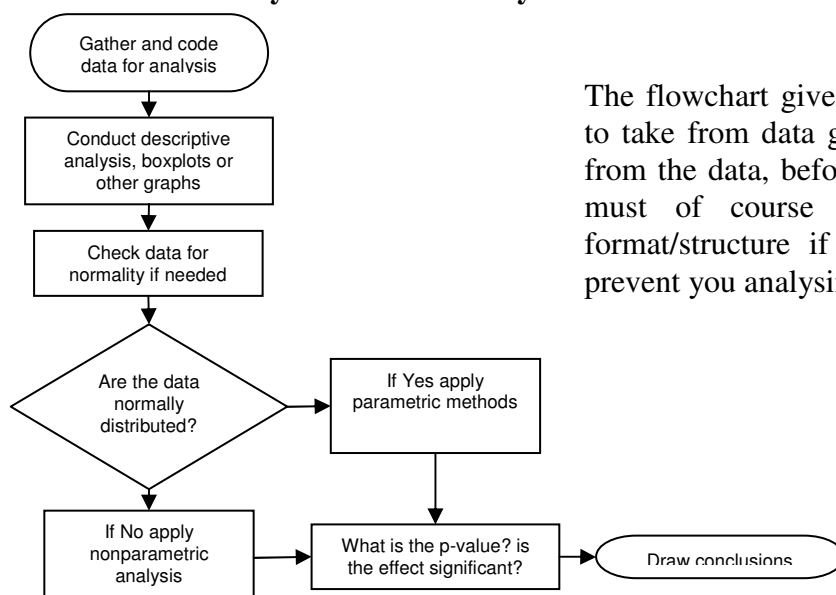
We can get SPSS to calculate the differences for each subject, then we can look at the change in balance between the exercise and non-exercise group. The data we analyse are no longer paired at this stage. We are looking for a difference between the groups. To get SPSS to do the calculation you can use the "Compute" command, it is under the Transform menu – it works just like a calculator – save and backup work before playing! (See appendix 1 for details.) The structure we then get is similar to the two independent groups of data example we considered earlier, we can ignore the two middle columns. We can now look to see if the "difference in sway area" is the same in both groups.

Three or more groups or conditions. Things look more complex when you have three or more groups or conditions but don't worry, it is essentially the same.

When you have **three or more groups** the grouping variable will simply have extra values, e.g. if there were four groups it would take the values 1,2,3 or 4. These would then be labelled as we did in the two independent groups of data example and analysed with descriptive statistics then with a one way ANOVA or the nonparametric equivalent.

If you have **three or more conditions** for the same set of subjects then the data will be paired (using the loosest definition of a pair). The structure will be similar to the within subjects experiment structure (simple paired data) above except that it will have more columns (variables), one more for each extra condition. These data could then be analysed with descriptive statistics then with a repeated measures ANOVA or the nonparametric equivalent.

What is the order you should tackle your data in?



The flowchart gives a rough indication of the steps to take from data gathering to drawing conclusions from the data, before you can analyse the data they must of course be stored in an appropriate format/structure if this structure is wrong it can prevent you analysing the data correctly.



More about Parametric or Nonparametric procedures.

In simple terms the parametric data analysis procedures rely on being fed with data about which the underlying parameters of their distribution is known, typically data that are normally distributed (the normal distribution gives that bell shape on a histogram). This generally makes the parametric procedures more sensitive, so people usually would prefer to apply these if possible. Nonparametric procedures don't care about the underlying data distribution and so are more robust, however we pay for this robustness in sensitivity. Nonparametric procedures are generally less sensitive so there is an increased chance of missing a significant effect when using the rough and ready nonparametric tests. The chance of detecting a significant effect that really does exist is called the statistical power of the experiment. Researchers would like this to be as high as possible, 80% or more is good.

When should we not use the parametric tests in favour of the less sensitive nonparametric equivalents?

Usually we would drop to nonparametric test if the data we are analysing are significantly different to a normally distributed data set; this might be due to the distribution or the presence of outliers. This would be even more appropriate if the sample size is quite small (e.g. below 15 or 20) since one outlier in 15 data points will have a greater effect than one outlier in 1500 data points.

Scores would typically be treated as nonparametric as would ordinal and nominal data.

What are the penalties of getting this wrong?

If you use a parametric test on nonparametric data then this could trick the test into seeing a significant effect when there isn't one. This is very dangerous, proper statisticians call this a "type one error". A type one error is a false positive result.

If you use a nonparametric test on parametric data then this could reduce the chance of seeing a significant effect when there is one. This is not ideal, proper statisticians call this a "type two error". A type two error is a missed opportunity, i.e. we have failed to detect a significant effect that truly does exist.

Of these two errors which is least dangerous? I feel that the type two error is least dangerous. Think of your research question as being a crossroads in knowledge. You are sat in your car at a fork in the road, should you go left or right? A type one error would be to go down the wrong road; you would be actively going in the wrong direction. A type two error would be to sit there not knowing which way is correct, eventually another researcher will come along and hopefully have a map.

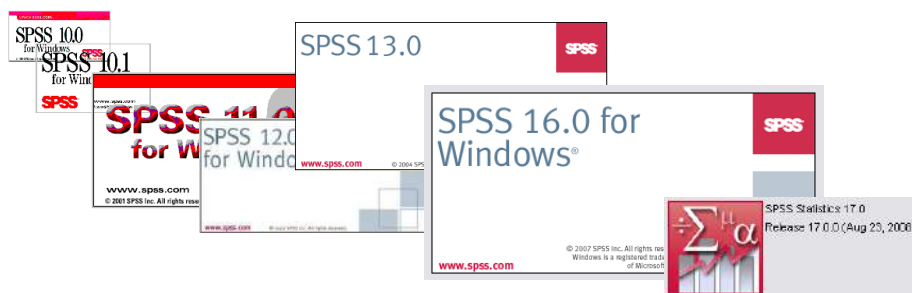
So to summarise; Using a parametric test in the wrong context may lead to a type one error, a false positive. Using a nonparametric test in the wrong context may lead to a type two error, a missed opportunity.

We might address this usefully again when thinking about interpreting p-values.



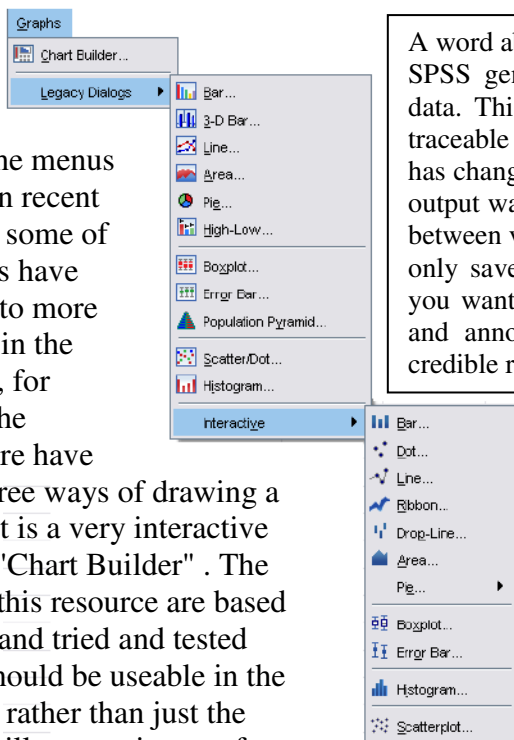
A few points about SPSS versions etc.

The SPSS software is upgraded regularly, at the time of writing we were just starting to use SPSS version 15 but this is now superseded by version 16 and version 17 is on its way. There are three main ways to get a copy of SPSS for a SHU student, it is on SHU PCs, for use at home students can purchase a version from the SHU learning centres and finally the latest version can be downloaded from spss.com for a free limited trial period.



When you first start the program it will ask you if you want to open existing data, if you are starting for the first time then you will probably want to type the data into the editing screen. It is also possible to import data from MS Excel if the structure is suitable (this can allow the user to key the data in using MS Excel at home then import it for analysis), this is pretty standard and hasn't changed radically throughout the generations.

One way that the menus have changed in recent versions is that some of the older menus have been relegated to more obscure places in the menu structure, for example over the generations there have been at least three ways of drawing a graph, the latest is a very interactive method called "Chart Builder". The instructions in this resource are based round existing and tried and tested methods that should be useable in the recent versions rather than just the latest, so you will see a mixture of new and old menus used, don't be frightened to try new methods but do be critical of the results, make sure they make sense before writing them up! On the later (version 16) dialog boxes the buttons have in some cases been moved around so watch out for this.



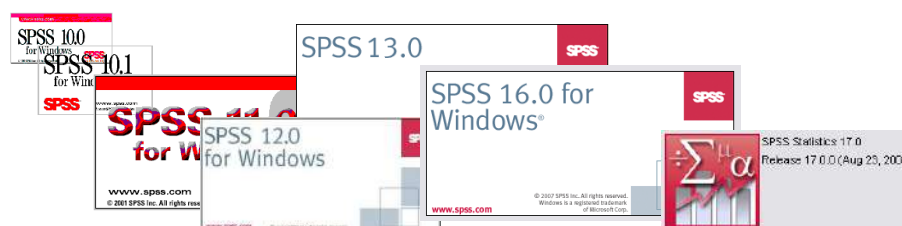
A word about the SPSS Output Viewer...

SPSS generates output in a window separate to the data. This is sensible to avoid confusion and give a traceable output file. Annoyingly the format of this file has changed in the recent versions and although earlier output was viewable in later versions this isn't the case between v15 and v16. My advice is that you should not only save the output but also copy and paste output you want to keep into a separate MSWord document and annotate this to enable you to build it into a credible results section when appropriate.



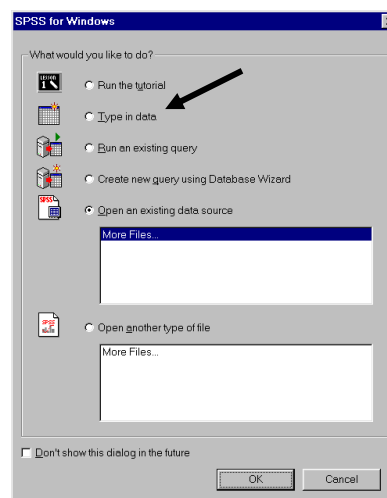
Part 1 - Creating descriptive statistics and graphs.

SPSS



Start SPSS now. To do this on a SHU PC click the **Start** button then choose **Programs, Specialist Applications, Statistics**, then **SPSS**, if you have it on a home PC the SPSS software will be directly under the programs menu.

You should see the SPSS copyright screen then a dialog box inviting you to decide what to do next. Some of the options on it are beyond us for now, the default option (already selected) is open an existing data source, later we will do this but for now select the option that lets you **Type in data**. Then click the **OK** button. (Or you can just click "Cancel".)



SPSS can display a number of different types of windows to interact with the user. The window you should see at the moment is the **SPSS Data Editor** window – this is where you can type in new data or see data that has already been put in. Just to stop things getting too simple this window has two “Views”, “Data View” and “Variable view”; this is actually quite useful, as you will see soon. One shows us the actual data the other shows us how it is stored. Do check you are in the correct view when you enter data!

When you draw a graph or work out a statistic (e.g. an average) you will get the results in a second window, this is essentially another program running, we can switch between the two by using the mouse on the Windows Taskbar at the bottom of the screen. It may seem an added complication at first but it works well in practice, it keeps the results separate from the data.

Task 1. Entering and saving Data.

To begin using SPSS we will type in some data and do a simple analysis.

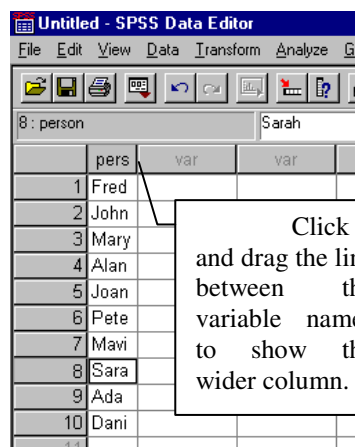
In the box on the right are ten people's names, type them into the first column.

You may notice a problem when you get to Peter. Peter has 5 letters in his name, unfortunately SPSS has assumed all the cases are similar to

Fred	27
John	22
Mary	54
Alan	49
Joan	67
Peter	16
Mavis	46
Sarah	20
Adam	21
Daniel	11



the first one and Peter has become Pete. We can alter this by switching to the **Variable View** (click the tab at the bottom of the SPSS window).



You should see a row of information about variable one (var0001), which is where we are storing these names. Change the **Width** from 4 to 12.

HINT: Click on the box with the number 4 in and increase it to 12

Go back to the **Data View** and type in Peter again.

Finish typing in the peoples' names then go back to the **Variable View** and change the name of this variable to something more meaningful, type in **person** instead of **var00001**.

Go back to the **Data View** - what has changed?

In the next column, type in the peoples' ages. You will notice that SPSS is much happier dealing with numbers than text. This second variable would be better named **age** than **var00002** – have a go at changing its name in the variable view. While in Variable view you can put in a more descriptive label for the variable that will appear on graphs etc, type this in the Label column. Unlike the Name the Label can contain spaces etc.

SPSS loves to give you more information than you need. The dialog boxes that appear on the screen in response to many menu commands also contain many options. The skill you need at this stage (as with most computer packages) is to *ignore* everything you do not need. This document will try to indicate what you need to notice. With more confidence you may choose to experiment with other options, but for the moment don't worry about anything on screen that you are not directed to use or look at.

Saving your work

NOTE: Graphs and analyses will not be saved unless you save them specially.

Before you save the work have a look at the top of the SPSS window, you will notice that your work is currently nameless, you will see **“Untitled – SPSS Data Editor”** in the blue title bar.

To save the data we have just typed in choose **Save** from the **File** menu.  The first time you use this command you will be prompted to give a filename, call it **people** and put your name at the end of the word people, my file is called **peopleandrew**. At the top of the Save-as dialog box is a small section labelled “Save in:” you can click the small button to the right of this to tell SPSS where to put the file if you want to store it somewhere other than that offered.

It is good practice to keep two copies of your data especially when working on original data. Keep one on a floppy disk or USB stick etc. and the other on your student homespace (F: drive - a network drive secured by your password). For the data we are using it doesn't matter too much. It does though matter a lot if you are working on data you have gathered, your data is unique.

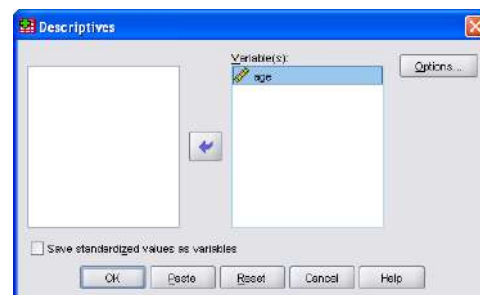


If it has already been saved once and you want to save changes to the data just select **Save** from the file menu and the data is updated on disk. Using the **Save as** command under the File menu lets you save a copy of the file on another drive or with another name. (The best way to control where files are and copy them to floppy is to learn to use Windows Explorer.)

IMPORTANT: this will only save the data in the current Window – Data and Results need to be saved separately.

Now we have some data in the system and it is safely saved we can play with it.

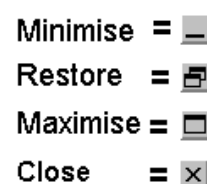
Click on the **Analyse** menu and choose **Descriptive Statistics** then **Descriptives**.



The button between the two windows let you choose the variables to be analysed, in our case the choice is simple, just click the  or  centre button to move the **age** variable over to the right then click **OK**.

After a bit of thinking SPSS should display the results in a separate window, you will see this appear in front of the Data Editor and a new button will appear on the Windows task bar at the bottom of your screen. The new window has a title, have a look in its title bar at the top of its window, what is it?

It should be something like “Output1 – SPSS Viewer”



Use the Windows task bar to switch between the SPSS data and results windows. I find it useful to maximise the window I am looking at, the buttons on the right of the title bar let you do this, the key I’ve put in the text here should help. You can also use Alt-Tab to toggle between windows.

Lets look at the output.

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
age	10	11.00	67.00	33.3000	19.04410
Valid N (listwise)	10				

You will see from the output that the “Descriptives” option gives you a statistical description of the data. It tells us how many cases there are (N=10) the Maximum, Minimum, a form of average called the Mean and a more complex statistic called the Standard Deviation, this last one gives us a numerical indication of how spread out the data are.

Now you have seen the output window you can close it (the cross in the top right corner). When prompted to save the output you can say **No**, we have already saved the data so if needed we could recreate the output. **When the output becomes more complex you will want to save it and probably copy some into MS Word** (see the note on versions).



Data for the tasks to follow.

The data for the tasks that follow is already stored in files for you; these files and much more are available in the BlackBoard system. The important files for downloading and using are stored in compressed format to speed up the transfer.

To access the online learning (Blackboard) site go to the SHU Student Portal.

(If you have difficulty getting onto Blackboard then the essential data is available at:
<http://teaching.shu.ac.uk/hwb/ag/resources/resourceindex.html>)

The next 30 or so pages cover a range of statistics called Descriptive Statistics. These methods use graphs and simple numerical methods to describe your sample. Later in the document we start to look at Inferential statistics, these allow you to gauge how strongly your findings in the sample are likely to relate back to the population you are studying, i.e. are the findings in the sample likely to be derived just by chance or do they evidence some real effect.



Task 2. Looking at the Data

For this task we are going to look at some data collected by an Occupational Therapy student, looking at how age affected OT students' participation in discussion in class. She counted how many times each student contributed orally in a period totalling 12 hours of classes. The students were from the 1st and 2nd years of the course, and were classed as young if under 21 and mature if 21 or over, making 4 groups altogether.

Do older students contribute more frequently in class discussion?

speaks	age	year
17	1	1
9	1	1
19	1	1
21	1	1
7	1	1
6	1	1
0	1	1
7	1	1
3	1	1
10	1	1
0	1	1
17	1	1
31	2	1
24	2	1
10	2	1
81	2	1
2	2	1
5	2	1
40	2	1
65	2	1
32	2	1
30	2	1
44	2	1
7	1	2
12	1	2
40	1	2
20	1	2
12	1	2
14	1	2
12	1	2
4	1	2
36	1	2
8	1	2
24	2	2
60	2	2
54	2	2
19	2	2
45	2	2
148	2	2
34	2	2
26	2	2
27	2	2
26	2	2
53	2	2

YOUNG	Y1	MATURE	Y1	YOUNG	Y2	MATURE	Y2
	17		31		7		24
	9		24		12		60
	19		10		40		54
	21		81		20		19
	7		2		12		45
	6		5		14		148
	0		40		12		34
	7		65		4		26
	3		32		36		27
	10		30		8		26
	0		44				53
	17						

The data are displayed here on the left in a slightly different format to the way we have formatted them for you to analyse in SPSS (on the right).

Look at the data. Can you understand it? Which group is which? How many students were in each group?

Does it show what you expected?

What does it tell you?

You may feel able to answer some of these questions, and less sure about others. Because the number of students is fairly small, it is possible to run your eye over the data and notice quite a lot. The techniques we are now going to use may help to clarify your ideas about this data, and would be even more useful with much larger data sets. Try to relate what we do now as much as possible to the feel you already have for these data.



Now open the data-file called **Studentss**. (Choose the **File** menu and select **Open, Data**. Click on the file you want to select and choose **Open**. (Data downloaded for the tasks will appear in a folder called "pd1qa" under the MyWork folder on drive F: (*The files may be on drive C: if you are working at home*))

The data should be displayed on your screen with a structure similar to the table on the right. These data are not paired so the structure above on the left is not suitable for analysis in SPSS (though it does fit better on paper). If you see four columns of figures on your screen you have opened the wrong file! These data represent independent groups of subjects, the first column is the number of times each person spoke, the second is whether they are young or mature (1=young, 2=mature) and the third column is their year of study.

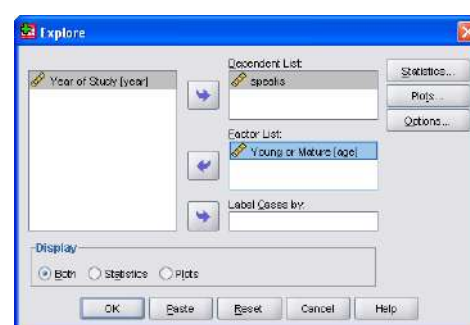


Look at the data on your screen, check that they are the same numbers as written here. We have structured them in a way that makes more sense to SPSS. This is a fundamental lesson to learn before typing data into SPSS. The SPSS system wants to see the data structured with each case (in this example each person) on a row. So in the file "studentss.sav" each row represents a student. The first column is how many times the student spoke, (called "speaks") and the other columns tell us what age group the student was in and what year of study.

To see what the numbers mean that we have used to represent the age category of the subjects click on the **View** menu then click **Value Labels**, do the same again to switch off the value labels. These **Value Labels** are set in variable view.

Exploring the data.

Click on the **Analyse** menu then **Descriptives** then **Explore**. The dialog looks alien at first because it uses some terminology that might be new to you. The dependant variable we are looking at is the number of times each student spoke, this is stored in the variable (column) called "speaks", transfer this one into the dependent list by selecting it and transferring it over with the little arrowhead button.



We want to know if age is a factor in the amount each person speaks so transfer the age variable to the factor list.

Click the OK button and the results should appear in the output window.

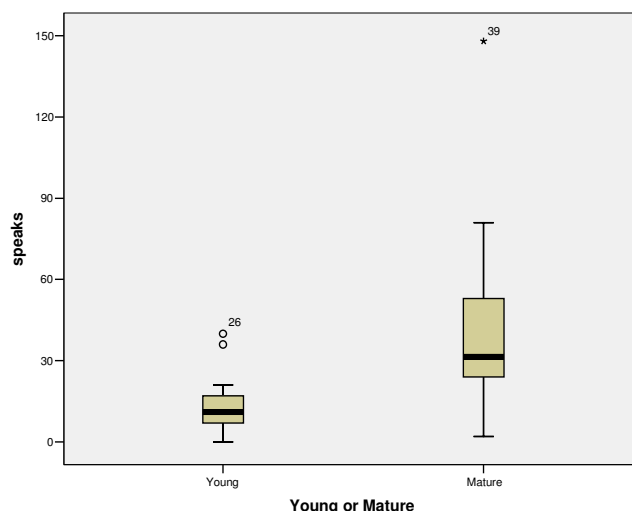
This command is a great way of getting a quick feel for the data, it does though generate a lot of output, some of which is of no use to us just now.

Descriptives					
		Young or Mature		Statistic	Std. Error
speaks	Young	Mean		12.77	2.166
		95% Confidence Interval for Mean	Lower Bound	8.27	
			Upper Bound	17.28	
		5% Trimmed Mean		11.99	
		Median		11.00	
		Variance		103.232	
		Std. Deviation		10.160	
		Minimum		0	
		Maximum		40	
		Range		40	
		Interquartile Range		11	
		Skewness		1.358	
		Kurtosis		2.020	
					.491
	Mature	Mean		40.00	6.603
		95% Confidence Interval for Mean	Lower Bound	26.27	
			Upper Bound	53.73	
		5% Trimmed Mean		36.43	
		Median		31.50	
		Variance		959.048	
		Std. Deviation		30.968	
		Minimum		2	
		Maximum		148	
		Range		146	
		Interquartile Range		29	
		Skewness		2.125	
		Kurtosis		6.469	
					.491
					.953

More on Drawing Boxplots

The Explore command is great for 2 or more groups of data, for paired data you can draw boxplots straight from the graph menu.

From the menus choose **Graphs** then **Boxplot...** (in the latest versions you'll need to use the legacy dialogs rather than interactive option on the menu to see it as it is here).





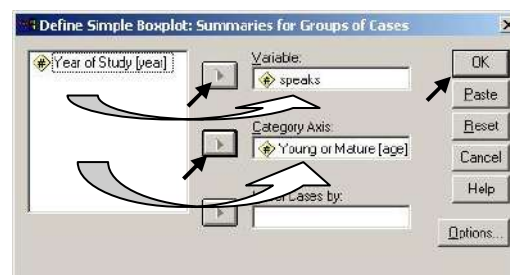
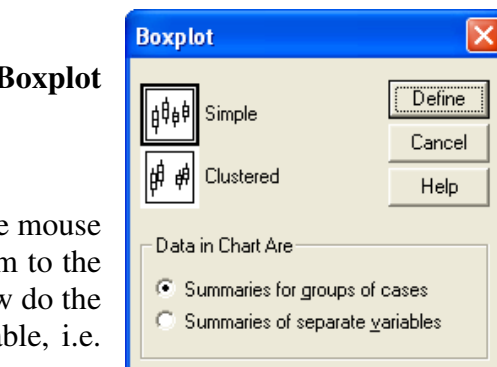
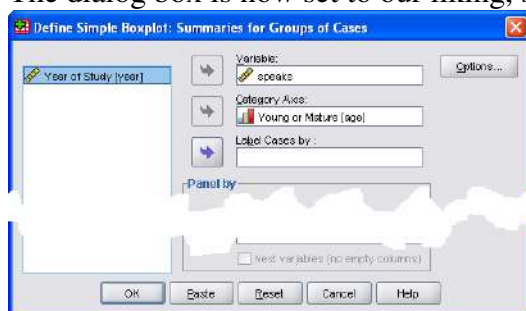
The Boxplot dialog tells us we are going to create a boxplot representing “**Summaries for groups of cases**” this is fine, our two groups are young and mature students. Press the **Define** button.

The next dialog box should be titled “**Define Simple Boxplot Summaries of Groups of Cases**”

We will use this dialog to pick the variables to plot.

To pick a variable select it from the left panel with the mouse and use the button between the panels to transfer them to the “Variable” panel. Do this for the *speaks* variable. Now do the same to tell SPSS which is our discriminatory variable, i.e. the one that tell us which age group each student is in.

The dialog box is now set to our liking, so click **OK**.



Remember you can switch between windows by clicking the buttons on the Task bar at the bottom of the screen.

Look at your boxplots. Can you see an asterisk or circle beyond the whiskers? In SPSS an asterisk represents an extreme outlier (a value more than 3 times the interquartile range from a quartile). A circle is used to mark other outliers with values between 1.5 and 3 box lengths from the upper or lower edge of the box. The box length is the interquartile range.

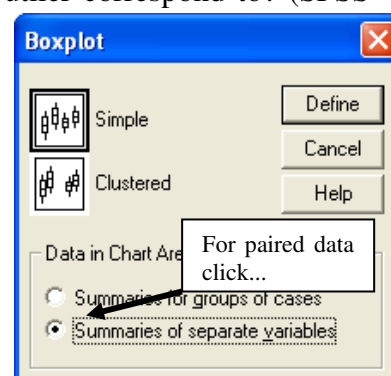
Which number on your data screen does the most extreme outlier correspond to? (SPSS gives a bit of a hint here!) Why is it an extreme outlier?

Look at the boxplots, which group has the highest median? What does this tell you about the groups?

Look at the boxplots, which group has the highest interquartile range (IQR)?

What does this tell you about the groups? Refer to the glossary for information about the interquartile range.

If the terms are unfamiliar to you read about the median and the interquartile range in the Glossary.



Summary: Boxplots are good for seeing the range and level of data and highlighting outliers. The box shows the IQR (Inter Quartile Range) and the bar in the box shows the median. Boxplots should be clearly labelled with the units of measure displayed.



Using Descriptive Statistics

From the boxplots it is hard to read the exact values of the median, quartiles, interquartile range and range. SPSS can calculate these easily.

Earlier we used the Explore command to calculate statistics for each group, young and mature. It is worth noting that if you have paired data you can put more than one variable into the dependant list and don't need to put any factors in, if you do this it works rather like the "Frequencies" method we looked at to begin with.

*If you want to do it again to recap; from the **Analyze** menu select **Descriptive Statistics** then **Explore**. The dependant list refers to the quantity we are measuring, in this case, the number of times people speak. In the factor list we put the factor that we are investigating, in this case "age".*

SPSS will calculate the stats for each group.

Mean (Young) _ _ _ _ _

Mean (Mature) _ _ _ _ _

From the output find the **Mean** and **Median** of each group. The mean and median are both forms of average, do they seem to agree?

Median (Young) _ _ _ _ _

Median (Mature) _ _ _ _ _

You can read (in the Glossary) that the **median** is a measure of central tendency. It gives us a kind of centre for each group, and allows us to say that students in one group 'on average' make more verbal contributions than students in another. The **interquartile range** is a measure of spread, on a boxplot it is the distance between the top and bottom of the box, and tells us something about how varied students in each group are. Look at the **mean** and the **standard deviation** (Std. Deviation also abbreviated to S.D. and sometimes s or σ) in your descriptive statistics. What do they tell you about the data? When you have had a good look, read about both in the Glossary.

The **standard deviation** is not the same as the **interquartile range**, but both are measures of spread or variation. When comparing datasets, the set with greater standard deviation will usually have the greater interquartile range.

You should get; Mean (Young)=12.77; Mean (Mature)= 40.00; Median (Young) =11.00; Median (Mature)=31.50

Another way of storing the data - a note to remember when putting in your own data.

The file we are looking at stores the data for each group using a discriminatory variable to tell us which group the case is from, this is preferred by SPSS, there is an alternative structure for this. To see this structure look at the file Students.sav The data it holds is just the same but is not as well suited to analysis by SPSS. It is important you pick the correct structure for your own data if you want to produce meaningful analyses.



More on Different types of data

To finish Task 2, read about nominal, ordinal, interval and ratio data in the Glossary. What kind are the data you have been studying in Task 2?

It is important that you understand the difference between data types, the type of data affects how it can be reasonably analysed.

For example the type of average we would use depends on the type of data, refer to the glossary to fill in the table below...

Example	Type of data	Mean, median or mode
We have the body weight of eight people and want to find an average, one person has a recorded weight considerably larger than any other, it could even be a typing error.		
We have 250 heights of female clients and want to give an idea of the average height.		
A researcher collects the type of housing that a sample of clients live in, single room, flat, terrace etc., what type of average can we use to talk about the typical type of housing for the sample of clients?		
Students are asked to score the taste of a new recipe of bun as like/dislike/don't know, what type of data have we collected and what average might you use?		

Answers; Body weight of 8 people including a possible outlier, these data are ratio but due to the small sample size and the possible error causing an outlier the median might be safer than the mean, usually the mean would be best for these ratio measurements. The 250 female heights are ratio data and the mean would be fine for these data. The housing type is at least categorical, however we might choose to rank the categories in order of size, e.g. flat, terrace, semi... and so on, this could pass for ordinal data with some caveats, if so the median might be employed, otherwise the mode is safest. Like/dislike/don't know, gives us three categories, even writing the categories as like/don't know/dislike, doesn't convince me that they represent ordinal data, it might be better to discount those who "don't know" and treat the remaining dichotomous variable. You could then analyse with percentages, e.g. "of those expressing a preference 73% preferred the new recipe." the percentage expressing how many expressed a preference could also be quoted.

The difference between Mean and Median

Open a new data file, we are going to type in a few figures. (from the menus choose **File, New, Data** – you will be prompted to save alterations to the last data you were editing.)

Put the following numbers in the first column;

7000, 7000, 7000, 7000, 7000, 7000, 7000, 7000, 7000, 100000.

Give the column the title 'Salaries' (you need to click onto the Variable View for this – notice that SPSS ignores the capital letter in a variable name). Back in Data View you may want to alter the column width by dragging the vertical bar next to the variable name.

salaries	
salaries	var
1	7000.00
2	7000.00
3	7000.00



The numbers represent the annual salaries of the 10 permanent employees of a small (mythical) private clinic. Which is the director's?

Run Descriptive Statistics to find the mean and the median. If you were the union negotiator for the employees of the clinic which of the two average salaries would you quote to the press? If you were the owner of the clinic which might you quote?

Find the inter-quartile range and the standard deviation. Can you sketch what the Boxplot would look like? Create the Boxplot on SPSS if you like.

Summary: Mean vs. Median - both are types of average. The mean is based on all the data values, however because of this it is prone to being unduly affected by outliers in the data, most noticeably when the sample is small. The median however is largely unaffected by one or two extreme outliers, even in small samples, it is simply the middle value.

An example: The table below is from the UK adoption statistics for the year 2003. (<http://www.dfes.gov.uk/rsgateway/DB/SBU/b000425/index.shtml>) Although we don't have the original data with the individual ages of the children, this has presumably been used to create the average, unfortunately we are not made aware whether the average used was the mean or median.

Age at adoption	2003
Under 1	240
1 to 4	2,100
5 to 9	1,000
10 to 15	180
16 and over	10
Average age	4 years 3 months

Have a look at the available summary of the data in the table, which type of average would the relatively small number of older children have the greatest effect on? Think about the effect of two extra children on the mean or median, one child under 1 and one age 16, how would they affect each type of average? What type of average do you think would be best for this type of data?

Thoughts on this example: The problem with using the mean on the data for this application is that a relatively small number of older children will increase the mean disproportionately.

If we want to convey a general figure for the age of adoption it might be better to either say a more general statement like "well over half the children adopted in 2003 were between the ages of one and four" this succinctly paints a picture of the figures, alternatively we could use the median rather than the mean, this would combat the tendency for the small number of much older children to skew the average higher.



Task 3 Standard Deviation (S.D.) what is it?

What is the Standard Deviation (S.D.) really measuring? What can it tell us about our data?

Name	German	Geography	IT
Fred	27	42	39
John	22	26	34
Mary	54	32	31
Alan	49	34	29
Joan	67	32	32
Peter	16	31	11
Mavis	46	34	29
Sarah	20	31	31
Adam	21	41	67
Daniel	11	30	30

The table above shows the German, Geography and IT results of a group of ten students. Use SPSS to help you fill in the shaded area below on these notes, i.e. the mean, maximum and minimum for each subject. (The data is stored in a file called “*std dev example.sav*” *If you can't remember how to open files re-read the instructions on page 8.*)

MEAN			
MAX			
MIN			

HINT:

- From the **Analyze** menu select **Descriptive-Statistics** then **Frequencies**.
- Select all the variables (get them from the left into the right pane).
- Click the **Statistics** button and select the options for mean, maximum & minimum, then click **Continue**.
- Uncheck the option to display frequency tables. Click **OK**.

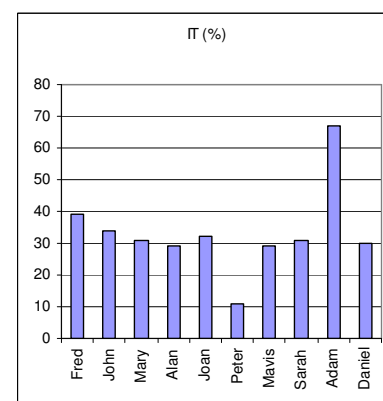
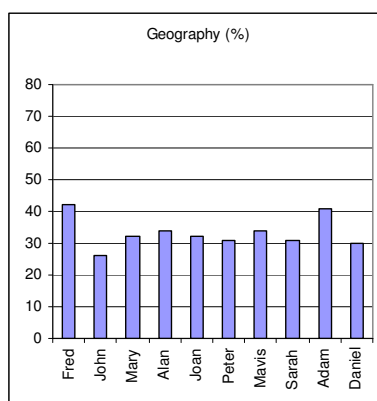
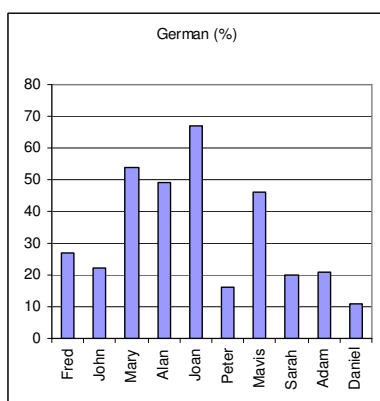
Looking at the figures for mean, maximum and minimum, what can you say about the three sets of figures?

Answer the questions below.

- 1 Which set(s) of figures has the largest range?
- 2 Which set(s) of figures has the largest number in it?
- 3 Which set(s) of figures contains the smallest number?
- 4 Which set of figures has the largest minimum?

Given the figures for mean, maximum and minimum it is hard to differentiate between the German and IT figures, the mean, (arithmetic mean) of the figures is the numbers all added together then divided by the number of numbers. However it gives no indication of the distribution of the marks within the sets of figures.

To do this we could graph the three sets of figures and see if that helps us (later we will create bar charts, for now just look at these).



Look at the three graphs above. Which two do you think are most similar?

I think the Geography and IT graphs but it's rather subjective. They do seem to have less variation in the values than the German results.

Question: How can we assess this in a fair, unambiguous way, to find out which of the three has the least widely deviating set of numbers?

Answer: Use the **Standard Deviation**.

The standard deviation of a set of numbers is a measure of how widely values are dispersed from the mean value. You can work the standard deviation (S.D.) out for a set of numbers manually if you are so inclined in a similar fashion to working a mean out; it just takes longer because the formula is much more complex! So let SPSS do it.

To work out the standard deviation of the numbers in each column use **Descriptive Statistics** then **Frequencies** from the **Analyze** menu.

Using the Frequencies option rather than Descriptives gives us a larger range of statistics available.

Select the three variables (get German, Geography and Information Technology (IT) from the left into the right pane).

Click the **"Statistics"** button and select the Standard deviation as well as mean, maximum and minimum, then click **"Continue"**.

Before pressing OK on the Frequencies dialog box, uncheck the option to display frequency tables then click **OK**

Because the figures for each subject are in the same units we can compare the standard deviations and see how widely dispersed the values are.

	German	Geography	IT
S.D.			

The values I got for the data are below. Look at the graphs and the S.D. values to decide if high S.D. values indicate large or small deviations in the data. High S.D. values indicate a greater spread of values.



To show this, create a new variable by copying the number 33.3 down ten cells. The total should be 333, the mean, median, maximum and minimum should all be 33.3, what is the standard deviation? (Have a guess before you calculate it.)

Now you've worked out the values for the standard deviation answer the following questions. The values I got are; German, 19.044, Geography, 4.877, IT, 13.849

1. Which set of figures, German, Geography or IT, is the least spread out?
2. Of the two subjects with the same mean, and the same range, which varies least?
3. Which of the three sets of figures, German, Geography or IT varies most?

I think the answers are:

Geography is the least spread out.
Of the two subjects with the same mean, and the same range, IT varies least.
German varies the most.

A real data example to look at: **Comparison of Visual Estimations and Mean Goniometric Measurements of wrist flexion and wrist extension.**

Load the file **goniometry.sav** the file contains estimates and measurements of wrist movement. The angle measurements were taken using a goniometer. Use SPSS to calculate the Mean, Median, Standard Deviation and Range for the estimated and measured flexion. (i.e. Flexion Estimate & Flexion Measurement)

Look at the figures you have calculated and decide...

1	Which column of flexion results appears most varied?	estimated	☐
		measured	☐
2	Was the tendency to underestimate or overestimate the flexion?	underestimate	☐
		overestimate	☐
3	On a Boxplot of these data, Which set of flexion results would you expect to have the biggest box?	estimated	☐
		measured	☐

Statistics

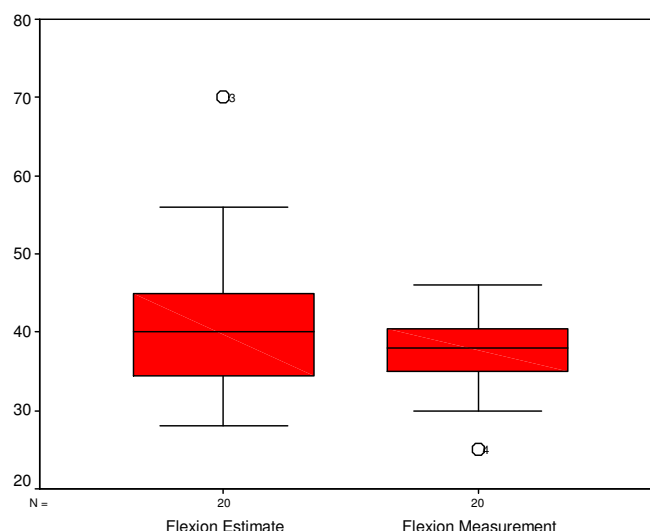
1. Which column of flexion results (estimated or measured) appears most varied? The adjacent results would lead me to say that the estimate is more variable. (SD is greater.)

		Flexion Estimate	Flexion Measurement
N	Valid	20	20
	Missing	0	0
Mean		42.25	37.40
Median		40.00	38.00
Std. Deviation		10.172	4.999
Range		42	21

2. Was the tendency to underestimate or overestimate the flexion? The above results show a slight over estimation, but it is quite a small difference and may be due to chance.



The Boxplot for the two variables allows a visual comparison of the level and spread. To get this boxplot you need to remember that these data are "Summaries of separate variables" rather than "Summaries for groups of cases". It looks like the estimate has the larger IQR to me! (The bigger box is the bigger Inter Quartile Range.)



Summary: Range, IQR & SD are all measures of spread. Only the SD takes all the data values into account, however this leaves it open to problems similar to the mean, i.e. a tendency to be swayed inordinately by extreme values. The range is extremely sensitive to outliers, since it is based only on the smallest and largest values. The Inter Quartile Range is again based on only two values, the upper and lower quartiles, these are on each end of the middle half of the data, therefore less effected by extremes.

A Simple example: A researcher is investigating the height of adult females living in two towns. She believes that the women from Youngville are, on average taller than those who live in Oldton., If the mean heights and Standard deviations were as follows;

Town	mean	Standard deviation	
Youngville	175cm	5.25	☐
Oldton	169cm	15.50	☐

Which sample varies most?

Thoughts on this example: The sample from Oldton seems more varied - it does perhaps lead us to think there are some differences in the samples other than the people in one town being taller.



Task 4 Histograms and the Normal Distribution

Using Histograms to look at the distribution of data.

We have already seen that two sets of figures may have the same mean but the data may be spread around the mean more widely in some populations than others.

Boxplots provide a simple graphical representation of how the values are distributed in the data. The Standard Deviation gives a numerical value to the level of spread.

A Histogram can give a picture of the data! It is a very powerful tool when used appropriately; it can let us see the distribution of the data. It does though need quite a large amount of data to give a nice bell shaped graph.

In this task we will use histograms to look at the shape of distributions, you might though want to apply this technique in other situations.

Heights of adult males. (Source: *Final Report of the Anthropometric Committee to the British Association (1883)*, p. 256.)

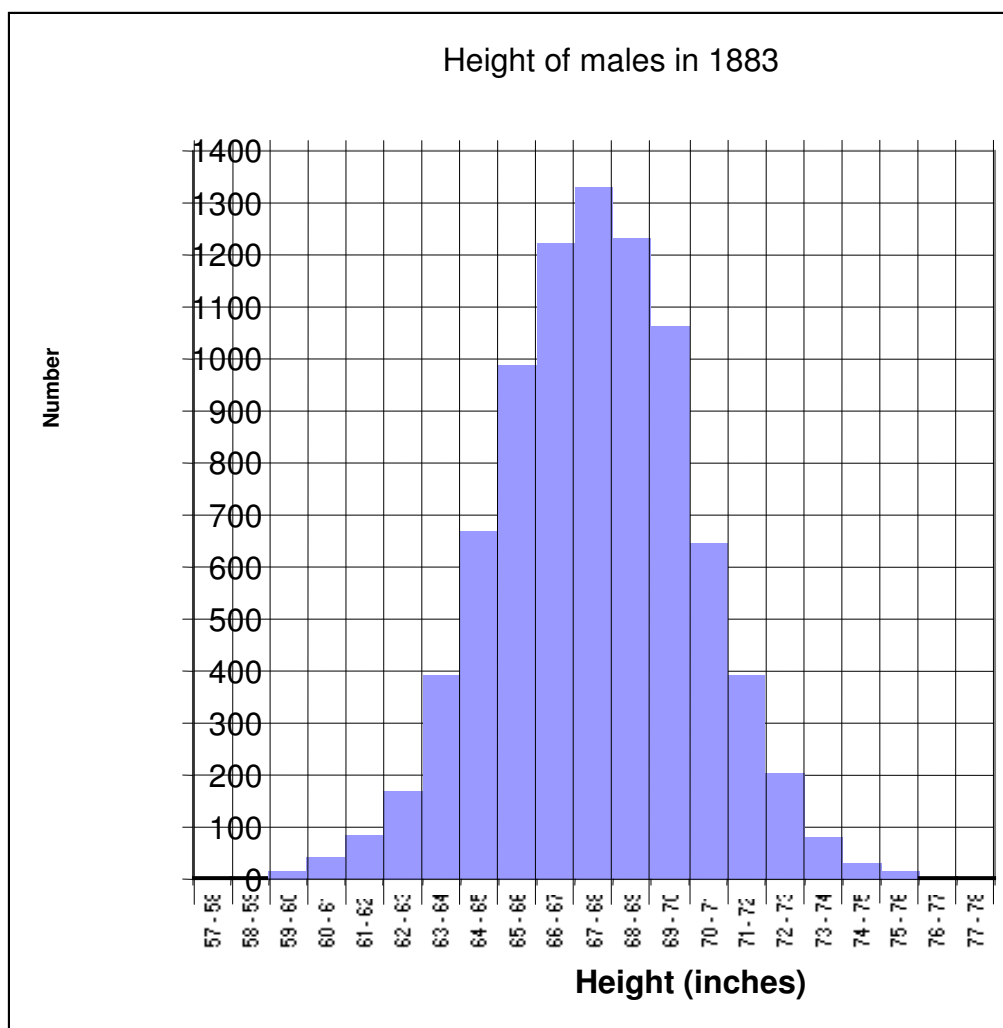
Height	Number
57 - 58	2
58 - 59	4
59 - 60	14
60 - 61	41
61 - 62	83
62 - 63	169
63 - 64	394
64 - 65	669
65 - 66	990
66 - 67	1223
67 - 68	1329
68 - 69	1230
69 - 70	1063
70 - 71	646
71 - 72	392
72 - 73	202
73 - 74	79
74 - 75	32
75 - 76	16
76 - 77	5
77 - 78	2

The data in the table gives heights of adult males in 1883. It represents the heights of 8585 adult males; the data is gathered in inches - this doesn't cause us any great problem since for this exercise we are concentrating on the shape of the distribution of heights. (If you really need to know, 1 inch = 2.54cm approximately)

The table is drawn from the heights of 8585 males. Rather than have a table with all 8585 heights it is summarised by giving the number of individuals in each height range, e.g. there were two people in the lowest range, covering people from 57 inches up to 58 inches. It isn't too clear from the table but we can assume that anyone who was exactly 58 inches tall would be in the 58-59 category.

Below the table is reprinted horizontally; on the next page is a histogram of the data it give a pretty good example of the bell-shaped Normal distribution.

Height	Number
57 - 58	2
58 - 59	4
59 - 60	14
60 - 61	41
61 - 62	83
62 - 63	169
63 - 64	394
64 - 65	669
65 - 66	990
66 - 67	1223
67 - 68	1329
68 - 69	1230
69 - 70	1063
70 - 71	646
71 - 72	392
72 - 73	202
73 - 74	79
74 - 75	32
75 - 76	16
76 - 77	5
77 - 78	2



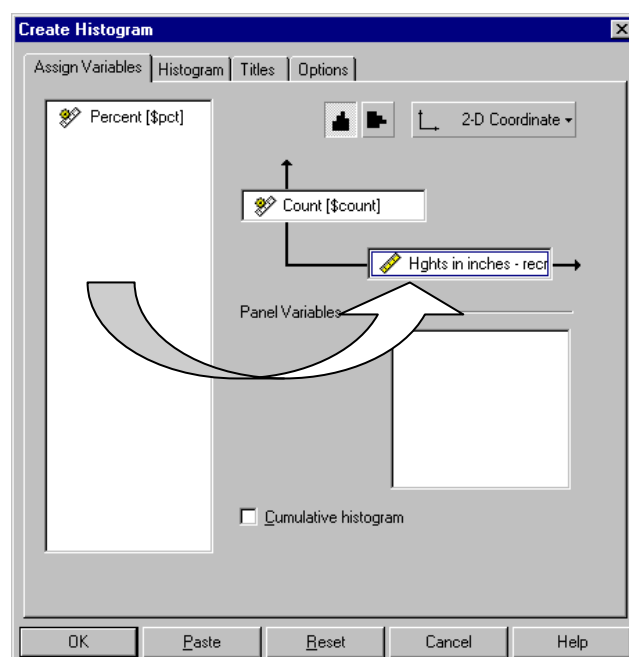
This manual version shows the typical bell shaped normal distribution. This distribution is sometimes referred to as a Gaussian distribution, for our purposes the two are similar enough.

Drawing the same graph in SPSS.

Load the file called **Reconstructed male heights 1883.sav** This file contains data that is similar to that from which the table you have seen was derived. The file contains 8585 heights, measured in inches.

We are going to create a histogram from the values in the variable called **hgtrein**

From the menus choose **Graph, (Legacy,) Interactive, Histogram.**





It is wise to press the **Reset** button in the Create Histogram dialog, to prevent the scales from previous data being used.

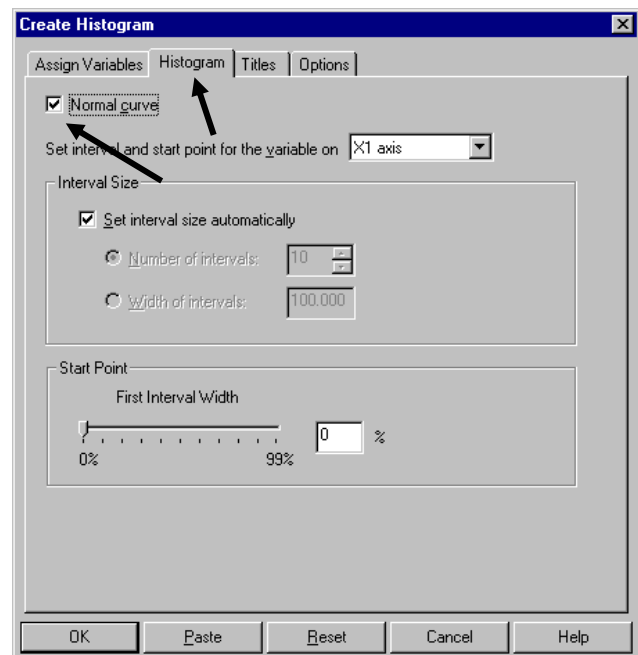


Drag the **hgtrein** (Heights in inches - reconstructed) variable over to the box representing the horizontal axis of the graph.

Click **OK** and wait to see the graph in the output viewer.

You should see a normal (bell shaped) pattern to the distribution of the data. This is typical in many natural distributions. The majority of subjects are clustered round the mean and the numbers of individuals in the categories more distant from the mean is far less, in this example there are less very tall or very short males.

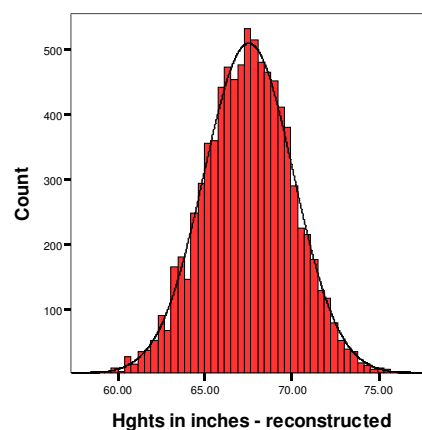
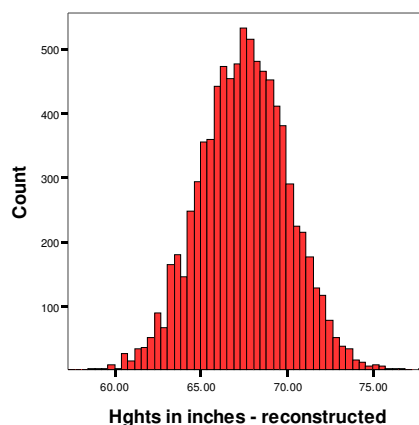
To see a normal curve superimposed on the graph go back to the Create Histogram dialog box (from the menus **Graph, (Legacy,) Interactive, Histogram**) then click on the **Histogram** tab and tick the "Normal curve" check box, then Click **OK**.



Are these data Discrete or Continuous? Read about Continuous and Discrete data in the glossary to help answer this.

The graphs below show the output you should see if you follow the instructions. The first two are the histograms without and then with the normal curve superimposed.

By the way the data are continuous do check in the glossary if you don't know why.





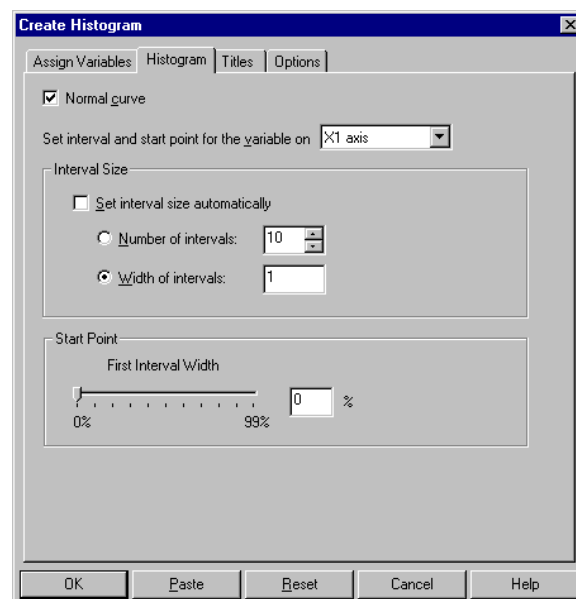
A tweak for the more confident to try...

On the Histogram tab of the Create Histogram dialog, switch off the automatic size interval setting and change the interval width from 100.000 to 1.

Compare 19th to 20th century heights.

The file **malehgt1990s.sav** contains heights for males of a more contemporary nature. There is less data so the bell shape may not be as smooth.

Follow the instructions again for creating a histogram using this data, work in mm or inches. If you choose inches then you can compare the histograms easily to the ones done earlier. Are people getting bigger?



The normal distribution is important, not just because it gives a pretty curve but also because many inferential tests assume normality in the data distribution.

Which of the following examples would you expect to be normally distributed?

Normally distributed?	Yes	No
Ages of people in a town.		
Heights of 20 year old men.		
Weights of one-year-old squirrels.		
The price of drinks in a bar.		
The life (in hours switched on) of light bulbs.		

	✓	✓	The life (in total hours switched on) of light bulbs.
✓	✓	✓	The price of drinks in a bar.
	✓	✓	Weights of one-year-old squirrels.
	✓	✓	Heights of 20 year old men.
✓	✓	✓	Ages of people in a town.
No	Yes		Normally distributed?

Another example of data with a discriminatory variable in:

This example reinforces the idea of a discriminatory variable. The file **Radiologist dose with and without lead combined.sav** contains data gathered to assess the effect of a lead screen to reduce the radiation dose to Radiologists' hands while carrying out procedures on patients being irradiated.

In the trials the lead screen was placed between the patient and the radiologist, the intended effect was to reduce the radiation dose to the radiologist, however there were fears that working through the screen would lengthen the procedure. We want to answer two questions with this data, one about the hand dose and the other about the length of time the examination took.



Look at the data, the variable called "screen" is the variable that lets you discriminate between procedures carried out with or without the lead screen. If there is a 1 in the screen variable column it means the procedure was carried out with the screen in place, if not the value is 0.

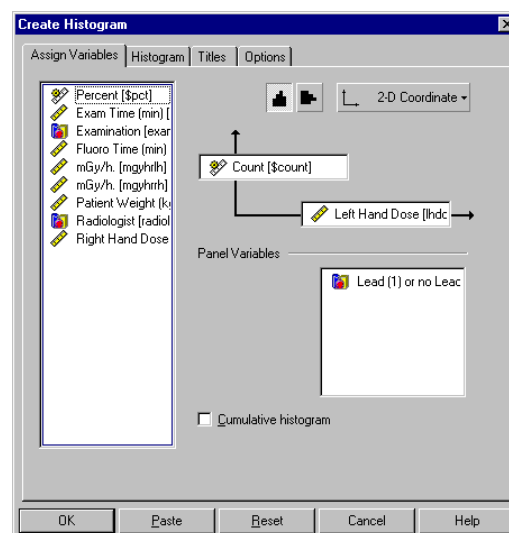
We can use this discriminatory variable to create two histograms at once, by using it as a panel variable.

The variable we are interested in is the dose to the radiologists' left hand, the left-hand would be nearest the patient so we will concentrate on the left-hand dose variable.

Draw an interactive histogram using the left-hand dose variable (**lhdose**) and the discriminatory variable (**screen**) as the panel variable.

What do the histograms show us about the data?

If you have time draw a similar histogram using the **extimmin** variable. Does this back up the fears about the increase in examination time?

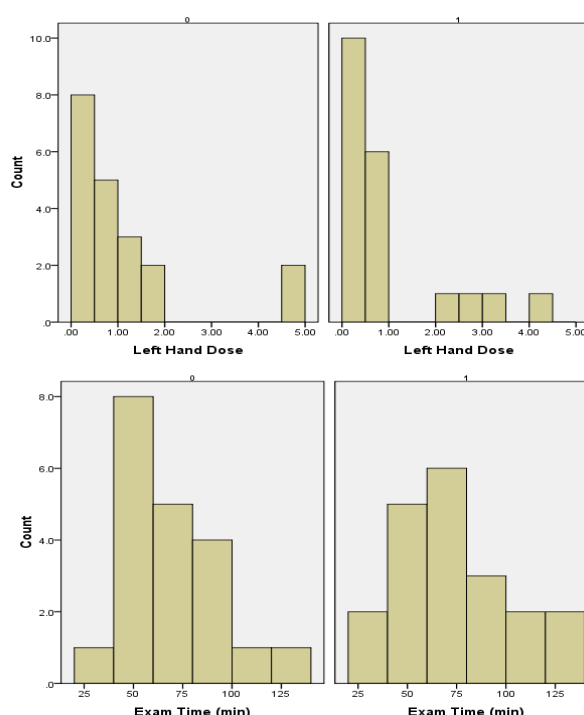


Summary: Histograms are for displaying continuous data, e.g. height, age etc, the bars touch, signifying the continuous nature of the data. The area of the bars represent the number in each range, the bars are usually of equal widths but this need not always be the case. Histograms should be clearly labelled and the units of measure displayed. The use of Histograms compared to Bar Charts is summarised after the section on Bar Charts.

The small sample size makes it difficult to draw conclusions, however it would appear that the screen has increased the number of radiologists receiving a lower left hand dose.

The examination time also appears to be altered, more examinations appear to be taking longer.

If you want to examine the data more it is worth looking at boxplots. Notice also on these graphs that the shield/no shield variable is left as 0 or 1 rather than labelled - it certainly doesn't help the readability of the output!





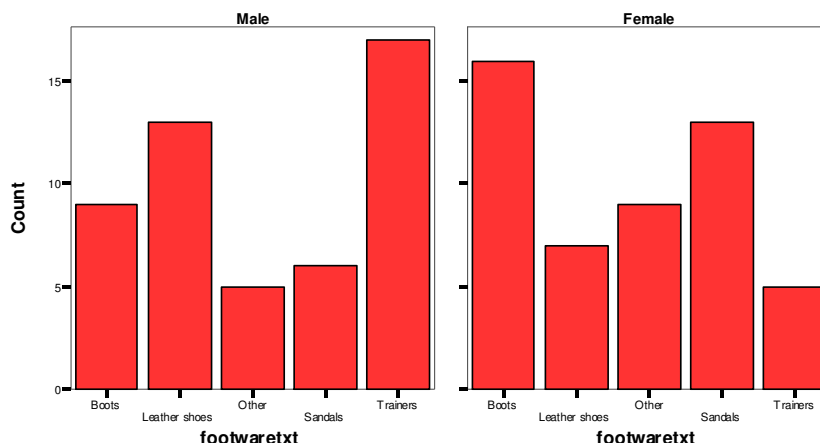
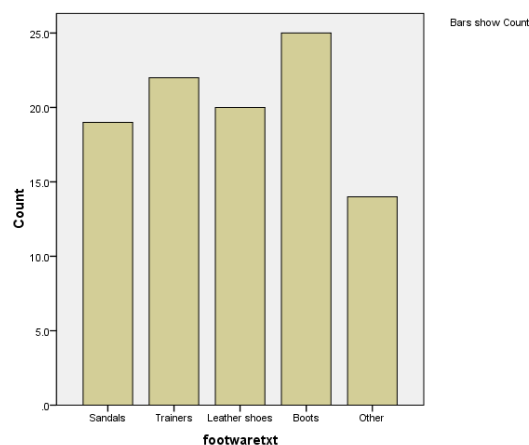
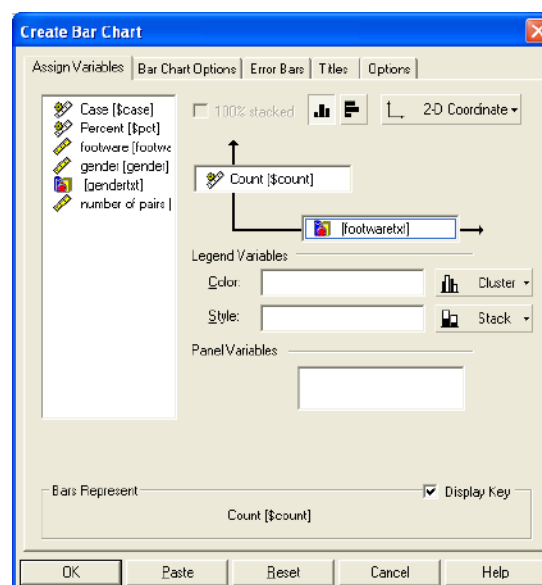
Task 5 Bar charts.

Bar charts and histograms look similar at first; there is however a definite difference in the type of data each is designed to show and this subtle difference is an important one if you are using them in your research. Bar charts are for non-continuous data, i.e. data in categories that are not related in any order. Histograms are for displaying continuous data.

To have a go at creating a bar chart open the file called *shoetypes.sav* this file contains data about the type of shoes worn at the time the data were gathered and number of pairs owned by a sample of 100 people. We can use SPSS to analyse the data by using bar charts among other methods.

Drag the "footwaretxt" variable to the horizontal axis then click OK. The graph above should appear. Try again but this time drag the "gender" variable over to the **Panel Variable** box and see what happens. (notice that if you use the numerical version of this variable you might get a "Convert?" dialog box, just say yes to this to continue. You might notice a different profile between the shoe portfolios across the genders. I must confess here that these data are purely fictitious; I have it on good authority that I've seriously underestimated the number of shoes for one of the genders!

It is worth noticing that the graph can be edited after it is drawn, just double click on the graph and then click into the labels you wish to alter. An alternative to "Counting" the numbers for the bar heights is to use percentages, this is done by dragging the "Percent" variable over to the vertical axis.





Summary: Bar charts are for non-continuous data e.g. the number of people from each of five towns, the bars do not touch. Bar charts should be clearly labelled and the units of measure displayed. Bar charts and Histograms look similar, however the type of data they should be used on is different. In a Histogram the bars touch each other, this denotes the continuous nature of the data being displayed. Bar charts should be used for discrete data. If you aren't sure about the difference between continuous and discrete data look it up in the Glossary.

Test yourself; Of the following which would best be displayed in a Histogram or Bar chart. Fill in the table below, put H for Histogram or B for Bar chart in the end column.

	H or B
1 The number of students in the age groups 18-27, 28-37,38-47 etc.	
2 The number of people living in each of three towns.	
3 The number of patients visiting an Optician with short sight, long sight and no sight defect.	
4 The marks of each individual student in a class.	
5 The number of students in each range of marks in 10% intervals.	
6 The number of men vs. women in a town.	

H or B	The number of students in the age groups 18-27, 28-37,38-47 etc.
H	This is continuous data, people can have any age in a continuous range - hence use a histogram.
B	The number of people living in each of three towns.
B	This is not continuous data, it is discrete - use a bar chart.
B	The number of patients visiting an Optician with short sight, long sight and no sight defect.
B	This is discrete - the data is giving the number of patients in each of three categories. Use a bar chart.
B	The marks of each individual student in a class.
B	The number of students in each range of marks in 10% intervals.
H	The number of men vs. women in a town.
B	continuous data - you could use a bar chart or in this case a pie chart may also be an option.



Percentages.

Lets do a simple example just to check the basic principal, sometimes it's a good idea to work through the principals on a simple example.

The table shows the spending money of my three children. To find out the percentage of the total spending money each individual child receives we must first work out the total amount.

Name	Spending Money per month	Percentage of total Spending Money
Tom	8.00	
Rachel	7.00	
Jodi	5.00	

To do this, simply add up all the money in the middle column.

$$8 + 7 + 5 = 20 \quad (\text{this tells us the total amount of money})$$

We could say that Jodi gets five twentieths of the total money. In figures this is $5/20$ or $\frac{5}{20}$

Tom gets eight twentieths of the total money. In figures this is $8/20$ or $\frac{8}{20}$

Rachel gets seven twentieths of the total money. In figures this is $7/20$ or $\frac{7}{20}$

We'll work on Rachel's money for the next bit...

If we want to convert this to a percentage we just multiply it by one hundred. A percentage means "*per hundred*" (cent means 100 – 100 cents make a dollar, 100 degrees on the Centigrade scale, 100 legs on a ... you get the idea!) so multiplying our fraction by 100 gives the fraction of 100.

We are really saying, "*Rachel gets seven twentieths of a hundred*". To work it out, first work out one twentieth, which would be 5 or 5% (since $5 \times 20 = 100$ we can deduce that one twentieth of 100 is 5). So each twentieth is 5%, Rachel gets seven twentieths of the total amount so that is 7×5 percent since each twentieth is worth 5% i.e. 35 percent.

The sum we have done could also be written as:

$$100 \times 7 \div 20$$

On a computer we would type $100 * 7 / 20$ because the multiply and divide symbols are not on the keyboard.

General rule for percentages of a total:

$$100 \times \text{the individual value} \div \text{the total of the values}$$

Have a go at filling in the "Percentage of total Spending Money" column in the table above. Check they add up to 100.

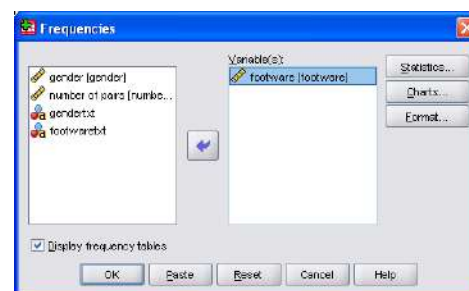
Summary: Percentages show proportions, it should be clear what they are percentages of.



Using SPSS to calculate the percentage of subjects in each group.

You can very quickly create summary percentages using the "frequencies" command, for example in the shoes file, what percentage of subjects were wearing each type of shoe? Clear any previous setting by clicking the "Reset" button then scoot the footwear variable into the variables box and just hit OK.

The valid percent column is the one to read, it will ignore any empty cells.

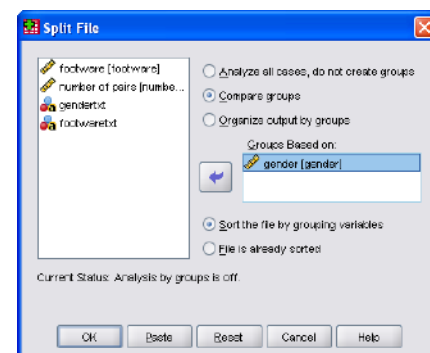


Does the percentage of footwear types differ in the different gender grouped; the bar charts seemed to imply this...

		footwear			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Boots	25	25.0	25.0	25.0
	Leather shoes	20	20.0	20.0	45.0
	Sandals	19	19.0	19.0	64.0
	Trainers	22	22.0	22.0	86.0
	Other	14	14.0	14.0	100.0
	Total	100	100.0	100.0	

Lets get SPSS to do everything twice, once for males and once for females, we can do this using the split file command. Choose **Data, Split file**.

Now calculate the percentages again as you did before.

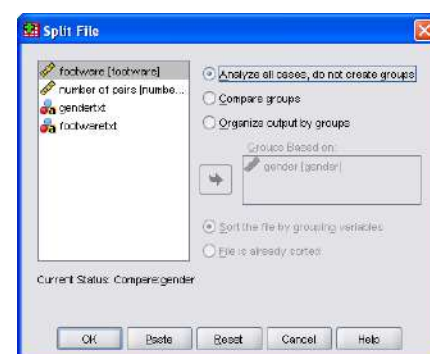


The output should now be split into two groups, one for Male and one for Female. Tables like this are rarely in the ideal format for inclusion in a dissertation or paper but can be copied and pasted into a word processor and manipulated there.

		footwear			
gender		Frequency	Percent	Valid Percent	Cumulative Percent
Male	Valid Boots	9	18.0	18.0	18.0
	Leather shoes	13	26.0	26.0	44.0
	Sandals	6	12.0	12.0	56.0
	Trainers	17	34.0	34.0	90.0
	Other	5	10.0	10.0	100.0
	Total	50	100.0	100.0	
Female	Valid Boots	16	32.0	32.0	32.0
	Leather shoes	7	14.0	14.0	46.0
	Sandals	13	26.0	26.0	72.0
	Trainers	5	10.0	10.0	82.0
	Other	9	18.0	18.0	100.0
	Total	50	100.0	100.0	

Remove the split once you have done with it. If you leave it on you may get some strange results. Choose **Data, Split file**. Then select the "Analyze all cases" option, then click **OK**.

Don't forget to switch this feature off when you don't need it!





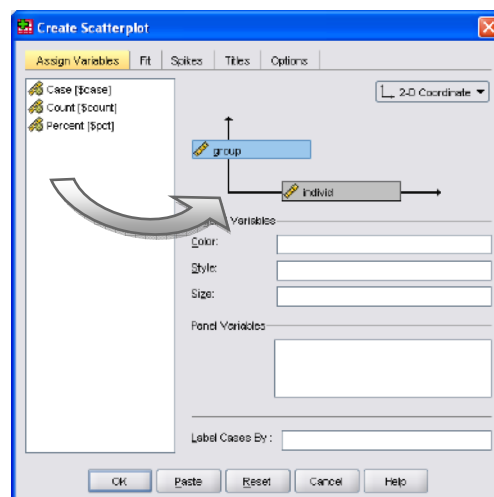
Task 6 Using Scatterplots to look for correlation

Scatterplots are used when data are paired: each point on a diagram represents a pair of numbers. Scatterplots need paired data.

- Open the data file called **Step**.

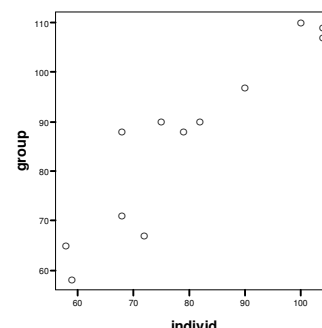
These data come from an experiment to see whether subjects could perform more step exercises in a fixed time in a group or on their own. A physiotherapy student collected them as part of a third year project.

Look at the data, you will see that the columns are of equal length, this is another indication that the data are paired. If we had the names of the twelve people who were the subject of the study we could put them in a third column, again with just twelve entries. Each row would then be one person's data, their name, the number of steps done when in a group and the numbers of steps done working alone. Sometimes you will see paired data where not all the columns have the same number of entries, this could have been so in this example if one of the subjects had failed to turn up for the group exercise.



We are going to draw a scatterplot for these two columns with the number of steps done individually on the x-axis.

To draw the scatterplot we will use the interactive graph system. Click on the SPSS **Graphs** menu then choose, (Legacy), **Interactive, Scatterplot**. Drag the “individual” variable to the horizontal axis and the “group” variable to the vertical axis. Click the **OK** button and your graph should eventually appear in the SPSS viewer.



Read about correlation in the Glossary and say what kind of correlation is involved here. The questions below may help.

Do the points appear to form a line? _____.

If they do is it a clear, quite thin line or more like a cloud? _____.

Does it slope up or down from left to right? _____.

Look at your answers and decide if there is a strong, weak or no correlation. Is it positive or negative? _____.

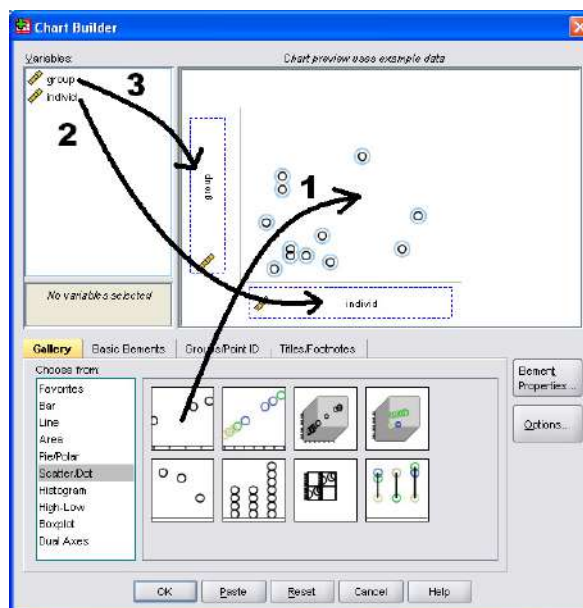
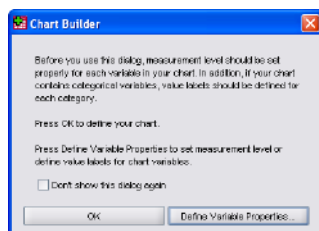
This shows fairly strong positive correlation.



The new "Chart Builder" method.

You might like to have a play with the latest way of creating charts, to recreate the scatter plot using the new "Chart Builder" feature under the Graphs menu. The intermediate advice about "measurement level" is important, but in this example no action is needed. Variables for this method have to be set at the correct measurement level for the type of graph you plan to use.

Drag the objects on the dialog in the order numbered on the illustration here.



Summary: Scatter plots are used to show paired data, where for example one person is tested under two circumstances, each individual will have a pair of readings. In this example a scatter plot can be used to indicate changes between the performance in different circumstances. Scatter plots are also typically used to show correlation. Scatter plots should be clearly labelled and the units of measure displayed.

An important note; Correlation does NOT show causality!

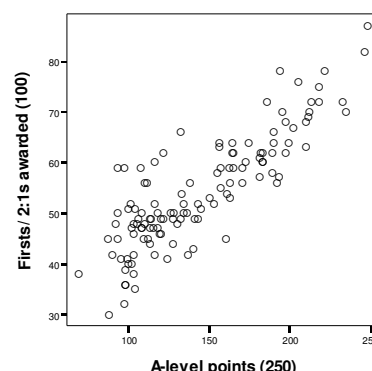
Example: The graph shows the proportion of Firsts or Upper Seconds as a measure of degree attainment, plotted against standards of A level passes obtained by new students, the data is from the Sunday times survey of HE. It is sighted as evidence that universities with an intake of students with "better" A-levels have an output of students with a higher percentage of "better" classed degrees.

Is this an appropriate way to show the data?

Is the graph labelled adequately?

What does it show?

Does this support the above argument?



Is this an appropriate way to show this data? Yes, this is paired data, one dot represents the data from one University. Is the graph labelled adequately? Not bad, but I would have liked an overall title and some indication about how the A-level scores are derived (is big = good on this scale?). What does it show? It shows that establishments with higher average A-level attainment students at intake tend to award a higher level of degree. Does this support the above argument? It appears to support the theory that universities with an intake of students with "better" A-levels have an output of students with a higher percentage of "better" classed degrees. However it is only an overall picture, it doesn't preclude the possibility that the worst A-level student could end up with the highest degree classification! It is looking at universities not students.



Task 8 Line graphs.

Load the file called “**Oxygen used walking**”

The data is just part of a large dataset collected by a student researching the effect of tibial malunion on oxygen expenditure during exercise.

For our purposes the data gives us a good example of a variable changing over time. The file contains the data from just one subject.

The subjects of the research, performed exercise (walking at a self selected speed) while their heart rate and oxygen consumption were monitored using an instrument to measure the oxygen uptake of individuals. The equipment used was the Cosmed K4. If you want more details on this instrument the company have a web site at:

<http://www.cosmed.it>

The variables in the file are:

vo2	Volume of O2 ml/min
vco2	Volume of CO2 ml/min
hr	Heart Rate beats per minute
seconds	time in seconds from start of procedure



The protocol employed to take the measurements consisted of:

- 5 minutes rest, to achieve baseline values for heart rate etc and enable the subject to get used to the equipment, followed by:
- 10 minutes exercise, (walking at a self selected speed) followed by:
- a second 5 minutes rest, to ensure baseline values return to the norm for the subject.

This is important when interpreting the graph we are about to draw.

Creating the line graph.

From the menus choose **Graphs, Interactive, Line**.

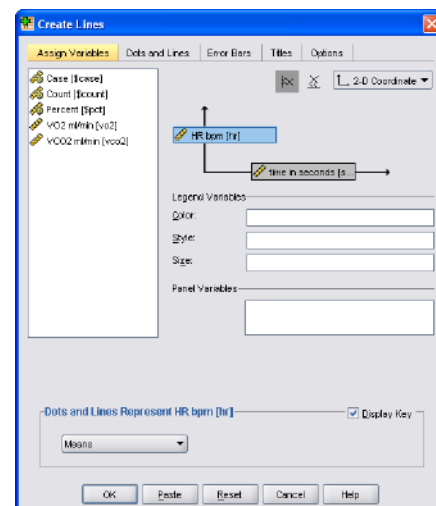
Drag the Heart Rate onto the Y-axis (the one going up)

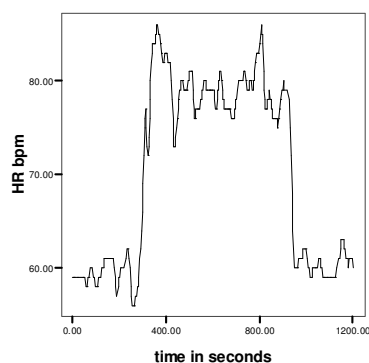
Drag the Time onto the X-axis (the one going across)

Press **OK**.

Look at the graph.

It is easy to see when the subject started and stopped walking!





However it looks as if there was a massive increase in heart rate on taking exercise unless you look at the figures. The graph is using a false origin. This magnifies the effect of differences in the data. I believe it is good practice to always draw a false origin to the attention of the reader, this doesn't always happen though, especially in areas like advertising and politics.

We will redraw the graph with no false origin.

You can select a previously used dialog box by pressing the Dialog Recall button.



This time switch off the “Display Key” option. ☒ → ☐

Click on the “Titles” tab and add a suitable title for the graph, e.g. “The effect of exercise on heart rate.” I feel it is good practice to put three basic pieces of information a graph, or any output for that matter:

-  A descriptive title, saying roughly what the graph is about (the axis labels should give detail such as units usually).
-  The date the data was current (especially important with ages or annual statistics).
-  The name or reference to the author or organisation responsible.

When your labels are as you want them click the **Options** tab.

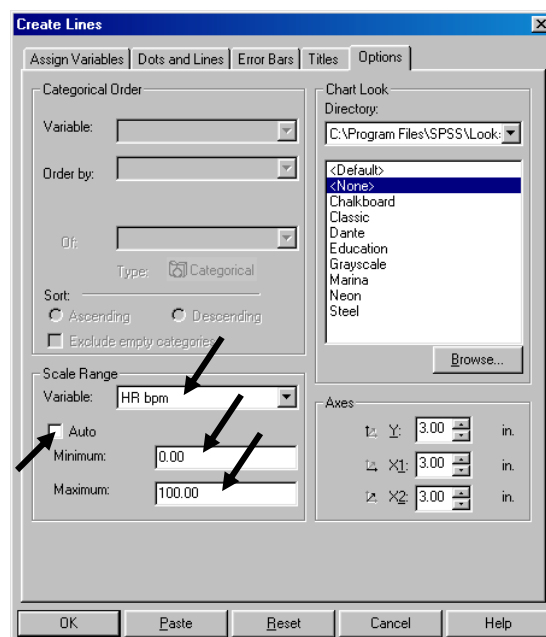
Change the “Scale Range” variable to **HR bpm**.

Switch off the “Auto” feature.

Set the minimum to 0 and the maximum to 100.

Press **OK**.

Sit back and admire your work!



Summary: Line graphs are ideal for showing the changes in a variable as another alters, e.g. changes over time. The independent variable goes on the x-axis and the dependant variable goes up the y-axis. More than one line is often shown on the chart allowing comparisons. Line graphs should be clearly labelled and the units of measure displayed.



Multiple line graphs.

For this exercise we will use the older graphing system and the data in the file called "*Children looked after.sav*". The variable names may look a bit strange at first, but if you move the mouse pointer over the top of the column a "tool-tip" should give the longer name. The data are from the Department for Education and Skills (<http://www.dfes.gov.uk/rsgateway/DB/VOL/v000454/index>) and give figures for children looked after by Local Authorities in England. We are going to draw a few graphs that might help us see some interesting features in the data.

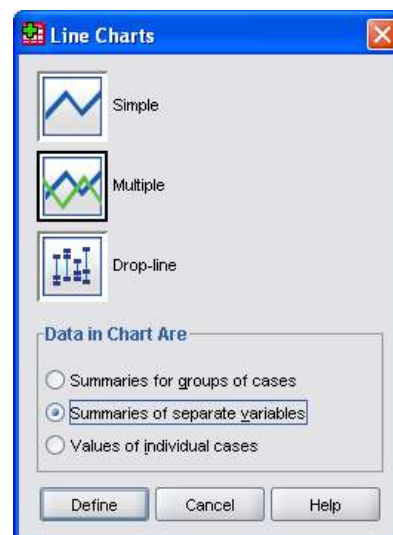
	year	bunder1	b1to4
1	1993	670	3300
2	1994	790	3300
3	1995	830	3600
4	1996	880	3900
5	1997	900	4300
6	1998	940	4600
7	1999	1100	4900

Choose **Line** directly from the "legacy" **Graphs** menu.

Select the option for **Multiple** lines and **Summaries of separate variables**.

Then press "**Define**".

Transfer the variables "Boys 1-4" and "Girls 1-4" to the top box and the "year" to the lower box then click **OK**.

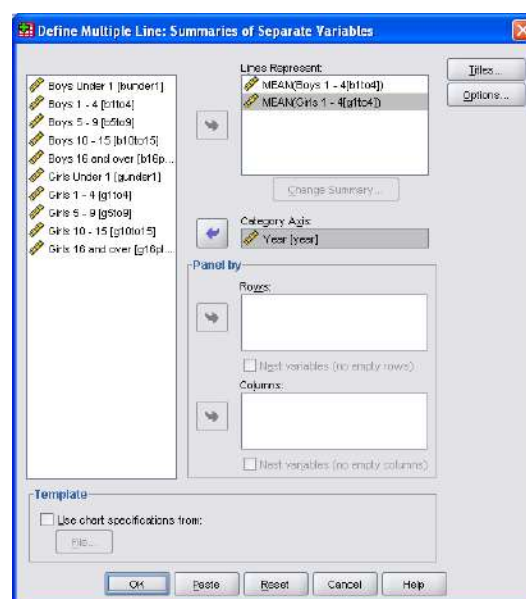


The graph that appears should let you answer the following questions;

1 In the 11 years covered by the data do the numbers of girls and boys aged 1 to 4 looked after by Local Authorities in England appear to increase?

2 Are the number of boys and girls in the age group 1 to 4 staying in roughly the same proportion, i.e. do they seem to increase or decrease together?

Now plot the data for the 16 and over age group, can you see any difference between the girls and boys?



The number of boys and the number of girls in the 1-4 age group has increased from around three to four thousand up to four or five thousand. They have increased together however; the number of girls is constantly a bit less than the number of boys.

Over the same years the difference between the number of boys aged 16+ looked after by the local authority and the number of girls in that age group has changed, the division between them has increased, there are now (at the end of this period) considerably more boys in this group than girls, this was not the case in 1993.



A final word about the structure of the data in the SPSS file...

In the example above we used the older type of graph menu - I did this because it suited the data structure that the DFES had supplied, if you are using data you have gathered yourself then you will need to decide on how it should go onto the computer.

The data we used in the example was already summarised or pre-aggregated. We weren't looking at the original data for the children but summaries of how many fell into various categories.

The original data for this would give each child one row of data, on the row would be their age (or age group) and their gender if a child was looked after for 3 years they would have 3 rows, one for each year. We would have 598220 rows of data!

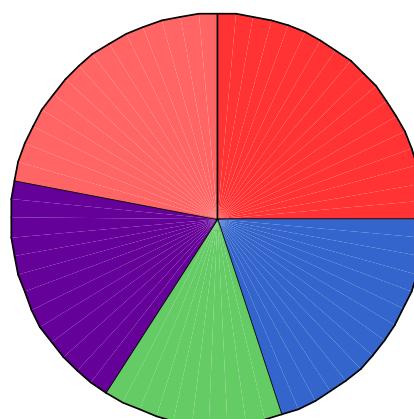
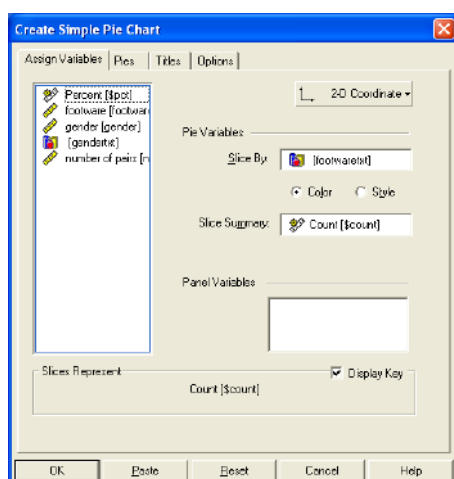
If you want a challenge or are reviewing these notes in desperation when faced with organising your own data in an analysable form then have a look at the file *Children looked after alternative format.sav* this goes some way to the ideal format, it gives the gender and age group in a separate variable - you could have a go at creating graphs similar to the ones above. The later methods are also worth investigating, it is largely a matter of personal preference; use the graph dialog that you feel happy with.



Task 9 Pie charts

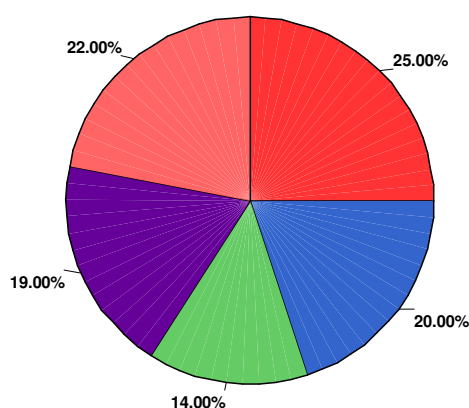
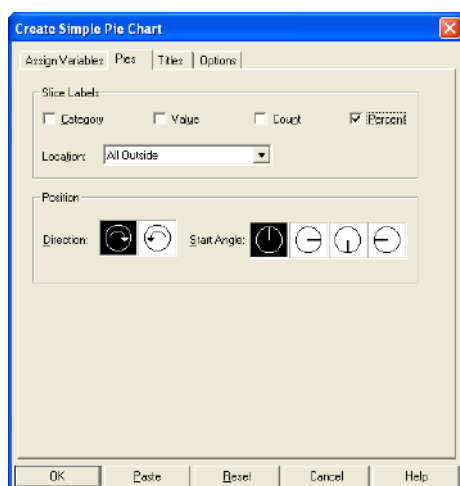
Simple Pie charts.

Pie charts are ideal for showing proportion. If you have raw data they can quickly show the proportion of subjects in various categories. To create the pie chart, Choose Pie, Simple, from the Interactive Graphs menu. (The example here uses the "*shoetypes.sav*" data.)



Pies show counts

We can improve this output by playing with the settings on the dialog box. For example we can show percentages on the chart. If the percentages don't display correctly try putting them inside the segments using the "location" dropdown on the "Pies" tab of the dialog. The best way to work out how to best exploit the system is to play with the options, do though make sure that the output you eventually get is appropriate and shows what you intended.



Pies show counts

For added fun drag the gender variable to the "panel variables" box. You should get two pie charts, one for each gender, this might help identify any differences between the gender groups in their choice of podiatric attire.



Pie charts that summarise pre-aggregated data.

Pie charts can be produced in two ways in SPSS (in common with many other types of graph). Rather than look at both methods we will concentrate on the Interactive Pie chart.

As well as this software oriented task we will also seek to draw some sensible conclusions about what type of data can best be displayed on a pie chart.

A simple example to get us going...

Load the file called **hip patient numbers**, this is a simplified version of the NHS hip fracture discharge data for 1997 to 1999 for England for patients aged 65 and over.

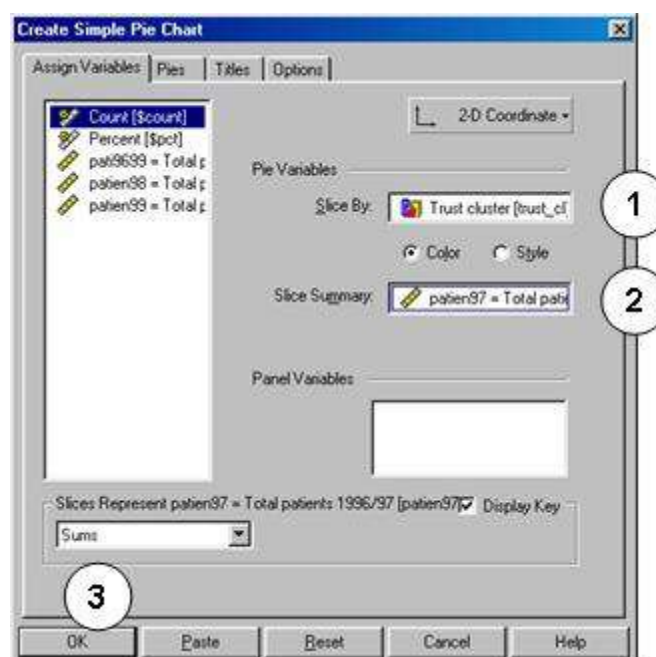
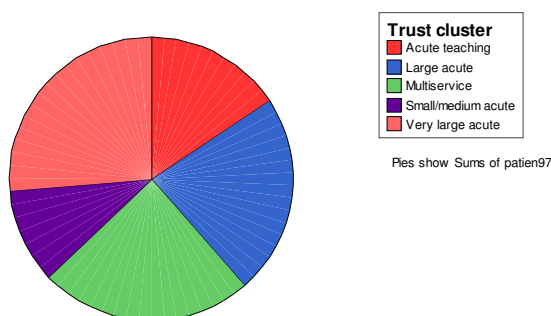
Trust type	1997	1998	1999	1997-9
Small/medium acute	4427	4447	4589	13463
Large acute	9329	9389	10345	29063
Very large acute	10774	10341	11528	32643
Acute teaching	6436	6564	6897	19897
Multiservice	9885	9485	11110	30480

Choose Pie, Simple, from the Interactive Graphs menu.

1. Drag the “**Trust Cluster**” variable to the “**Slice By**” box, this will tell SPSS to make each slice of the pie represent one type of trust (Small/medium acute, Large acute, Very large acute, Acute teaching, Multiservice).

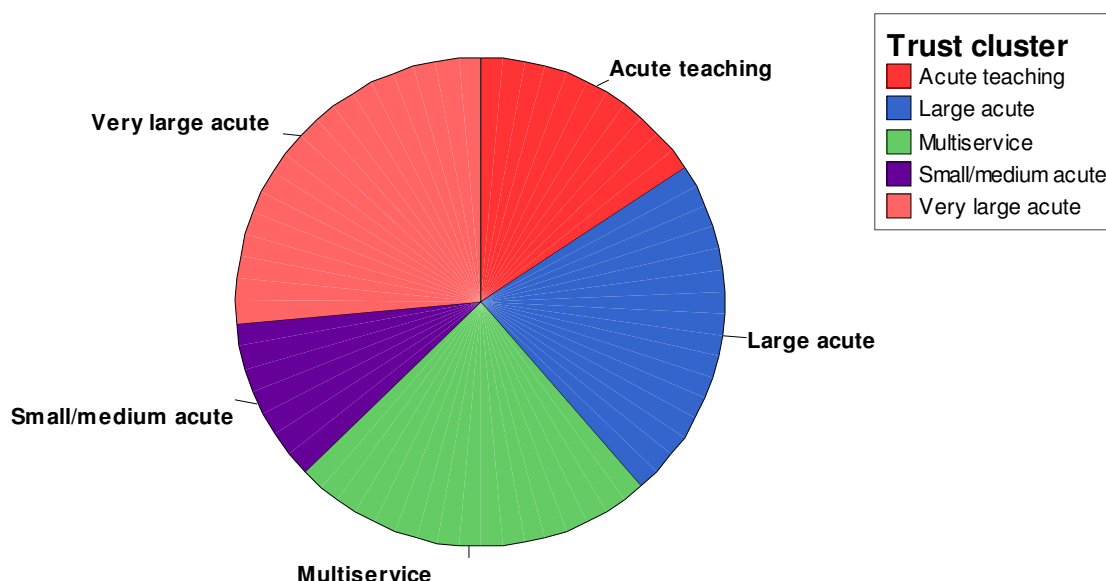
2. Drag the “**Patient 97**” variable to the “**Slice Summary**” box.

3. Press the **OK** button.



The default settings give a pie chart with the key to the slices at the side.

Playing with the various settings on the dialog box is probably the best way of exploring the features. In the chart below I added the cluster names as slice labels. If the labels don't display correctly try putting them inside the segments using the "location" dropdown on the "Pies" tab of the dialog.



Of the four types of data; nominal, ordinal, interval and ratio data (details in the glossary) what type is the **Trust Cluster** variable storing?

A pie chart that summarises the data.

Open the file called **hip fracture discharges**.

The Region variable tells us what region the hospital is in. Make each slice represent a region by dragging the **Region** variable to the “Slice By” box. Make the “Slice Summary” contain the **Patient 97** variable, i.e. the number of patients in 1997.

SPSS should produce a pie chart with one slice per region.

Multiple pie charts.

You may like to have a go at creating more than one pie chart at a time. This can be a useful tool for quickly seeing differences. We can have a go at seeing regional differences in the **hip fracture discharges** data.

Assign the variables as follows;

Slice By: typesize (typesize = Trust type/size)
Slice summary: patient97 (patient97 = Total patients 1996/97)
Panel variable: region (region = NHS region)

Look at the resulting graphs, they are basically the same graph as the one you produced earlier, however the data has been broken down into regions, on graph has been drawn for each. Can you think of any reasons why they are different from the first pie chart you created? Would you expect all the regions to be the same?

Summary: Pie charts, are used to show proportion, e.g. the number of votes cast for each party in an election. The pie should add up to 100% of the observed data. Pie charts should be clearly labelled and the units of measure displayed.



Part 2 - Inferential Statistics.

So far, what we have done has been descriptive statistics, a phrase that has a meaning beyond the menu in SPSS. Descriptive statistics are about what you can say about the data you have collected from your sample, including graphical representations.

We move on to inferential statistics when we try to draw conclusions about a background population from the sample. For instance, rather than just describing how our patients reacted to a drug, we try to predict that future patients will react in similar ways.

From Sample to Population...

In many areas we may choose to study it would not be practical to measure the entire population so we do our statistics on a sample of the population.

Statistics can be categorised into two areas, ones that describe and ones that infer.

So far we have looked at descriptive statistics, these describe the data and include the various forms of average we may use to quantify the level of data and statistics such as the standard deviation, which measure spread.

If we want to draw conclusions about an entire population from our sample we enter the realm of inferential statistics. This is where we draw inferences about the entire population from what we can see in the sample we have taken. Inferential statistics generally make the assumption that the sample is randomly drawn from the population we are studying. If the sample isn't random it would then seem sensible to restrict the inference to the segment of the population that the sample represented.

There are some issues around how well we can make the leap from sample to population, our aim is to be able to quantify how good an estimate our sample based statistics are at telling us things about the population.

More evidence generally makes for greater certainty so we can expect that bigger samples might be better, actually predicting what size sample might be needed to safely see a potential effect is yet another branch of stats and we won't cover now except to note that when we get a p-value that is not significant we generally can't use it as evidence of no effect, it simply means we didn't find the effect, this could be either because it doesn't exist or that our experiment wasn't powerful enough to detect a relatively small effect..

Inferential statistics give us some indication of the reliability of our inference about the population we make by analysing the sample.

Because the types of data we are collecting and analysing may be of different types (Ratio, Ordinal etc) and differently distributed in the population, we have an armoury of different tests for different situations, we will try just a few common ones but I'll also point to the others.

The mathematics behind what we are going to do is not the focus of our study so we'll get SPSS to do it for us so we can concentrate on what the statistics are telling us!



Task 10 A Parametric test

We have already looked at drawing informal inferences, for instance the data in the very first SPSS task we did where we analysed data from young and mature students. In the groups we studied, we concluded that the mature students talked more. This might have made you think that mature students at that university would be likely to talk more than the younger students would in future intakes. This would be drawing an inference. I.e. by looking at what happened in our example we can infer what might happen in the future.

We are now going to look at formal hypothesis testing. This is quite tricky to understand. It is worth spending time discussing after this task. You need to be clear why a null and an alternative hypothesis are required. You do not need to understand the details of the calculations that the computer does, but you do need to understand what the results mean.

Open the **waheig2S** worksheet.

This holds data on the heights of women of different ages (women, age, height). It consists of data on thirty individuals between 20 and 24 years old and another thirty aged between 50 and 54 years old, 60 in all, collected in 1980. (This is not paired data, these are 60 different women not the same 30 measured twice with 30 years between!) This issue of data structure can be confusing to new users of SPSS, especially if you've used MS Excel in the past, it might seem sensible to have a column for each age group, however would not be a very logical structure. The structure in the **waheig2S** file, with all the heights in one column and the adjacent column to hold a variable that tells us whether they are in the 20-24 or 50-54 age group is preferred by SPSS. The variable that tells us the group each research subject is in is often called a discriminatory variable or grouping variable, it lets us discriminate between the groups of people.

For the purpose of this task imagine that as a researcher on height you had only been able to collect a sample of this size, but it was a truly random sample. You want to know whether it provides enough evidence to decide between two hypotheses:

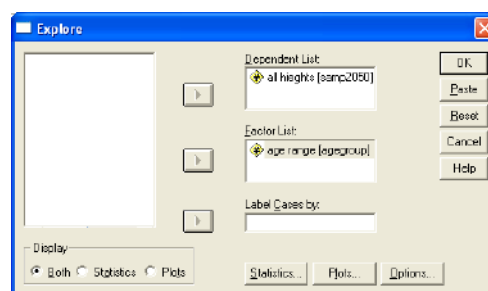
- alternative hypothesis: **the women in the younger age group tend to be taller than those in the older age group.**
- the null hypothesis: **there is no real difference between the heights of women in the age range 20-24 and 50-54 in Britain in 1980.**

Usually you will see the null hypothesis written first.

To do this we use a *test* to find how likely we would be to get results like these if the null hypothesis were true. (The null hypothesis is the one that says there is no effect.)

Before we conduct the test we'll have a quick look at the descriptive statistics. (From the menus choose Analyze, Descriptive Statistics, Explore.)

Put the height variable in the "Dependent list" box and the age groups in the "Factor list" box then click OK.



Look at the means for each group. They are certainly different. (162.5 and 159.083) and that the younger group has the larger mean height (162.5). These figures are *sample means* they simply describe the samples. They may make



us think that younger women are generally taller, but we have no idea if this really is the case. Since our sample was drawn at random from the population we would expect it to be a fair cross-section of each age group, however because it is a random sample we could have, just by chance, picked a non-representative sample. To find out if our inference about the population is reasonable we need to engage with inferential statistics. This means we can get some idea of how likely it is that such a difference in mean heights could be due to chance and based on this we can decide whether we accept or reject our hypotheses.

We are going to use a two-sample t-test to test whether the mean of the younger women in the population is greater than the mean of the older women. The Independent samples t-test is used on two different samples, not paired data, to test for differences in the population means. Check in the Glossary for examples of paired and non-paired data, also the SPSS v.11 help has a good example, to see it press the *help* button on the *Independent-Samples T Test* dialog box. (This is a *parametric* test - more about *parametric* and *non- parametric* tests later.)

Note that the mean of the *sample* of younger women *is* greater, but we are testing whether the data supports our belief that there is a real difference between the *populations* from which the samples are drawn.

The p-value we will get will give us the probability of getting this much difference in the means by chance, *if* there were really no difference between the populations. We are going to run an Independent samples t-test to test the means of our two samples. This test will test the Null Hypothesis, even though it's the alternative we are interested in – statisticians are funny like that!

To analyse the data using a t-test SPSS wants the data structured as we noted earlier, there may be several different files with different amounts of data and differing structures all with female heights in the data we are interested in are stored in the file called **waheig2S**. Make sure this is the file you are looking at, have a quick look at the structure. The structure of the data is an important issue, if the structure is inappropriate then SPSS might not be able to sensibly analyse or test it. The next few paragraphs address this again, it is fundamental to getting results from the SPSS system.

You will notice that there are sixty rows. SPSS will often refer to the rows as cases. A database system would call them records. This layout may seem just an excuse for complication, but it does make sense, really, let me explain...

The structure allows each persons' data to be stored on a different row; this is a real advantage if we have data where we are storing quite a few different things about each person. We can tell which age group the people belong to by looking in the second column (**agegroup**) of course you don't know which group is which, I've simply put a 1 or 2 in the column. In research this can be a useful trick, the researcher cannot tell the different groups she/he is studying apart, so there is less likelihood of bias in the work.

To find out what my coding means click to the **Variable View** and click the Values button for the **agegroup** variable - this should let you see the text values that I have attached to the numbers. These will be used in output to make it more readable - very important when



writing up research.

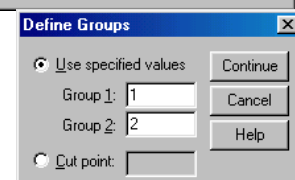
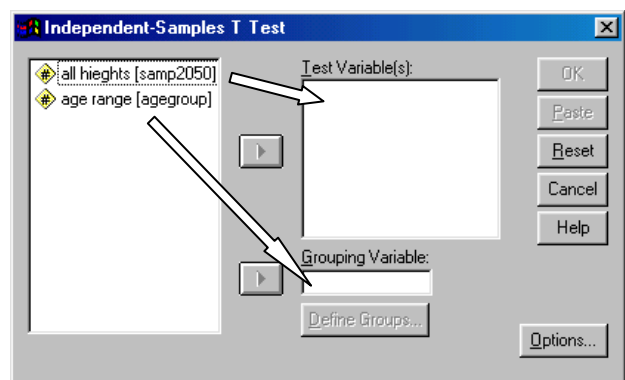
	Name	Type	Width	Decimals	Label	Values	Missing
1	samp2050	Numeric	11	2	all hieghts	None	None
2	agegroup	Numeric	8	1	age range	(1, age 20 - ...)	None

Click the **Cancel** button when you've seen enough and go back to **data view**.

We will now have a go at the test, hold tight...

From the menus choose **Analyze, Compare Means, Independent-Samples T-Test**.

We are testing the means of the heights of the two groups, so put the variable with **all heights** in it to the Test Variable(s) pane and the **age range** variable to the **Grouping Variable** box.



You will notice that the “age group” variable appears with question marks after it, this is because SPSS doesn’t know how to use the values in this variable to discriminate between the groups, click the Define Groups button to tell it, then click **Continue**. Then click **OK**.

T-Test

Group Statistics

	age range	N	Mean	Std. Deviation	Std. Error Mean
all hieghts	age 20 - 24	30	162.5000	5.4536	.9957
	age 50 - 54	30	159.0833	5.1897	.9475

SPSS calculates the means for each group in the sample. If you like, check them against the values you got earlier.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
all hieghts	Equal variances assumed	.094	.761	2.486	58	.016	2.4167	1.3745	.6654	6.1679
	Equal variances not assumed			2.486	57.858	.016	3.4167	1.3745	.6653	6.1681

The significance figure here lets us know which of the two rows of figures to look at. SPSS has tested the variance of the two groups and given us two sets of figures, the one we use will depend on whether the variance is the same for each group. If the figure here is less than 0.05 use the lower set of figures. In this case use the upper ones (since there is not a significant difference in variances). If in doubt, or unsure then use the lower row, where equal variances are not assumed.

The figure we are after is under **Sig. (2-tailed)**.

If this figure (the p-value) is less than 0.05 we can reject the null hypothesis (The null hypothesis says there is no difference). In our example we can reject the null hypothesis (0.016 is less than 0.05), so we can accept the alternative hypothesis that says there is a significant difference between heights of the two groups of women. The descriptive statistics will enable us to say in which direction the difference lies. Later we will see how to interpret the results of this 2-tailed procedure for a one tailed alternative hypothesis.



The results you see should be like the ones above: Notice also the alternative spelling of the word “heights”, it’s worth remembering that SPSS has no spelling checker!

You can read about significance testing, hypothesis testing, parametric tests and p values in the Glossary and there are plenty of books and websites that will help you broaden your understanding. Write some short sentences saying what the p-value tells you about the data. Which hypothesis would you accept?

I stated earlier that my hypothesis was really a “one-tailed hypothesis” this, put simply, is what we have when our null-hypothesis can only be rejected in one way. I was saying in the alternative hypothesis that the younger women were significantly taller than the older women, not just that they were significantly different in height but we don't know if they are taller or shorter. The two tailed test was testing that the mean heights of the two groups were not significantly different. The low p-value let us reject this.

Can you have a difference in means without it being significant? Yes! if you don't believe me re-do the last test using the file *waheig2Stest.sav* which contains samples with the same means, but altered to not give a significant p-value.

Now have a look at **waheig1S** and run the test on that very large sample. What happens to the p-value? Why? What does this tell you about the hypotheses?

If you have time you may want to experiment with deleting some of the rows of **waheig2S** to see what kind of p-values you might have got if your sample had been even smaller.

Put simply a larger sample gives us more certainty because our decision is based on more evidence.

Task 11 Using a Non-parametric Test

Open the **studentss** worksheet (note the extra “s”). The file has all the numbers representing the number of times each student contributed in the variable called “speakn” and the age group in the variable called “grp”. Each row of this data represents a student, the number in the “speakn” column is the amount they contributed and the number in the “grp” column tells us their age and year grouping. The middle column is just some text to help you see which group is which, if you go to variable view you will see the “grp” variable labels similar to the ones explained in the previous task.

We can take the observed data as a sample of all student contributions to classes over the whole year. We want to know whether the mature first year students do really contribute more on average, or whether the data we collected only showed this by chance.

We will use a *non-parametric* test called the Mann Whitney test to test whether first year mature students contribute more than younger first year students do.

Non-parametric tests don’t depend on many assumptions about the underlying distribution of the data (e.g. whether it is normally distributed or not.). They are used widely to test small samples of *ordinal* data. The test decision chart later in this document gives structure for deciding which test to use, there is also there is also plenty of advice on the web.



The SPSS help system describes the Mann-Whitney U test as a non-parametric equivalent to the t-test. It can be used to test whether two independent samples are from the same population (i.e. are they of a similar level).

Write down a null hypothesis and an alternative hypothesis below (for help have a look at the ones in the previous example - remember the null hypothesis is the one that says there is no change). Remember, we want to know whether the mature first year students contribute more than young first year students.

- the null hypothesis: _____.
- alternative hypothesis: _____.

Have a go at filling in the blanks before looking in the box below.

A possible pair of hypotheses;

the null hypothesis: there is no real difference between the number of contributions made by young and mature first year students
alternative hypothesis: the younger students tend to contribute less frequently than the mature students in their first year. (as indicated by the descriptive statistics)

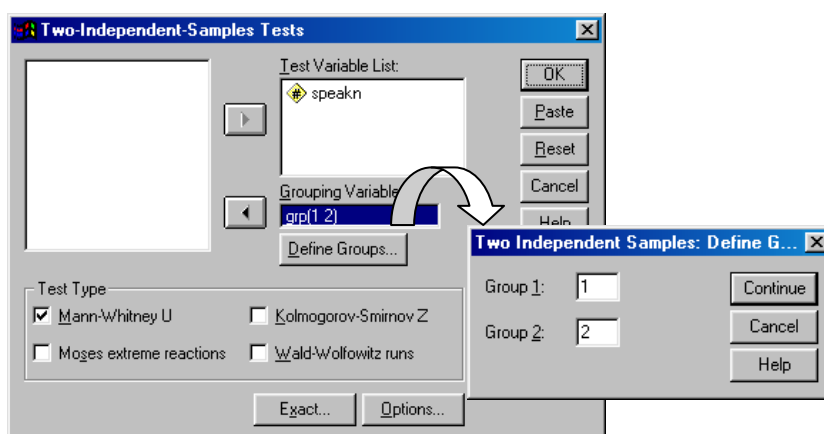
Applying the test.

From the menus choose **Analyze, Nonparametric Tests, 2 independent samples.**

Notice that although we have three variables in the data only the two numeric variables are shown in the left pane of the dialog box.

Don't forget to define the groups (1 = Year1 young and 2 = Year1 mature).

Notice that the Mann-Whitney test is ticked.



Click **OK** and see the results appear.

Below are the results you should see,



NPar Tests Mann-Whitney Test

Ranks				
	GRP	N	Mean Rank	Sum of Ranks
SPEAKN	1	12	8.46	101.50
	2	11	15.86	174.50
	Total	23		

Test Statistics^b

	SPEAKN
Mann-Whitney U	23.500
Wilcoxon W	101.500
Z	-2.618
Asymp. Sig. (2-tailed)	.009
Exact Sig. [2*(1-tailed Sig.)]	.007 ^a

a. Not corrected for ties.

b. Grouping Variable: GRP

Interpreting the output.

In the previous task (the t-test) we used a value in the output to enable us to decide if the result was significant. We will do the same in this test, however I thought you could do the work this time!

The three boxes on the next couple of pages contain statements from the SPSS help system. SPSS is rather like a car salesperson, they never tell you whether a car is a good car or a bad car, they just tell you things about it!

SPSS and the statistics it calculates will not tell us whether mature students definitely talk more than young students but it will give us a set of indicators which can help us to decide if this is likely to be true.

Read the first box, about **Observed Significance Level** then answer the question below it before continuing to the second and third boxes.

Observed Significance Level

Often called the **p** value. The basis for deciding whether or not to reject the null hypothesis. It is the probability that a statistical result as extreme as the one observed would occur if the null hypothesis were true. If the observed significance level is small enough, usually less than 0.05 or 0.01, the null hypothesis is rejected.

Question

The null hypothesis says (if you can read upside down): there is no real difference between the number of contributions made by young and mature first year students. We would like to reject this in favour of our alternative view of the world, which says there is a difference. Do we want the p-value to be;

- A As big as possible
B As small as possible

Put your answer in the box, then read on.

I hope you put the letter B in the box, it is quite confusing, if you got it wrong read through it again.

Frustratingly, SPSS gives us two p-values with this test, *Asymp. Sig.* = 0.009 and *Exact Sig.* = 0.007 The next two boxes tell us which value we should use from the output. Bear in mind that our sample is quite small, which figure from the results should we use as our p-value?



Asymptotic Significance (Asymp. Sig.)

The significance level based on the asymptotic distribution of a test statistic. Typically, a value of less than 0.05 is considered significant. The asymptotic significance is based on the assumption that the data set is large. If the data set is small or poorly distributed, this may not be a good indication of significance.

Exact Significance (Exact Sig.)

The significance level based on the exact distribution of a test statistic. When the data set is small, sparse, contains many ties, is unbalanced, or is poorly distributed, it is preferable to calculate the significance level based on the exact distribution.

What is the p-value in this case ?

Refer back to the first box, decide whether we can reject the null hypothesis.

Another question for you.

Tick the statement you think the test supports.

☐

We can reject the null hypothesis in favour of our alternative hypothesis, which says there is a difference in the amount young and mature first year students contribute.

☐

We cannot reject the null hypothesis, the data does not support the view that there is a difference in the amount young and mature first year students contribute.

I think there is evidence to strongly support our alternative hypothesis, $p = 0.007$ (Exact Sig.), i.e. we can reject the null hypothesis, we seem to have an effect.

In this example we can be more specific. The descriptive statistics would suggest that the alternative hypothesis should be “mature students speak more than young students” – it is appropriate to use the exact significance figure. Since our alternative hypothesis is really one-tailed we can quote a p-value of 0.004 (half of 0.007 to three significant figures.) as Brace, Kemp & Sneglar (2000) explain in relation to the Wilcoxon Signed Ranks Test we will meet soon. However, be careful of one-tailed tests, a reviewer might question the justification for a one-tailed interpretation! If in doubt always stick to two tailed tests.

If you have time, test to see if;

- first and second year young students contribute differently in class,
- mature and young second year students contribute differently in class.

Remember; a p-value less than 0.05 indicates a significant effect, if it is over 0.05 then we can't reject the null hypothesis and can't claim to have detected a significant effect.

Key to groupings.

First year young students = Y1 = 1

Mature first year students = M1 = 2

Young second year students = Y2 = 3

Mature second year students = M2 = 4



Task 12 Testing Paired Data

Neither of the last two tests have involved paired data. You can read about paired data in the glossary.

Open the worksheet **Step**.

The data in this file come from an experiment to see whether subjects could perform more step exercises in a fixed time in a group or on their own. A physiotherapy student collected them as part of a third year project.

Paired data often occur in ‘before and after’ situations. They are also known as ‘related samples’. These data are paired, it’s the same person doing step exercises under two different conditions.

To deal with these we use a paired t-test (parametric) or the Wilcoxon test (non-parametric).

Write down a null hypothesis and an alternative hypothesis. (Decide on these yourself. Remember the null hypothesis is the one that says there is no change.)

We will use the non-parametric test, as we have no good reason to think number of steps completed is normally distributed and because the dataset is small. A non-parametric test is not as powerful as a parametric test. (Later in this document I have put in an SPSS technique to help check if data are normally distributed but for now we’ll assume a non normal distribution.)

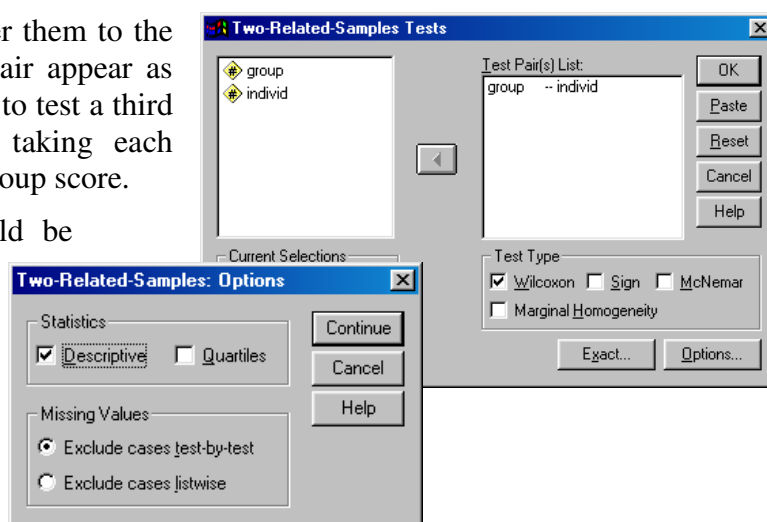
Applying the test.

From the menus choose **Analyze, Nonparametric Tests, 2 related samples**.

Select both variables then transfer them to the Test pairs list. Notice that the pair appear as “group – individ”. SPSS is going to test a third column of figures derived by taking each individual score away from the group score.

The Wilcoxon check box should be ticked.

In addition to the test we will also ask SPSS to provide some descriptive statistics. Press the **Options** button, select “**Descriptive**” then **Continue** and **OK**.



We will only be testing our null hypothesis, which says there is no difference between the number of steps done individually or in a group. If we have evidence to reject the null hypothesis we can look to our descriptives to indicate which way the evidence lies.



NPar Tests

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
GROUP	12	86.67	17.87	58	110
INDIVID	12	79.92	16.40	58	104

Wilcoxon Signed Ranks Test

Ranks

	N	Mean Rank	Sum of Ranks
INDIVID - GROUP Negative Ranks	10 ^a	7.25	72.50
Positive Ranks	2 ^b	2.75	5.50
Ties	0 ^c		
Total	12		

a. INDIVID < GROUP

b. INDIVID > GROUP

c. GROUP = INDIVID

Test Statistics^b

	INDIVID - GROUP
Z	-2.631 ^a
Asymp. Sig. (2-tailed)	.009

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

What do the results show?

First what are we testing? my hypotheses were...

Null hypothesis: *Group or individual conditions made no difference to the number of steps subjects completed.*

Alternative hypothesis: *Subjects had a tendency to complete more steps under group conditions than under individual conditions.*

The alternative hypothesis is in this case a “one tailed” hypothesis, a “two tailed” version would be “*Subjects had a tendency to complete a different number of steps under group conditions than under individual conditions.*” That would leave two options for the difference to be more or less. We will deal with this issue.

The bottom row of the Test Statistics box indicates a p-value of 0.009 which being smaller than 0.05 allows us to reject the null hypothesis – the data supports our alternative. However this is a two-tailed significance figure, we can halve this for a one tailed test to 0.0045. ($0.009 \div 2$) As explained by Brace, Kemp & Sneglar (2000). This may seem like a bit of a fix, especially when it makes the result appear to more strongly support our



finding, that's one reason I have put in an external reference to share the blame. However if you look at sampling statistics, particularly the way that the statistics provided by samples are distributed around the real population statistic it can be shown that this is justified. (If you want to delve deeper into this, one of the best explanations of this I have come across is in *Statistics Without Tears* by Derek Rowntree (1981).)

If we were including these findings in a report, we could put something like...

Subjects had a tendency to complete more steps under group conditions than under individual conditions. ($p = 0.0045$, one-tailed (Wilcoxon signed ranks test.))

There are other figures we could put in but for our purposes here this is enough.

There is a convention that p-values below 0.05 are called significant, p-values below 0.01 are called highly significant, and p-values below 0.001 are called very highly significant. They are often marked *, **, and *** respectively in tables of results.

It can be quite difficult comparing small decimal numbers, read the paragraph above and decide which of the statements below are correct with reference to our derived one-tailed p-value of 0.0045. Tick a box below.

- | | | |
|---|--|--------------------------|
| 1 | The above results are significant. | ☐ |
| 2 | The above results are very highly significant. | ☐ |
| 3 | The above results are not significant. | ☐ |
| 4 | The above results are highly significant. | ☐ |

It is important to note that a high p-value does not mean that the alternative hypothesis is false, but only that your data do not provide good evidence for it. (I think the term "highly significant" is appropriate.)

If the alternative hypothesis is really true, large samples are more likely to give statistically significant results than small ones. This can be an issue in, for example, drugs testing. A company can test a large number of subjects and therefore reduce the p-value, their findings would be significant. However that doesn't make the effect of the drug more significant. Don't confuse statistical significance with clinical significance.

It is also important to note that a low p-value does not prove that your results are not due to chance, but only that they are unlikely to be due to chance. (It is worth noting that if you apply 20 different tests to samples from a large set of data you are likely to get at least one result significant at 0.05 even if none of the alternative hypotheses are true. Remember 0.05 is the same as 1 in 20. Our 0.05 p-value can be said as, the results would happen by chance, five times out of one hundred or once in every twenty.)

Task 13 Correlation

Open the file **Heathip**. This file contains data from a student project on the effect of heat on hip stretches. The first column gives the subject's height, and the second column gives



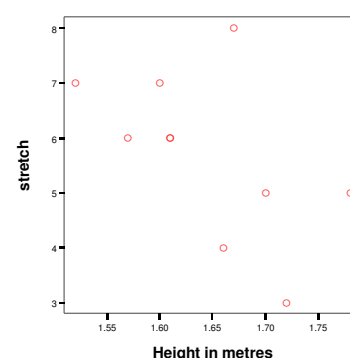
the increase in hip extension after stretching exercises. (Other columns relate to the discomfort experienced, and the stretch and discomfort when heat is used.) The student was interested in whether the increase in stretch was related to the subject's height.

Look at the data and decide for yourself if this file consists of paired data?

If you aren't sure then think about the question "is this looking at the same people under different conditions?"

The conclusion I hope you came to being that the file is made up of paired data, each row has one patient's data on it. The data represents measurements taken under two conditions.

Plot a scatter diagram with height on the x-axis and stretch (without heat) on the y-axis. (To draw the scatterplot click the **Graphs** menu, then choose **Interactive, Scatterplot**. Drag the "height" variable to the horizontal axis and the "stretch" variable to the vertical axis. Click the **OK** button and your graph should eventually appear in the SPSS viewer. (Refer back to the previous books if you need help on drawing and interpreting scatter plots.)) Describe what you see as clearly as possible by answering the following questions.



You can read about correlation in the Glossary and say what kind of correlation is involved here. The questions below may help.

Do the points appear to form a line? _____.

If they do is it a clear, quite thin line or more like a cloud? _____.

Does it slope up or down from left to right? _____.

Look at your answers and decide if there is a strong, weak or no correlation. Is it positive or negative? _____.

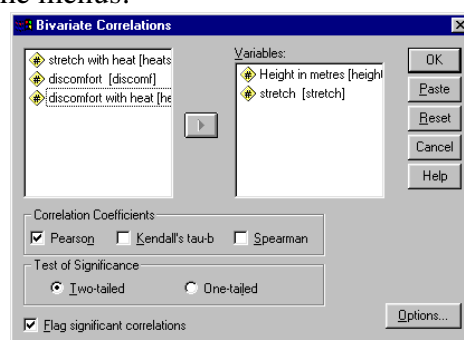
Calculating the Pearson's correlation coefficient.

We are going to get SPSS to calculate the Pearson's correlation coefficient to get a numerical indication of any correlation between the variables **height** and **stretch**. When we have the figure, read about the correlation coefficient in the glossary, satisfying yourself that you understand the connection between the correlation coefficient and the diagram.

The Pearson correlation coefficient is a form of Bivariate Correlation. To obtain Bivariate Correlations in SPSS choose from the menus:

Analyze, Correlate, Bivariate...

Select the variables **height** and **stretch**.





You may notice that there is more than one method of calculating correlation coefficients, we will use Pearson's. (The Spearman is for nonparametric data, scores, lickert scales etc.)

Leave the test of significance set as it is. (You can select two-tailed or one-tailed probabilities. If the direction of association is known in advance, select One-tailed. Otherwise, select Two-tailed.)

Run the test and look at the results.

SPSS provides the results in the form of a matrix and as usual provides lots of extra information to confuse new users, but don't get annoyed, its only trying to help!

We are really being shown four correlations;

HEIGHT with HEIGHT	HEIGHT with STRETCH
STRETCH with HEIGHT	STRETCH with STRETCH

But "HEIGHT with HEIGHT" and "STRETCH with STRETCH" will of course be perfect correlation. (Correlation coefficient =1).

Correlations

Correlations

		HEIGHT	STRETCH
HEIGHT	Pearson Correlation	1.000	-.548
	Sig. (2-tailed)	.	.101
	N	10	10
STRETCH	Pearson Correlation	-.548	1.000
	Sig. (2-tailed)	.101	.
	N	10	10

A strong correlation gives a number near to 1, weak is near 0. A minus sign means a negative

Later we will see why the results can be usefully presented in a matrix, but for now we will concentrate on the possible correlation between HEIGHT with STRETCH.

Look at the information in the square indicated by the arrow, it tells us three things;

Pearson Correlation -.548
Sig. (2-tailed) .101
N 10

We have a weak negative correlation.
The correlation is not significant at the 0.05 level.
There were ten pairs of data.

Be careful not to confuse correlation and significance. On the next page we look at this in more detail.



Significance in perspective.

There are a few things here that may be of interest about correlation and significance;

- Looking for correlation is different from looking for increases or decreases – we will address this in more detail soon.
- Correlation does not necessarily mean a causal relationship. Just because two values appear to go up and down together does not mean one is causing the other.
- The Pearson's coefficient is designed primarily for looking at linear relationships. Two variables can be related, but if the relationship is not linear, Pearson's correlation coefficient is not an appropriate statistic for measuring their association.
- The number of observations as with other statistics effects the significance.

This last point can be demonstrated quite nicely with the data in this file. What we are about to do is not in any way statistical good practice so please don't save the file with these changes, just in case you go back to it and assume it is genuine. If you have time do the following...

Click and drag with the mouse to select (highlight) all the data. (50 cells in all, 10 rows and 5 columns.)

Choose **Edit** then **Copy** from the menus.

Click in the cell on row 11 under the HEIGHT column and choose **Edit** then **Paste** from the menus.

SPSS thinks we now have twice as much data. We are of course conning it, purely to see the effect of sample size on significance – this is not something we would do in real research. Run the correlation coefficient again and compare the results. (This is a rather false exercise since although we are doubling the amount of data we are not adding different values so we are unlikely to make a realistic change to the variability.)

Has the correlation coefficient changed?

Has the result become more significant?

What is the effect if you delete all but the first two rows?

To get back to the unadulterated version of the data open the file **Heathip** again but **DO NOT** save the copy you have altered when asked.

Check there are now only 10 rows when you reopen the file.

If you have time you may like to see if stretch and discomfort are correlated.

To see how a larger matrix of results looks try the three variables *Height*, *stretch* and *discomfort*. When more than one variable are being examined the matrix output is an advantage.



Looking for Correlation is different from looking for increases or decreases

Open the file **Step** and draw the scatterplot. Plot a scatter diagram with **individ** on the x-axis and **group** on the y-axis.

To draw the scatterplot click the **Graphs** menu, then choose **Interactive, Scatterplot**. Drag the “individ” variable to the horizontal axis and the “group” variable to the vertical axis. Click the **OK** button and your graph should eventually appear in the SPSS viewer. (Refer back to Tasks 6 and 7 in the previous book if you need help on drawing and interpreting scatter plots.)

Last time we looked at this diagram we were looking at whether subjects had increased or decreased their number of steps when in a group instead of individually. We also looked at the shape of the points on the plot to see if there was correlation between doing well individually and in a group. i.e. did people who performed well individually also perform well in a group?

Look at whether subjects who did more steps (compared with the others) under individual conditions, also did more under group conditions.

Write down what the diagram tells you about this. _____
Describe the kind of correlation (if any), you see. _____
(Is it strong, weak, positive, none, negative?) _____

We are going to find the correlation coefficient. However you may notice that there is more than one option on the "Bivariate Correlations" dialog. The ones we are interested in are Pearson and Spearman.

The Pearson correlation coefficient assumes that the variables are normally distributed, it is a parametric test. The Spearman correlation coefficient assumes that the variables are not normally distributed, it is a non-parametric test. There are other issues that may effect your choice of test but for now we'll stick with normality.

Generally parametric tests are considered more powerful, they carry more weight with statisticians. In this case I've no reason to expect the number of steps an individual can do to be normally distributed, so select the Spearman test. (It is not good practice to select both then try to argue that the one that gives the results you want is the best to use!)

NOTE: Checking for normality: You can graphically compare a sample to a normal distribution with the Q-Q plot. In the Q- Q plot the normal distribution is represented by a straight line (the bell shape is squashed flat), your data is plotted round it. Data points from a normal distribution would appear close to the line. Q-Q plot is under the Graphs menu. See the section later on methods to test whether your data is normally distributed for a full account.

Correlation: Descriptive and Inferential Statistics



When you describe the correlation that you see in the scatter diagram or calculate the correlation coefficient you are doing descriptive statistics: you are talking about the *sample*.

When you infer from the correlation in the sample that correlation also exists in the *population* then you are doing inferential statistics.

This can be done informally by looking at the pattern and deciding whether, given the number of dots and the amount of slope they show, this could have been due to chance or not.

You can also use the formal method of carrying out a hypothesis test to obtain a p-value for the likelihood that the results are due to chance.

Here it is important to notice that the strength of *evidence* for correlation in the background population is not the same as the strength of the *correlation*. A few points in a very straight line could be due to chance, whereas a very large number of points in a scattered shape, which shows some slope, would be unlikely to occur by chance.

P-values a summary.

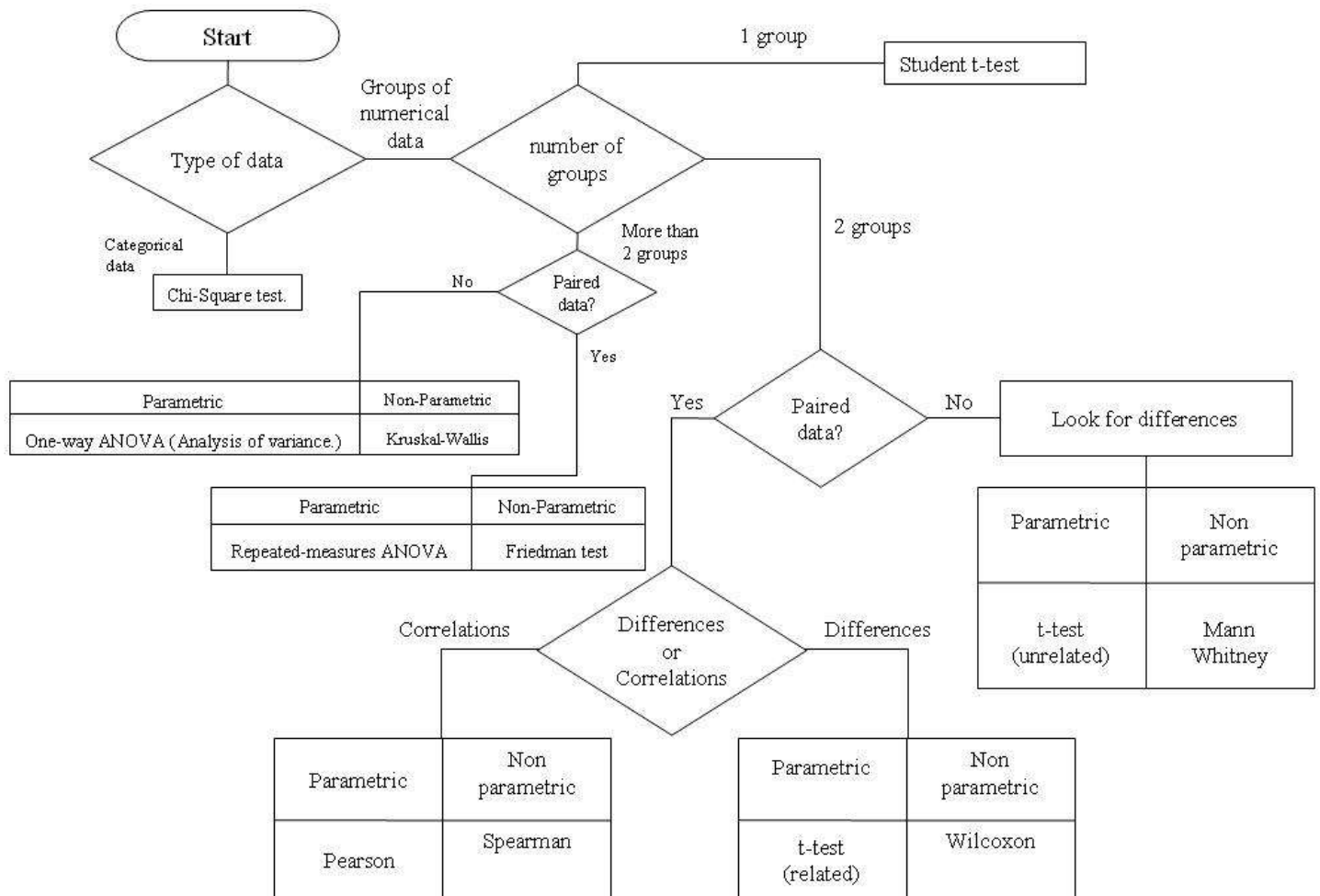
"P-values do not simply provide you with a Yes or No answer, they provide a sense of the strength of the evidence against the null hypothesis. The lower the p-value, the stronger the evidence. Once you know how to read p-values, you can more critically interpret journal articles, and decide for yourself if you agree with the conclusions of the author. " - TexaSoft, (1996-2001)

What have we learned so far?

By now you should have a good idea about what p-values mean and perhaps some feeling for the vagaries of sample based statistics and how some tests are more appropriate in some situations.

This is important knowledge for a consumer of stats, someone who is basing their activity on the best available research. When you read research, a basic understanding of the techniques underlying the data analysis is important if you are going to be able to judge the value of the research for yourself.

The next part of this document covers how a researcher might select tests, after this are a couple of examples, there are also more examples later in the document. The other tests covered later in this document are Chi-Square and ANOVA.



Test decision chart.

The above chart is similar to ones you may see in statistics textbooks.

The various tests we have seen in SPSS have restrictions. Typically non-parametric tests make fewer assumptions about the distribution that the data come from. Parametric tests generally assume the data come from a normal distribution (that's the bell shaped one that appears so often in nature). The non-parametric tests don't assume a normal distribution, and are more suited to smaller samples. Non-parametric tests are also called distribution-free tests because they don't assume that the data are from a known distribution.

There are tests that can be carried out on data to help you decide whether the data are normally distributed and so help decide what tests are useful. (These would be non-parametric tests, since we don't know if the data is normally distributed or not.)

Other decisions are about whether the data is paired or not. Often, paired data occurs in before and after situations, e.g. where the same thing is measured under two conditions.

On the following page are two rather contrived examples of research,

- Work out whether the data is paired or not.
- Think about whether the data is parametric; if you're not sure think how the researcher might find out.



- Think about what kind of information the researcher is trying to extract, e.g. are they looking for differences or things changing together?
- Decide on the tests to use by using the flow chart above to help.

1 Oil pollution and plant growth.

A biologist wants to study the effect of oil pollution on plant growth. He sections off two separate square metres of a lawn, the areas have the same grass type, are on similar soil and get the same amount of sunlight etc. He randomly measures 20 blades of grass from each square metre. One of the areas is then sprayed with a 10:1 mixture of lead free petrol and two-stroke oil. The grass is then left to grow for 2 weeks. After a week another 20 blades of grass are measured from each area. These are not necessarily the same blades of grass measured before – there were worries that marking the grass may alter the effect of oil pollution.



He expects to see a greater increase in the height of the non-oiled grass.

What would he compare and how?

He opts to compare the two readings taken before, they should be similar he hopes. This he says will show that the two areas are comparable. He then is intending to compare the samples derived after the week's growth, if the oil is restricting growth the polluted sample should be shorter, and since the other factors are unaltered any difference can be attributed to the presence of oil.

What test should he use to compare the samples?

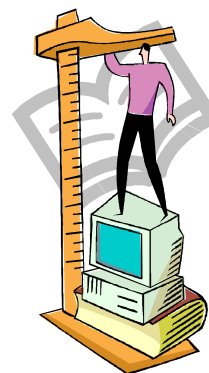
What, if any, possible loopholes can you see in the methodology?

2 Can computer games improve arithmetic ability?

A researcher believes that playing a certain computer game can improve his ability to do arithmetic.

She sets up a regime where she plays the game for approximately a half hour period each morning and afternoon. Each evening she takes a simple timed arithmetic test. The test is different material each time.

Her results comprising of daily data for 10 days, include the cumulative amount of time spent on the game, the test score and the time taken to complete the test.



She actually always got all twenty questions correct each evening in the test, however she feels that she completed the tests taken later in a shorter time. I.e. she got quicker.

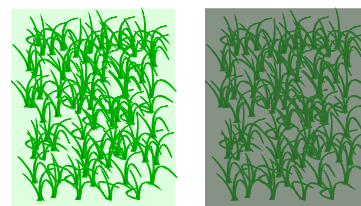
What are her options for testing the data?

What, if any, possible loopholes can you see in the methodology?



1 Oil pollution and plant growth.

- Work out whether the data is paired or not. *The researcher would not be gathering paired data. To do this he would need to tag the blades of grass being measured and re-measure them after the non-control group had been treated.*
- Think about whether the data is parametric; if you're not sure think how the researcher might find out. *The height of grass will probably be similar to the heights of humans in that it will be normally distributed – a QQ plot could help check this. (I am of course referring to the shape of distribution of the heights of grass and humans! most grasses are shorter than most humans.)*
- Think about what kind of information the researcher is trying to extract, e.g. are they looking for differences or things changing together? *The test will hopefully show that there is no significant difference between the untreated samples before treatment, but after treatment there will be a significant difference. This would suggest using an independent samples t-test for parametric data (or Mann Whitney test if you decide the data is non-parametric or you are not sure and want to play it safe).*
- Loopholes? *I can think of two possible problems;*
 - 1 *The experiment only holds up if there is no other difference between the patches of grass.*
 - 2 *We could inadvertently be measuring the petrol-oil mixture's effect on the pests that effect growth rather than the direct effect.*



2 Can computer games improve arithmetic ability?

- Work out whether the data is paired or not. *This data is paired; each day has a value for both total amount of game use and speed of completing the test correctly.*
- Think about whether the data is parametric; *he cannot be sure the data is normally distributed, a Q-Q plot could help here or one of the tests for normality, the sample is quite small so I would err on the side of a non-parametric test if I were advising the researcher.*
- Think about what kind of information the researcher is trying to extract, e.g. are they looking for differences or things changing together? *She could look for a correlation between the time spent on the game and the speed of test completion. If here feelings are correct then there will probably be a negative correlation between the variables, i.e. as the total time on the game went up the time for doing the test went down. The non-parametric test for correlation is the Spearman test.*



Loopholes? *Could doing the tests each night be increasing her ability?*



The Chi-Square Test. (χ^2)

(*Chi is pronounced "ky" as in "sky".*) The Chi-square test for independence can be used in situations where you have two categorical variables. It works with the "simplest" form data. Data such as *gender* or *country*, or data that has been placed in categories, such as *age group*.

What can you apply the test to?

The SPSS documentation states that the test can.... "Use ordered or unordered numeric categorical variables (ordinal or nominal levels of measurement)." and on assumptions about the data.... "Nonparametric tests do not require assumptions about the shape of the underlying distribution. The data are assumed to be a random sample. The expected frequencies for each category should be at least 1. No more than 20% of the categories should have expected frequencies of less than 5."

Cross-tabulation

To begin we will look at one of the most common methods of analysing this categorical type of data - cross-tabulation – this method is useful in its self.

Let's answer a question about our University...

Does the ratio of males to females in each school in SHU reflect the overall ratio in the university? (or put another way is there a larger than expected number of one gender in some schools?) The data we have available are from a survey of students done in 2001. It will inevitably not cover all the students, so we could question the randomness of the sample. However for our purposes I thought it would be more interesting than studying the geographical distribution of red and grey squirrels!

The data records the students' gender and their school.

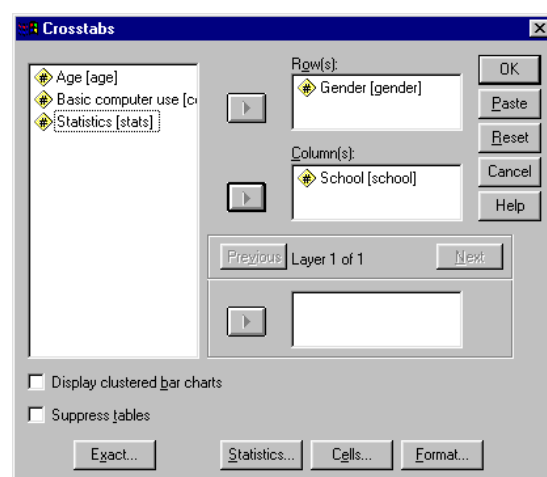
Load the file called "**Students data 2001.sav**"

You will see that the data is all numeric. If you want to know what the numbers represent you can look under the Variable View to find out, but this isn't necessary for our purpose. The crosstab system automatically labels the output!

Choose, **Analyse**, **Descriptive Statistics**, **Crosstabs** from the menus...

Put the **Gender** variable under Row(s)

Put the **School** variable under Column(s) then Press **OK**.





The table we are interested in is the Cross-tabulation table rather than the Case Processing Summary (the Case Processing Summary tells us about incomplete records etc.) Look at the cross-tabulation table, I hope you'll agree it's a useful form of analysis. (If you need to do similar analyses but don't have access to SPSS Microsoft Excel has a similar system but calls it a Pivot Table)

Gender * School Crosstabulation

Count		School									Total
		EDS	HSC	SCI	SED	SLM	CMS	SSL	ENG	SCS	
Gender	Male	54	38	125	250	284	342	53	54	142	1342
	Female	141	283	125	57	307	56	132	1	142	1244
Total		195	321	250	307	591	398	185	55	284	2586

A quick look at the figures can give us some idea about the male - female representation in the schools. Look at Health and Social Care (HSC), how does it compare to Engineering (ENG)? What does the total column at the end say about the overall ratio of males to females in this data?

If the genders were represented equally in each school then we would get the same number in the male and female rows for each column, however this would lead to equal numbers in the total column. This can't be the case since in the figures in the totals column show slightly more males than females overall in the university, if the distribution across the schools reflects the overall ratio in the university we would expect there to be slightly more males than females in each school.

SPSS can work out the expected values in the schools for us.

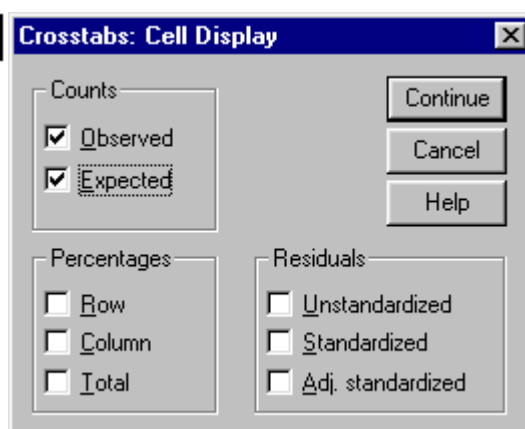
Get back to the **Crosstabs** dialog box.

Cells...

Click the **Cells** button, then select "**Expected**" under the counts section.

Click **Continue** then **OK**.

The resulting output has the expected number in the cell along with the observed number.



Gender * School Crosstabulation

			School									Total
			EDS	HSC	SCI	SED	SLM	CMS	SSL	ENG	SCS	
Gender	Male	Count	54	38	125	250	284	342	53	54	142	1342
		Expected Count	101.2	166.6	129.7	159.3	306.7	206.5	96.0	28.5	147.4	1342.0
	Female	Count	141	283	125	57	307	56	132	1	142	1244
		Expected Count	93.8	154.4	120.3	147.7	284.3	191.5	89.0	26.5	136.6	1244.0
Total	Count	195	321	250	307	591	398	185	55	284	2586	
	Expected Count	195.0	321.0	250.0	307.0	591.0	398.0	185.0	55.0	284.0	2586.0	

Have a look at the output. There are inevitably going to be differences between the expected and observed frequencies in the real world. Our question is, "are they significant or just attributable to chance?" For example in the School of Education (EDS) we would expect 101 males and 94 females, rather than the 54 males and 141 females we have seen in the data.



To answer this question we can conduct a Chi-square test to test if the data really does support our feeling that the ratio of males to females differs from school to school.

A null hypothesis H_0 for this would be that "*there is no difference in the representation of the sexes across the schools within Sheffield Hallam University*"

An alternative hypothesis H_1 for this would be that "*there is a difference in the representation of the sexes across the schools within Sheffield Hallam University*"

You could probably write the hypotheses more eloquently, feel free to have a go! but we get the idea.

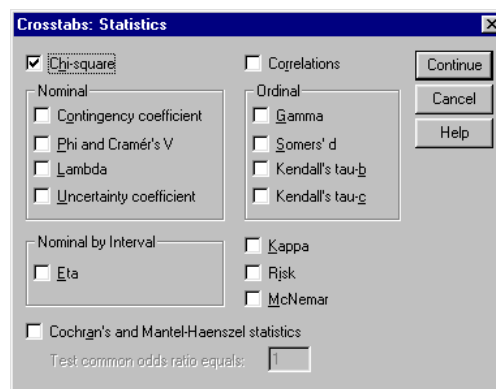
Get back to the **Crosstabs** dialog box.



Click the **Statistics** button, then select "**Chi-square**".

Click **Continue** then **OK**.

The output now has another table, this one gives us a significance value.



Look at the Pearson Chi-Square Sig. figure.

In this case it's .000 - pretty low.

How would we write this down?

We could write something like, "the Chi-Square test carried out on the data was significant at the 0.001 level (2-tailed $p < 0.0005$) of significance. ($\chi^2 = 635.56$, $df = 8$) so we conclude that there is a significant difference in the representation of the sexes across the schools", i.e. it is unlikely that the variables are independent. In this case I think we knew anyway!

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	635.561 ^a	8	.000
Likelihood Ratio	709.332	8	.000
Linear-by-Linear Association	106.694	1	.000
N of Valid Cases	2586		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 26.46.

Example What about age and gender? Are the sexes equally distributed across the age groups? My feeling is that males are more prevalent in the younger age group and females are more highly represented in the older age groups. Am I right?

Construct the test properly, write your hypotheses first then test them. Check in the student companion when you've finished.



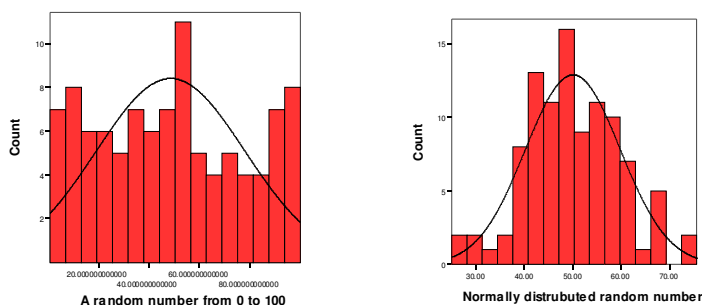
Three methods to test whether data is normally distributed.

Open the file "tests for normality.sav"

1 Draw a histogram of the data and get SPSS to superimpose a normal curve.

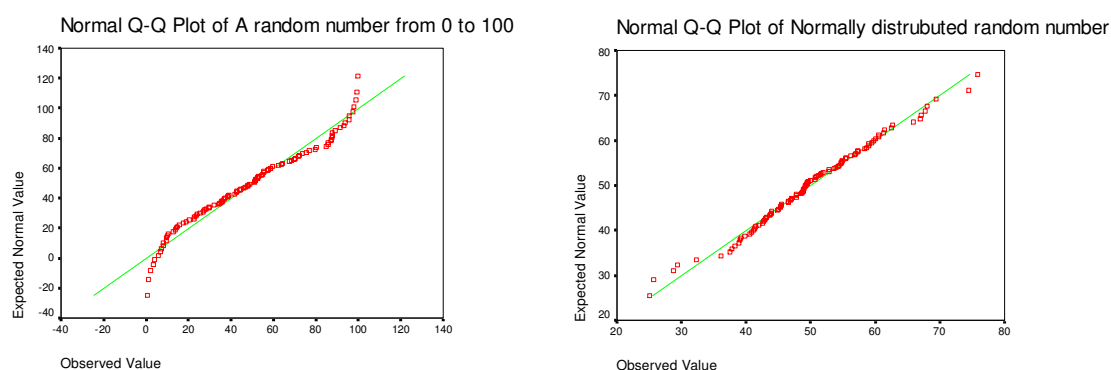
Use **Graphs, Interactive, Histogram**, then assign the variable you are investigating to the horizontal axis, leave the vertical as “count”, then click on the histogram tab at the top of the dialog box and select “normal curve” then click **OK**.

You are unlikely to see perfect bell shapes with smaller sample sizes, to help you compare, the two printed here are both sets of 100 random numbers, on the left they are generated with equal probabilities of being any number between 0 and 100 but for the numbers on the right they were generated with a normal distribution.



2 Checking for normality with a Q-Q plot

You can graphically compare a sample to a normal distribution with the Q-Q plot. In the Q-Q plot the normal distribution is represented by a straight line (the bell shape is squashed flat), your data is plotted round it. Data points from a normal distribution would appear close to the line. Q-Q plot is under the Graphs menu – they are also offered in the “Explore” facility.



Observed values of a single numeric variable are plotted against the expected values if the sample were from a normal distribution. If the sample is from a normal distribution, points will cluster around a straight line.



3 A method of testing if a variable contains normally distributed data.

Yes, you guessed it, there is a test for normality – a test to see what test you can use!

From the menus select **Analyze, Descriptive Statistics, Explore...**

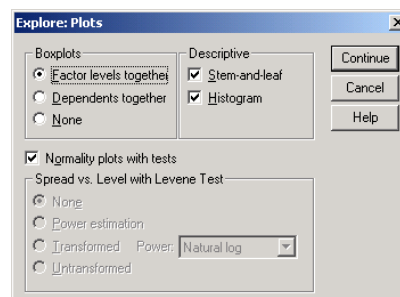
Put the variables you want to check for normality in **Dependant list** box.

Click on the **Plots** button.

Click to select **Normality plots with tests**.

You may also want to see a histogram of the data – if so click to select one.

The results of the Kolmogorov-Smirnov test are similar to other tests, the important figure is the significance. If the figure under the Sig. column is more than 0.05 then the data appear to be from a normal distribution. i.e. they are not significantly different to a normally distributed set of figures.



If the size of the sample is small, the Shapiro-Wilk test is more appropriate. You will see in my example below that Shapiro-Wilk is less inclined to sit on the fence! (The non-normally distributed set of figures "random number 1-100" has a p-value well below 0.05, i.e. it is significantly different to a normally distributed variable.)

Tests of Normality

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
A random number from 0 to 100	.083	100	.085	.953	100	.001
Normally distributed random number	.057	100	.200*	.990	100	.650

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

You can also do the one sample Kolmogorov-Smirnov as a non-parametric test under the analyze menu, if you do this you may see different values in the output. I am not totally convinced, but I think this is because the version in the "Explore" command includes the Lilliefors Significance Correction. (now we know why they say "lies, damn lies and statistics"!)



Some examples to get your teeth into.

1 Are we getting taller?

The file “**Combined heights.sav**” contains the heights of males from the 19th and 20th centuries.

Your mission is to find out if men are taller in the late 20th century are taller than men were in the 19th century.

What assumptions are you making about the data? How it was gathered and pre-processed.

What are your hypotheses?

What test(s) will you use?

Are there any graphical methods that might be useful?

Have a look at the file “**male heights 1883 and 1990.sav**”, it holds similar data. Run your test(s) again, are the results the same. If there are differences what are they and why?

2 Is the staff-student ratio in universities linked to teaching quality?

The file “**The Sunday Times 2001 university league table.sav**” holds data about English universities on 2001. Is the quality of teaching connected in any way to the staff-student ratio?

What assumptions are you making about the data? How it was gathered and pre-processed.

What are your hypotheses?

What test(s) will you use?

Are there any graphical methods that might be useful?



3 Are females taller than males?

Cast your mind back to the beginning of this document, we were talking about samples and how inferential statistics lets us use random samples to draw inferences about the whole population they represent.

Below are two random samples drawn from male and female populations, the heights are in millimetres.

Are males taller, on average than females?

Female sample; 1560, 1610, 1660, 1640, 1580, 1580, 1670, 1630, 1610, 1660.

Male sample; 1685, 1765, 1675, 1754, 1704, 1695, 1765, 1629, 1786, 1620.

What tests will you use? – How will you decide?

How will you structure the data when you type it into SPSS?

What will your null and alternative hypotheses be?

What level of significance will you accept?

4 Radiologist left hand dose.

The file “Radiologist dose with and without lead combined.sav” has in it data gathered to assess the effect of a lead screen to reduce the radiation dose to Radiologists hands while carrying out procedures on patients being irradiated.

In the trials the lead screen was placed between the patient and the radiologist, the intended effect was to reduce the radiation dose to the radiologist, however there were fears that working through the screen would lengthen the procedure.

If there is a 1 in the *screen* variable it means the procedure was carried out with the screen in place, if not the value is 0.

Test this data, did the shield work in significantly reducing left hand dose? What about the right hand?

Did it take longer to examine patients with the shield in place?

Did it take longer to examine heavier patients? Are there any problems in using this data to look for such a trend?



Some ideas about the examples to get your teeth into. (NOTE: In these examples we could question whether our samples are randomly drawn from the population we wish to infer our findings to. It is worth remembering that inferential statistics assume the sample is randomly drawn from the population.)

1 Are we getting taller?

What assumptions are you making about the data? How it was gathered and pre-processed. *The data is old and we can't be sure how it was collected, the 19th century group may have been from a different area than the 20th century males.*

What are your hypotheses?

Null, H_0 , *there is no significant difference between the heights of males in 19th and 20th centuries.*

Alternative, H_1 , *there is a significant difference between the heights of males in 19th and 20th centuries. (this would be 2-tailed)*

What test(s) will you use? - *I did the explore technique, the Q-Q plots and the Kolmogorov-Smirnov and Shapiro-Wilk statistics support the hypothesis that the data is normal, therefore used and independent samples t-test. Remember there are 2 groups here so put the year into the factor list otherwise the distribution of all the heights will be used not the distribution of heights from each century.*

What level of significance will you accept? *For this example 0.05 would do for me.*

Are there any graphical methods that might be useful? *Histogram to see the distribution and maybe boxplots to see level.*

Have a look at the file “**male heights 1883 and 1990.sav**”, it holds similar data. Run your test(s) again, are the results the same. If there are differences what are they and why? *More data will lead to lower p-values if there is a difference.*

2 Is the staff-student ratio in universities linked to teaching quality?

The file “**The Sunday Times 2001 university league table.sav**” holds data about English universities on 2001. Is the quality of teaching connected in any way to the staff-student ratio?

What assumptions are you making about the data? How it was gathered and pre-processed. *We can test the data for normality and choose out test based on that.*

What are your hypotheses?

Null, H_0 , *there is no significant correlation between staff-student ratio in universities linked to teaching quality.*

Alternative, H_1 , *there is a significant correlation between staff-student ratio in universities linked to teaching quality.*

What test(s) will you use? *its going to have to be one of the correlation tests, the non-parametric one is Spearman's.*

Are there any graphical methods that might be useful? *A scatterplot.*



3 Are females taller than males?

What tests will you use? – How will you decide? - *I did the explore technique, then the Q-Q plots and the Kolmogorov-Smirnov and Shapiro-Wilk statistics support the hypothesis that the data is normal, therefore used and independent samples t-test.*

How will you structure the data when you type it into SPSS? *Look in the file “male and female height samples.sav” to see how I did it.*

What will your null and alternative hypotheses be? *Null would state that there was no significant difference between male and female heights; Alternative would say that males are higher. This would be a one-tailed test. You could do a 2-tailed test and simply say male and female heights are different. It depends largely on how much you feel you know about the species you are studying.*

What level of significance will you accept? *For this example 0.05 would do for me.*

4 Radiologist left hand dose.

The file “Radiologist dose with and without lead combined.sav” has in it data gathered to assess the effect of a lead screen to reduce the radiation dose to

Radiologists hands while carrying out procedures on patients being irradiated.

In the trials the lead screen was placed between the patient and the radiologist, the intended effect was to reduce the radiation dose to the radiologist; however there were fears that working through the screen would lengthen the procedure.

If there is a 1 in the *screen* variable it means the procedure was carried out with the screen in place, if not the value is 0.

Test this data, did the shield work in significantly reducing left hand dose? *The Q-Q plot etc told me the data were not normal, so I used a Mann-Whitney Test (non-parametric, independent samples). I used a two tailed hypothesis, I wasn't sure whether the dose would increase or decrease. I found no significant difference.*

What about the right hand? *Using the same test I was surprised to find a significant difference.*

Did it take longer to examine heavier patients? Are there any problems in using this data to look for such a trend? *There appeared a significant trend that heavier patients took longer to examine, however the dataset covers several different types of examination, it could be that the lighter patients were undergoing quicker examinations.*



Part 3 ANOVA and all that.

Analysis of Variance - one-way ANOVA

An experimenter is interested in evaluating the effectiveness of three methods of teaching a given course. A group of 24 subjects is available to the experimenter. This group is considered by the experimenter to be the equivalent of a random sample from the population of interest. Three subgroups of eight subjects each are formed at random; the subgroups are then taught by one of the three methods. Upon completion of the course, each of the subgroups is given a common test (exam) covering the material in the course. The resulting test scores are given in the following table.

Note: You may see the term "Factor" used in some texts, it generally means a variable consisting of one or more levels (treatments) thought to be a cause of variation in a dependent variable. So in our example we are saying that the method of teaching is the factor causing any differences between the groups.

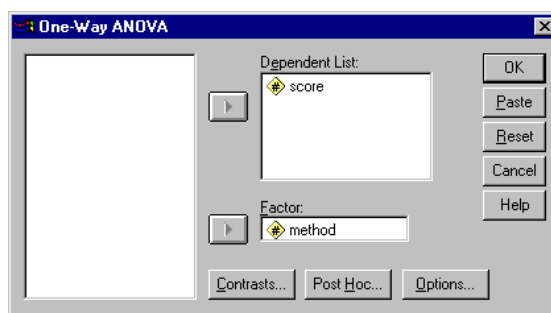
Method 1	Method 2	Method 3
3	4	6
5	4	7
2	3	8
4	8	6
8	7	7
4	4	9
3	2	10
9	5	9

SPSS would prefer the data to have a discriminatory variable:

Score	Method
3	1
5	1
2	1
4	1
8	1
4	1
3	1
9	1
4	2
4	2
3	2
8	2
7	2
4	2
2	2
5	2
6	3
7	3
8	3
6	3
7	3
9	3
10	3
9	3

The data is in the file "anova one way example.sav"

Choose **Analyse, Compare Means, One-Way ANOVA...**



ANOVA

SCORE	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	50.083	2	25.042	6.053	.008
Within Groups	86.875	21	4.137		
Total	136.958	23			

There is a significant difference (0.008) among the three methods of teaching. So it is appropriate to proceed to a posthoc (*a posteriori*) test.



Our test shows there is a significant difference among the three methods of teaching. We therefore proceed to a posthoc (*a posteriori*) test to find out where the difference is.

To do this return to the "One Way ANOVA" dialog box. It should be how you left it.

Choose **Post Hoc...**

Choose **Tukey**.

Post Hoc Tests

Multiple Comparisons

Dependent Variable: SCORE
Tukey HSD

(Selecting a post test is not simple; generally, to compare groups with each other choose the **Tukey** test.)

(I) METHOD	(J) METHOD	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	.13	1.017	.992	-2.44	2.69
	3	-3.00*	1.017	.020	-5.56	-.44
2	1	-.13	1.017	.992	-2.69	2.44
	3	-3.13*	1.017	.015	-5.69	-.56
3	1	3.00*	1.017	.020	.44	5.56
	2	3.13*	1.017	.015	.56	5.69

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

Using the **Tukey** test, we can conclude that method three is the most effective method of teaching. SPSS has put a * by the significant differences and this shows method 3 was significantly different to the other two methods. I've put a ring round the actual Significance figures.

The example above is based on information from the Faculty of Social Science, University of Western Ontario.

It might also be worth noting that this example could well have been done using a non-parametric version of ANOVA, e.g. the Kruskal-Wallis one-way non-parametric ANOVA. You could try this (see below) and see if the result is different. You'll get a p-value = 0.018 (Asymp. Sig.), notice that the nonparametric test isn't quite as sure about the p-value, it is a bit bigger but still significant, this is the price we pay for using the more robust nonparametric test. The test has less statistical power - more on this below.

There are plenty of useable internet resources that will help you advance your repertoire of statistical techniques, the example above is based on one of them, so your knowledge doesn't need to be just a subset of what this document contains! For example the summary below is from a document available on the web "SPSS Version 9 for Windows - an introduction." By Dr Hugh D Jones.

Basically, ANOVA answers the question "Is there a significant difference between the samples (is any one different from the others)?" If there is not (Sig. >0.05) then there is no need to go any



further. If there is then you might want to know which sample(s) is different from each other. A supplementary (Post-hoc) test is carried out to investigate differences between the samples.

An example of a nonparametric one way ANOVA (Kruskal-Wallis)

The data above used in the one way ANOVA example are scores not "real" measurements. We have used a test to try to "measure" some sort of ability, however this isn't quite the same as measuring someone's height or weight, and we can't measure it directly and so are using the score from a test to estimate it. The issue for us as statisticians (yes I'm sorry, but if you've got this far you are) is to decide if the resulting score, in this case the exam score, is OK to investigate using a parametric method. The danger is that if we assume the data are parametric and we are wrong then we can get erroneous results, possibly a false positive result, (i.e. a type one statistical error).

So what do we do?

Read the appendix about parametric and nonparametric tests. One method of deciding is to check the data for normality, parametric tests expect to see data that are reasonably close to normally distributed., if they are not, e.g. if they contain outliers, and especially if the sample is small, this can cause problems with their maths and give incorrect results.

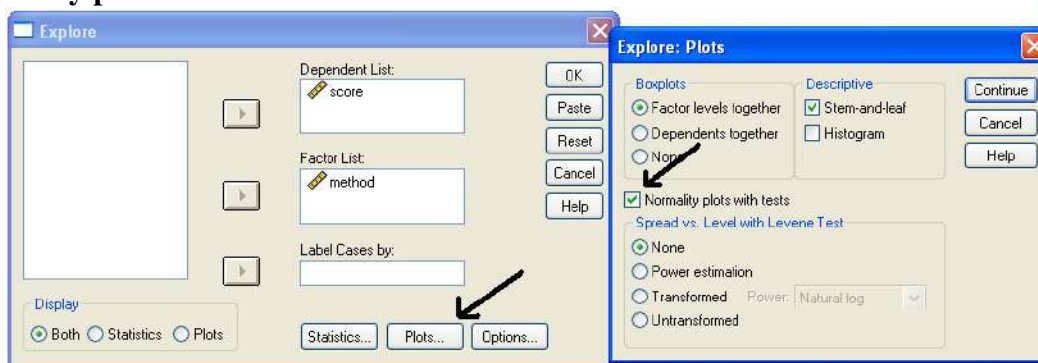
SPSS lets you test data for normality using a variety of methods (see the section "Three methods to test whether data is normally distributed.") we will use the test for normality to see if our data are normally distributed. Here's how...

Open the file we used earlier, "**anova one way example.sav**"

These data are really three different sets of scores, one set for each group, so when we test them for normality we need to remember this, if we treat them as one group then any differences between the groups might lead us to think the data aren't normally distributed when the data from each group is. It is the normality of each group that matters.

To test the data for normality chose **Analyze, Descriptive Statistics, Explore...**

Put the "score" variable in the **Dependant list** box. Click on the **Plots** button. Click to select **Normality plots with tests**.





The output gives p-values (in the Sig. column in the output) that will be less than 0.05 if the data you fed in are significantly different to normally distributed data.

Tests of Normality

method	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
score 1	.243	8	.181	.874	8	.166
2	.248	8	.159	.922	8	.450
3	.193	8	.200*	.920	8	.428

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Both the Shapiro-Wilk and the Kolmogorov-Smirnov tests offer an opinion on the normality of the data. There are various caveats that should accompany such tests but they do offer a consistent guide, a Google search for the two tests will bring up plenty of information about the issues around normality testing but for us the tests offer useful guidance in the parametric/nonparametric decision when we add in the other sensible advice which is to "make as few assumptions as possible", i.e. err on the safe side.

If the p-value in the Sig. column is below 0.05 we should play safe and opt for non parametric tests. In our case with three sub sets of data, any of the three Sig. values being below 0.05 would stop us applying the parametric ANOVA to the data.

In our case the data are not significantly different from normal. However just to show you the method we can apply the nonparametric version of one way ANOVA to these data; this isn't a bad thing to do, just a little over cautious in this case. It is preferable to using a parametric test when it should be a nonparametric test.

The Kruskal-Wallis test (one-way non-parametric ANOVA).

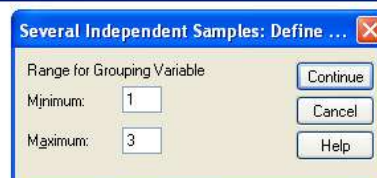
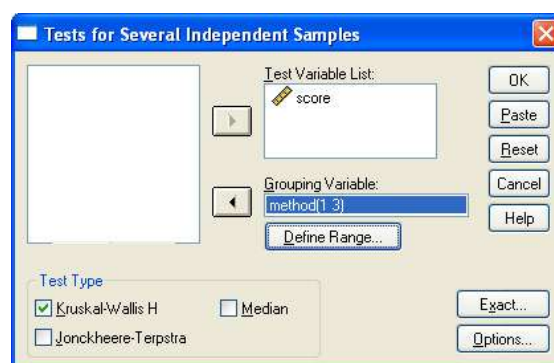
Choose, Analyse, Non-parametric tests, Independent samples. Put the "score" variable in the test variable list and the method as the grouping variable. Before the test can work you need to define the groups, in our example they are numbered 1 to 3. The table below shows the relevant output.

Test Statistics(a,b)

	score
Chi-Square	8.077
df	2
Asymp. Sig.	.018

a Kruskal Wallis Test

b Grouping Variable: method



Notice that the nonparametric test still says that there is a significant difference between the groups ($p=0.018$) however it isn't quite as well convinced as the more sensitive ANOVA. This is a good illustration of the minor penalty that you pay for the more rugged



nonparametric tests, they are less likely to catch a small effect that does exist, i.e. they are less powerful.

So to recap; generally scores would be better treated by nonparametric methods. In this example we did find them to be normally distributed and used them as an example in applying a one way ANOVA and its nonparametric equivalent, the Kruskal-Wallis test. Finally, the two tests agreed but we noticed a slight difference in how certain they were.

Analysis of Variance - Repeated measures ANOVA.

One group of subjects walked in three conditions; no crutches, elbow crutches and axillary crutches, the energy used was measured (indirectly by looking at the oxygen used).

Subject	NormalKj100m	AxillaryKj100m	ElbowKj100m
1	0.89164	1.26126	0.92267
2	0.49598	0.65153	0.75821
3	0.3352	0.39911	0.31486
4	0.50821	0.36173	0.30608
5	0.68024	0.5537	0.6027
6	0.34827	0.43575	0.63945
7	0.10649	0.09443	0.1274
8	0.33581	0.38588	0.42368
9	0.38625	0.41426	0.28196
10	0.27579	0.1876	0.343

You can think of the repeated measures ANOVA as an extension of the paired t-test, rather like the One-way ANOVA is an extended version of the independent samples t-test.

Notice that there is no discriminatory variable in this data set, this is because there is only one group, each person was measured in all three conditions.

Q. How would we tackle analysing these data?

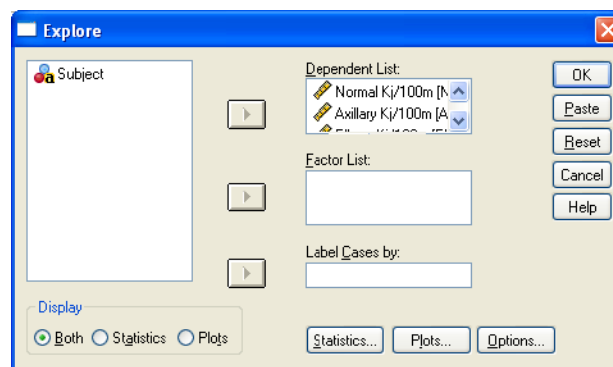
A. Pretty well the same as any other;

1. Create some simple descriptive stats (mean S.D. etc.).
2. Check the data for normality to check it is sensible to run parametric tests.
3. Run the test (Repeated measure ANOVA or nonparametric Friedman test) and interpret the result.

You can use the Explore command for the descriptives, choose Analyse, Descriptive Stats, Explore

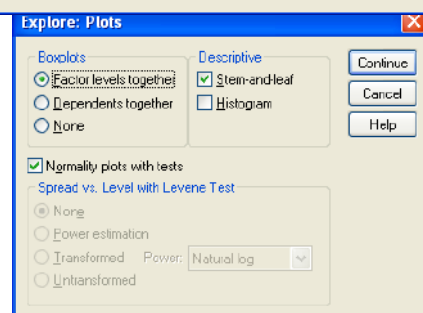


Scot the three variables containing the energies into the "Dependant list" box, then hit the "Plots..." button and tick the option for "Normality plots and tests". Click Continue then OK and the output should appear. You should see a table of stats giving such things as Mean, Median, Standard deviation (SD) etc. Just ignore the ones you don't understand.



The table giving the results of the "Tests of Normality" is telling us in this case that there is some doubt about the normality of the Axillary variable.

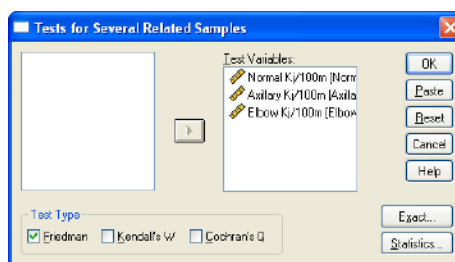
The Shapiro-Wilk test suggests there is a significant difference between the distribution of the variable called "Axillary Kj/100m" and a normal distribution. (Sig. (p) = 0.031) We should therefore consider not using parametric methods on this variable but use non parametric tests.



	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Normal Kj/100m	.190	10	.200(*)	.935	10	.495
Axillary Kj/100m	.248	10	.081	.827	10	.031
Elbow Kj/100m	.198	10	.200(*)	.940	10	.550

A nonparametric repeated measures ANOVA - the Friedman test.

To run the test Choose Analyse, Nonparamtric tests, K-related samples and put all the three variables representing the three repeated measures into the "Test Variable" box. Then click OK. The result tells us that in this case there is no significant



Test Statistics ^a	
N	10
Chi-Square	.800
df	2
Asymp. Sig.	.670

a. Friedman Test

difference detected between the conditions ($p = 0.670$ which is larger than 0.05). In reality we can only say we failed to detect a significant difference it could be that there is a difference but it's so small that this experiment wasn't powerful enough to detect it.

A parametric repeated measures ANOVA. (One-factor within subjects ANOVA)

If the data are suitable you can use the more sensitive repeated measures ANOVA. The following data shows the results of an experiment where subjects jumped three times. These are repeated measures. Each subject jumped three times, the height was recorded, the column labelled Jump1 has each subjects first jump in it, the column labelled Jump2 has each subjects second jump in it and so on.



This is a typical layout for paired data, paired data doesn't mean only two columns, we can have three or more columns of paired data.

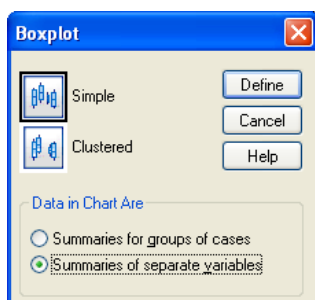
Subject	Jump1	Jump2	Jump3
1.00	6.00	43.90	45.70
2.00	7.00	48.00	46.20
3.00	8.10	52.70	51.60
4.00	4.10	47.80	48.30
5.00	2.10	50.50	50.40
6.00	5.10	55.60	58.50
7.00	6.40	47.10	51.30
8.00	7.90	50.80	52.30
9.00	8.70	40.20	40.60
10.00	4.00	34.00	31.20
11.00	4.90	37.90	39.50
12.00	4.70	57.00	57.50
13.00	0.60	45.40	44.40
14.00	7.80	39.00	42.00
15.00	4.20	45.20	45.80
16.00	5.10	44.70	45.60
17.00	2.90	44.90	45.90
18.00	0.20	52.30	53.40
19.00	7.40	33.10	33.70

Before applying the ANOVA to these data we can first usefully get some descriptive stats from them and check for normality.

The simple way to do this is use the Explore command (Analyse, Descriptive Stats, Explore). Put the three variables containing the Jump into the "Dependant list" box, then hit the "Plots..." button and tick the option for "Normality plots and tests". Click Continue then OK and the output should appear.

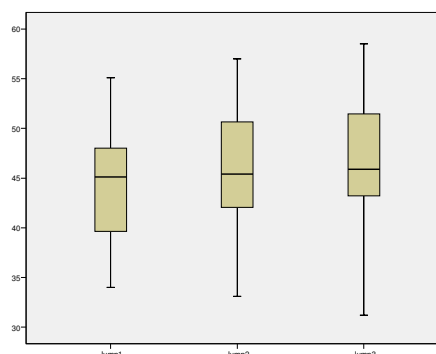
Means, Standard Deviations etc. should appear and the normality tests. This time the Sig. figures (p-values) for both the Shapiro-Wilk and the Kolmogorov-Smirnov tests are well above 0.05 and we can conclude that the three variables are not significantly different to normally distributed data.

You might notice that the boxplots supplied with the explore command are on separate graphs, this doesn't aid comparison, you can get all three on one

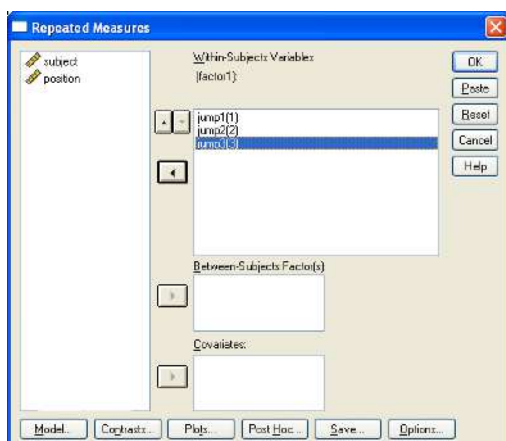


easily by using the graphs menu. Choose Graphs, (then Legacy Dialogs if on a late version) then Boxplot. Choose the option for "Summaries of separate variables" for these paired data then click Define. Put the three Jump variables in the "boxes represent" area and hit OK. You should get

a pretty boxplot to illustrate your research! (The labels could be better, you can add more detail by double clicking the graph while in SPSS to edit it.)



The results of the normality tests tell us that in this case it is safe to apply the parametric repeated measures ANOVA. Life doesn't get much better than this!



Choose, Analyse, General Linear Model, Repeated Measures. The "Repeated measure; define factors" dialog should appear. Put 3 in the number of levels box because we have measured our victims 3 times. Click the **Add** button. The text "factor1(3)" should appear in the larger box. Now click **Define**. The next step seems complex but isn't, highlight the three Jump variables and send them into the box with the question marks in. Click OK to see the output.



Making sense of the repeated measures ANOVA output.

Lots of tables will be presented to you as a result of following the instructions above. Two are of importance; Mauchly's Test of Sphericity and Tests of Within-Subjects Effects. The first one, Mauchly's Test of Sphericity, tells us which line of the second one to read.

Mauchly's Test of Sphericity^b

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon ^a		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
factor1	.767	4.512	2	.105	.811	.880	.500

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

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
factor1	Sphericity Assumed	36.171	2	18.085	7.233	.002
	Greenhouse-Geisser	36.171	1.622	22.302	7.233	.005
	Huynh-Feldt	36.171	1.759	20.558	7.233	.004
	Lower-bound	36.171	1.000	36.171	7.233	.015
Error(factor1)	Sphericity Assumed	90.016	36	2.500		
	Greenhouse-Geisser	90.016	29.194	3.083		
	Huynh-Feldt	90.016	31.670	2.842		
	Lower-bound	90.016	18.000	5.001		

In the tables above, Mauchly's Test of Sphericity is giving us a Sig. figure of .105 this is greater than 0.05 and so we can assume sphericity. Don't worry too much about what sphericity is simply use it to decide which line in the next table gives the p-value we want. Sphericity is about the variability of the data we are using, the test for non spherical variables are conducted with slightly different maths so it is important we read the correct line in the output dependant on whether the data analysed are spherical or not. In our example we can read the p-value from the top of the table called "Tests of Within-Subjects Effects" i.e. the line that assumes sphericity, giving a p-value of 0.002. We can therefore say that there is a significant difference between the three columns of figures and therefore between the height of jump attained in at least one attempt and at least one other. If the Sig. figure for Mauchly's test had been below 0.05 we would have read the p-value from the second row of the second table, i.e. the row labelled "Greenhouse-Geisser", this would have been 0.005 in this case.

Post Hoc tests for repeated measures ANOVA.

Where do the significant differences really lie? To answer this we can apply a post hoc test. Dienes (2001) states that "you can perform a post hoc test called "Fisher's protected t" easily enough. This just means you use repeated-measures (i.e. paired) t-tests to see which pairs of levels are significantly different (go to "Analyze", "Compare Means", and then "Paired-Samples T Test"). However, you only perform the t-tests if overall significance is found. Further, only use this procedure if you have no more than three levels." So applying paired t-tests shows us that jump1 is significantly different to jumps 2&3 but jumps 2&3 don't seem to be significantly different from each other.

Pair	P-value
Jump1 - Jump2	.033
Jump2 - Jump3	.076
Jump1 - Jump3	.005

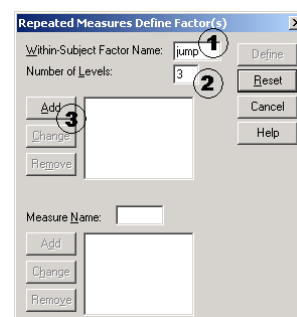


Mixed designs.

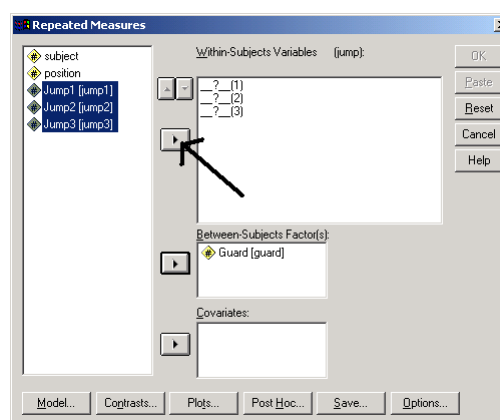
We can think of the repeated measures ANOVA as an extension of the paired t-test, and the One-way ANOVA as an extended version of the independent samples t-test. A mixed design is when we have for example two groups of subjects measured repeatedly, e.g. two treatments, each treatment group being measured before and a couple of times after treatment. An excellent example of this technique is covered in Kinnear and Gray's book "SPSS for Windows Made Simple" under the chapter "Experiments of Mixed Design"

The file "*Three jumps.sav*" has the three jumps data in it, these data were gathered for an investigation into the jump height of players in various positions in basketball, the players are split into Guard, Forward and Centre, however for our example we are going to use broader categories of Guard and non-Guard. This means we have both repeated measures, and independent groups. I.e. our repeated measures (jump height, measured three times for everyone) were taken within two groups (players who were guards and those in other positions).

The method for analysing this is similar to the repeated measures ANOVA we saw earlier. Choose Analyse, General Linear Model, then Repeated Measures From the menu. First - put in a term to describe the within subjects effect (our repeated measures of jump – I put the word “jump” in). Second - how many measures (levels) are there?, we have three. Third – press the “Add” button. The text “jump(3)” or similar should appear in the upper white box.



Hit the “Define” button and then in the next dialog transfer the three variables representing our repeated measures to the upper square, the blanks in the box should change to text like “jump1(1)” etc. Finally get the between subjects variable into the lower box. Take a big breath and hit OK.



The output should look incredibly confusing, if it doesn't then consider a career in astrophysics. To appreciate it we have to realise that SPSS spends most of its time locked away on the hard disk so when it gets to display something on the screen it's got lots to say. We'll ignore much of it. The first table (Within-Subjects Factors) starts simply and tells us we had three within-subjects factors, Jump1, jump2 and jump3. The next table (Between-Subjects Factors) tells us our between-subjects factor is whether they were a guard or not, it recons we have 11 guards and 8 non-guards. Ignore the third table (Multivariate Tests) – apologies if a relative of yours designed any of these. The next table, “Mauchly's Test of Sphericity” is relevant, it lets us know which figure to read in the fifth table. In the tables above, Mauchly's Test of Sphericity is giving us a Sig. figure of .158 this is greater than 0.05 and so we can assume sphericity. This means that in the next table, “Tests of Within-Subjects Effects” we can read the rows labelled “Sphericity Assumed”, if the Sig. figure was below 0.05, e.g. 0.023 we would read the next row, labelled “Greenhouse-Geisser”.



From this table we read two p-values from the Sig. column on the right, 0.004 and 0.070, (I hope at this point you appreciate me finding an example with plenty of different p-values so you know where I'm looking!)

This tells us that the factor “jump” i.e. whether it was their 1st 2nd or 3rd jump had a significant effect on the jump height (p=0.004) but the interaction “jump * guard” is not significant at the 5% level (our p-value isn't below 0.05).

Measure: MEASURE_1						
Source		Type III Sum of Squares	df	Mean Square	F	Sig.
jump	Sphericity Assumed	28.977	2	14.488	6.397	.004
	Greenhouse-Geisser	28.977	1.659	17.470	6.397	.008
	Huynh-Feldt	28.977	1.924	15.062	6.397	.005
	Lower-bound	28.977	1.000	28.977	6.397	.022
jump * guard	Sphericity Assumed	13.008	2	6.504	2.872	.070
	Greenhouse-Geisser	13.008	1.659	7.842	2.872	.082
	Huynh-Feldt	13.008	1.924	6.761	2.872	.073
	Lower-bound	13.008	1.000	13.008	2.872	.108
Error(jump)	Sphericity Assumed	77.008	34	2.265		
	Greenhouse-Geisser	77.008	28.197	2.731		
	Huynh-Feldt	77.008	32.706	2.355		
	Lower-bound	77.008	17.000	4.530		

Kinnear and Gray (1997) have a good example of this technique, our example here gives a potted view, if you plan to use it seriously do follow up their example and explanation, it's from proper statisticians!



Part 4, Reliability and sensitivity

The two issues don't necessarily have to sit together but were useful to put together to form a short section in their own right rather than be lost in the appendices.

They started life as individual notes and references at the end have been left for convenience as well as added to the full document list.

We can though see a link between the concepts, for example if rather than looking at the inter rater reliability (e.g. do two different people agree?) we were comparing a diagnosis to a clinical finding upon for example an operation (i.e. did we find what we thought was there?) similar techniques might be employed.

Also it is worth checking the definitions of inter-rater reliability and intra-rater reliability.

Inter-rater reliability deals with the issue of reliability between different people (raters).

Intra-rater reliability deals with whether one rater is consistent, i.e. when they re-look at the same subjects do they rate them in a similar way again.



Inter-Rater Agreement using the Intraclass Correlation Coefficient.

Imagine that a student wants to find out if a certain exercise can improve performance. To measure performance they decide to use a simple measured jump. However to be sure that he can sensibly repeat the measures after the exercise regime has been completed he wants to estimate the reliability of his measurement method. Although we are looking at the measurements from two attempts at the task (jump length in cm.) for each of six research subjects, it is a similar task to two judges each assessing each of six items. The assumption we make here (and its not a bad one) is that if we can re-measure the same thing reliably then when we re-measure after the treatment any differences we see are likely to be due to the treatment not just variation in the measurements.

Jump1	Jump2
157.2	170.1
179.0	169.3
168.7	180.2
154.2	152.5
99.2	104.5
108.4	115.3

How does our intrepid student assess the reliability of measurement?

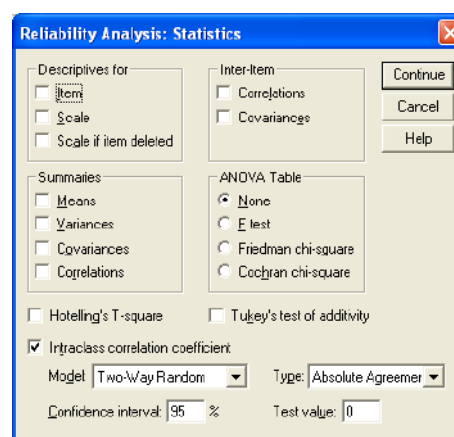
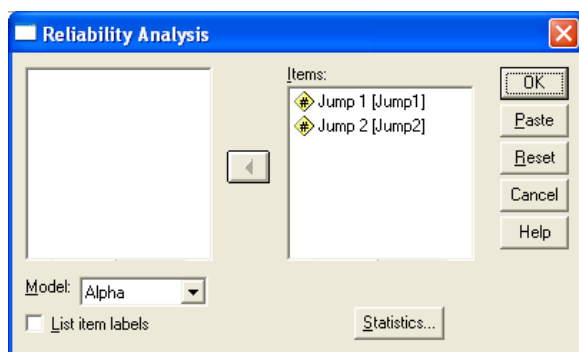
He could, if he had just two columns, simply look for correlation between the Jumps; this however would not be an ideal method. Look at the table on the right, of six subjects each performing two jumps, each second measurement is ten times the first, the

<i>Jump1</i>	<i>Jump2</i>
<i>157.2</i>	<i>1572.0</i>
<i>179.0</i>	<i>1790.0</i>
<i>168.7</i>	<i>1687.0</i>
<i>154.2</i>	<i>1542.0</i>
<i>99.2</i>	<i>992.0</i>
<i>108.4</i>	<i>1084.0</i>

correlation coefficient would be one (Correlation Coefficient=1) suggesting massive correlation, which there is, the two columns increase and decrease together. However if these were your measurements you'd struggle to say you could reliably repeat the measurements and get the same answer!

To get round this problem we can use the Intraclass Correlation Coefficient (ICC).

SPSS can calculate it for you. Choose; Analyze, Scale, Reliability analysis from the menus. Put the relevant variables into the Items box then Click Statistics, (see the pictures below of how the menus look). Tick the box for the Intraclass correlation coefficient, choose the Two-Way Random model and the Type as Absolute Agreement. Click Continue then OK. A shed load of output should appear but as usual in SPSS you only need a few bits.





Intraclass Correlation Coefficient

	Intraclass Correlation ^a	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.962 ^b	.791	.994	55.606	5	5	.000
Average Measures	.981	.883	.997	55.606	5	5	.000

Two-way random effects model where both people effects and measures effects are random.

a. Type A intraclass correlation coefficients using an absolute agreement definition.

b. The estimator is the same, whether the interaction effect is present or not.

Interpreting the results.

The Intraclass Correlation Coefficient (ICC) in this case is 0.962 we use the single measures because the figures we fed SPSS were raw measurements not an average of several attempts.

This value, 0.962 shows a considerable amount of agreement!

Cronbach's Alpha.

Cronbach's Alpha is another measure of reliability, and agrees with the ICC here about the high level of reliability between the two measures. (Note that a reliability coefficient of .70 or higher is considered "acceptable" in most Social Science research situations using Cronbach's Alpha.)

Reliability Statistics

Cronbach's Alpha	N of Items
.982	2

The method also works if you have taken more than 2 measures of the same thing.

Jump1	Jump2	Jump3
157.2	170.1	160.3
179.0	169.3	157.7
168.7	180.2	176.4
154.2	152.5	153.2
99.2	104.5	100.3
108.4	115.3	118.2

The data here gave an ICC of 0.961

If you want to check you can use the method, try it with these figures to check you get the same result.

Reference;

Wuensch Karl L. , (2002), Dr. Karl L. Wuensch's SPSS-Data Page, last accessed 27/9/2006 at <http://core.ecu.edu/psyc/wuenschk/SPSS/SPSS-Data.htm>

UCLA Academic Technology Services, SPSS FAQ: What does Cronbach's alpha mean? last accessed 10/11/2006 at; <http://www.ats.ucla.edu/STAT/SPSS/faq/alpha.html>



Inter rater agreement using Kappa.

Do two people agree about a diagnosis?

This section is based on an example from Altman (1995) based on data from Boyd (1982).

The crosstabulation below shows the classification of 85 xeromammograms by 2 radiologists. Each radiologists rated each of the 85 xeromammograms.

Radiologist 1 * Radiologist 2 Crosstabulation

		Radiologist 2				Total
		Normal	Benign	Suspected cancer	Cancer	
Radiologist 1	Normal	21	12	0	0	33
	Benign	4	17	1	0	22
	Suspected cancer	3	9	15	2	29
	Cancer	0	0	0	1	1
Total		28	38	16	3	85

Look at the diagonal from top left down to bottom right in the table, it has the numbers 21, 17, 15, 1. these are the cases where the raters (our radiologists) agree. In fact they agreed in quite a lot of the cases, 54 out of the 85. about 64% of the time. This level of agreement though is slightly false, since if they had both simply guessed the diagnosis (and I am assured they don't!) there would have still been a degree of agreement just due to chance. The Kappa statistic takes this into account, it can be done manually with relatively few maths skills, if you want a go then refer back to Altman's excellent book. If you want to do this the easy way use SPSS.

The data on the left are those used in the example. the numbers 1-4 are codes for the classifications of the xeromammograms as below.

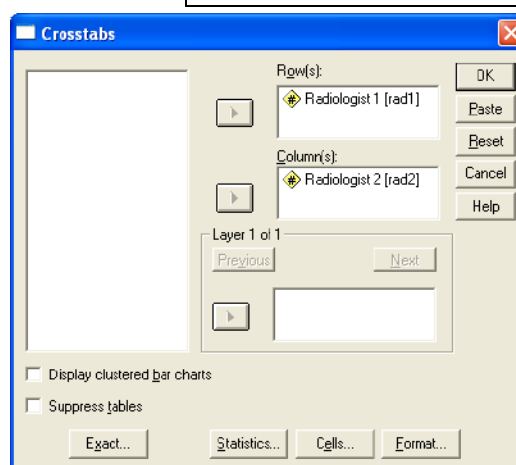
You can copy and paste the data into SPSS if you want to play with the technique.

1	Normal
2	Benign
3	Suspected cancer
4	Cancer

Choose; Analyse, Descriptives, Crosstabs and put the variables over to the rows and columns. Click OK to get the crosstab above. To calculate the Kappa (κ), Click statistics on the crosstabs dialog and choose "Kappa".

The value for Kappa in this case is 0.473 indicating moderate agreement.

Try it yourself; the data are in the file called "Inter rater agreement using Kappa.sav"





Calculating the sensitivity and specificity of a diagnostic test.

The table below is a 2x2 crosstabulation (contingency table) representing the findings of a diagnostic test when compared to the actual disease state. I.e. a comparison of what the test indicated and the real facts. The four cells TP, FP, FN & TN would have in them the number in each category, they will total the number of cases investigated. The cross-tabulate command in SPSS or the pivot table feature in MS Excel can calculate the matrix values.

Test for disease	Actual Disease state	
	<i>Positive</i>	<i>Negative</i>
<i>Positive</i>	True Positive (TP)	False Positive (FP)
<i>Negative</i>	False Negative (FN)	True Negative (TN)

Sensitivity=TP/(TP+FN)

Specificity=TN/(FP+TN)

Prevalence=(TP+FN)/(TP+FN+FP+TN)

Positive Predictive Value =TP/(TP+FP)

Negative Predictive Value =TN/(FN+TN)

Positive Likelihood=SENS/(1-SPEC)

Negative Likelihood=(1-SENS)/SPEC

Overall Accuracy = (TP + TN)/(TP + FP + FN + TN)

A convenient calculator is available at; <http://www.intmed.mcw.edu/clincalc/bayes.html>

Altman and Bland (1994) offer the following definitions;

Sensitivity is the proportion of true positives that are correctly identified by the test.

Specificity is the proportion of true negatives that are correctly identified by the test.

Positive predictive value is the proportion of patients with positive test results who are correctly diagnosed.

Negative predictive value is the proportion of patients with negative test results who are correctly diagnosed.

An example; N = 45+5+15+35 = 100,

	Disease	No Disease
Positive Test	TP: 45	FP: 15
Negative Test	FN: 5	TN: 35

Sensitivity = TP/(TP+FN) = 45/(45+5) = 45/50 = 0.9

Specificity = TN/(FP+TN) = 35/(15+35) = 35/50 = 0.7

sources;

Altman, D. G. and Bland, J. M., (1994), Statistics Notes: Diagnostic tests 1: sensitivity and specificity, *BMJ*;308;1552-

Altman, D. G. and Bland, J. M., (1994), Statistics Notes: Diagnostic tests 2: sensitivity and specificity, *BMJ*; 309;102-



Appendix 1 **Copying information from SPSS to other programs.**

You can copy the raw data from SPSS to MSWord.

Using the mouse, click and hold down the mouse button on the first cell of the data you want to copy then keep the button down and move the mouse up to the last cell. This should select the data you want. You can release the mouse button – the selection should stay black. Next choose **Copy** from the **Edit** menu. Start Microsoft Word. Open the document you want the data to be pasted in to and choose **Paste** from the **Edit** menu in Word.

The data may not be in the format you want. Select the data within the word document, then click on the **Table** menu and select **Convert Text to Table...** check the settings in the dialog box seem appropriate and click **OK**. The data should then appear as a table in word.

Copying results tables into MS Word.

In the SPSS Output Viewer, click on the results you want to copy, e.g. a table of descriptive statistics and choose **Copy** from the **Edit** menu.

Go back to Word (click on the Word button on the Task bar at the bottom of your screen,) and choose **Paste** from the **Edit** menu in Word.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
AGE	10	11.00	67.00	33.3000	19.0441
Valid N (listwise)	10				

This pastes the data in as a table.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
AGE	10	11.00	67.00	33.3000	19.0441
Valid N (listwise)	10				

For a better layout you can choose **Paste Special** from the **Edit** menu in Word and choose to paste as a **picture**, this looks better but is harder to edit and may take up more space when stored on disk.

Copying graphs into MS Word.

In the SPSS Output Viewer, click on the graph and choose **Copy** from the **Edit** menu.

Go back to Word and choose **Paste** from the **Edit** menu in Word.



You may need to use; *Format, Picture, Wrapping, Square* while the picture is selected, to get the text to go round the graph, then size the graph using the drag-handles at the corners. To move the graph in the document just click and hold in its middle and drag it about with the mouse. Remember also there is a cropping tool that lets you make the finished product better.

Saving data in an alternative format, e.g. to MS Excel.

To save the current data editor contents in Excel format, go to the SPSS data editor and choose **Save As** from the **File** menu. Select **Excel (*.xls)** from the save as type drop-down box, type in a new filename and press **Save**. Always check the work by opening it in Excel, make sure all the records have transferred.

Opening an Excel file in SPSS

Make sure Excel doesn't have the file open.

Choose **File, Open**, then where the dialog box says "**Files of type**" click the arrowhead button and choose **Excel** as the file type.

SPSS should then show Excel files in the current folder.

The data in the Excel file needs to be in an appropriate format for SPSS i.e. each row being a case with one single row at the top of the Excel version containing the variable names for SPSS to use. Also SPSS may alter the variable names to fit in with its own regime - details should appear in the output window - check this. Also check that the transfer has worked and the data still makes sense! Check the data and how it is stored by looking at the "Variable view" in SPSS. It can sometimes happen that data is stored as text or string format when it is really numeric, getting this wrong can mean you can't analyse your data, so be sure to check it in variable view.

(NOTE; It is worth noting that a new version of MS Office running under Windows Vista is becoming available, the menu structure in this appears completely different to the standard version. I've not yet had to get to grips with this new beast but am told that holding the Alt key gives access to the original menus for folk who don't want to relearn the product, but I haven't tried it.)



Appendix 2 **More about parametric and nonparametric tests.**

Generally we only use parametric tests on data that we can make assumptions about the normality of the underlying population on. However the issue of sample size can be reason enough to fall back on nonparametric tests.

Typically a non parametric test looks just at the ranks of the scores, this means that it simply looks to see where the scores appear when judged against each other, a bit like the "Top 20" is for singles. It doesn't matter how many records each song sells, the one that sells most is number 1, and it could have sold three more than the number two single, or three million! Because of this the nonparametric tests are more robust, they are less likely to be affected by extreme results.

In a big sample one or two extreme results would make little difference even to a parametric test, that does look at the actual value, rather than ranked order, however when the data set is small (and different statisticians will argue what is small!) one bad reading can cause great harm, consequently it can be safer to drop down to a nonparametric test. We tend to use the term "drop to" because the penalty you pay for using the non parametric test is that they are less sensitive, we term this statistical power, this, put simply, means that a nonparametric test is less likely to detect an effect if it really does exist.

If you do both and the p-values agree then it implies that the data are suitable for using a parametric test, since we could argue that widely differing p-values indicate that we should use the nonparametric test since the nonparametric can be used regardless, however this is hard to justify as a strategy for selecting a test.

If in doubt you can quote the p-value from the non parametric on the basis that this being the less powerful method, if it passes it then it would pass the other anyway! (though this isn't always the case) the other argument is "if in doubt make the least number of assumptions about the data" this also would make the nonparametric the safe bet.

Another reason for using a nonparametric test could be a small sample size, the way the tests work make them less susceptible to "wobbly" data.

Some ways to choose....

- You could use a method that is widely accepted in the literature for dealing with data of your type - this may especially be the case if you are following a method previously used in a different context.
- You can test your data to look for evidence of normality by drawing a histogram and looking for the normal curve shape or doing an inferential test to see if the data are distributed significantly differently from data in a normally distributed variable.
- You could play safe and use non parametric tests anyway. these will do the job but may be less sensitive or powerful, i.e. they may have less chance of detecting an effect if there is one. One argument for adopting this strategy is if the sample is small. Statisticians are unlikely to say what a small sample is - partly because the answer varies depending on how big the effect is that we are looking for and partly because they are just like that.



If I were to offer an opinion about sample size, then numbers under twenty can certainly be called small (some older sources would say under 30), though many effects will show up with such amounts of data, below 10 is small. There are some clever calculations that can be done to estimate a sample size needed to show a given difference but generally they rely on estimating the difference in means likely, the standard deviation of the data the required statistical power and the acceptable p-value. Such calculations, though increasingly popular are I feel beyond the scope of our endeavours at present. A good web resource and resource for further reading has been developed by Russ Lenth. The site is at; <http://www.cs.uiowa.edu/~rlenth/Power/> the extensive notes below the calculator on his page give sound reading for researchers planning to carry out sample size calculations, many of Russ' publications also make this difficult topic accessible to a wider audience of less statistically minded readers.



Appendix 3 Creating a new variable in SPSS based on an existing variable.

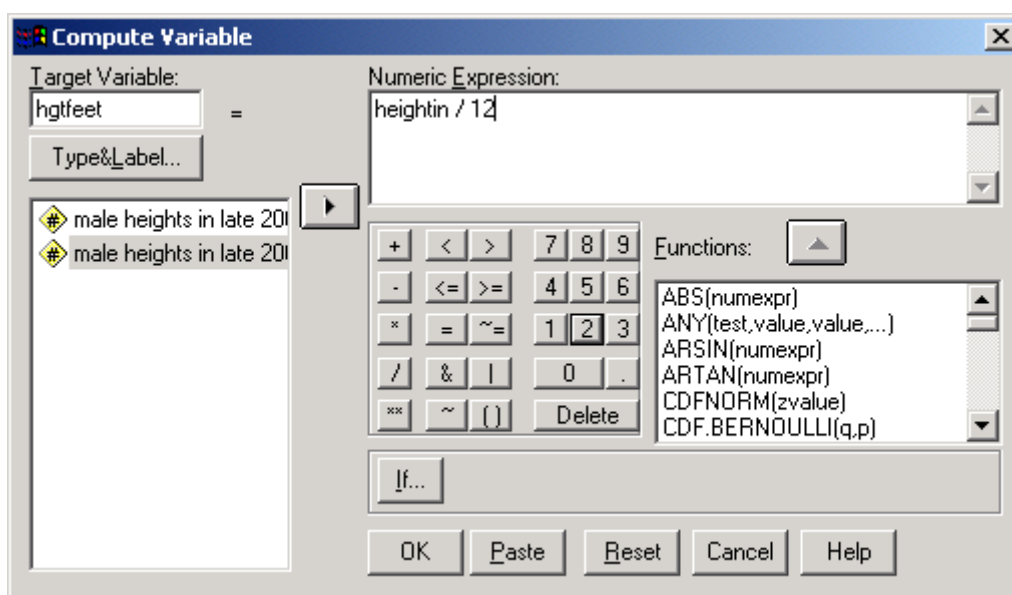
Open the file **malehgts1990s.sav**

The heights are given in MM and inches, lets pretend we want the height in feet.

Use Transform, Compute from the menus, this brings up the "Compute Variable" dialog box.

This is what to do...

1. Type in a new variable name (8 characters max) in the "Target Variable" box.
2. Click on the variable that has the data in you are starting with (*heightin*) and transfer it to the "Numeric Expression" box using the arrow button. 
3. Put the maths round the variable to transform it using the calculator on the dialog box - e.g. divide it by twelve to make it into feet.
4. Click OK and a new variable should appear on the screen. The original data should be unaltered. Check the results to make sure it's what you expected. (In our case it is feet and tenths not feet and inches.)



The built in functions can also be useful. To get information on each one click it with the "other" mouse button.



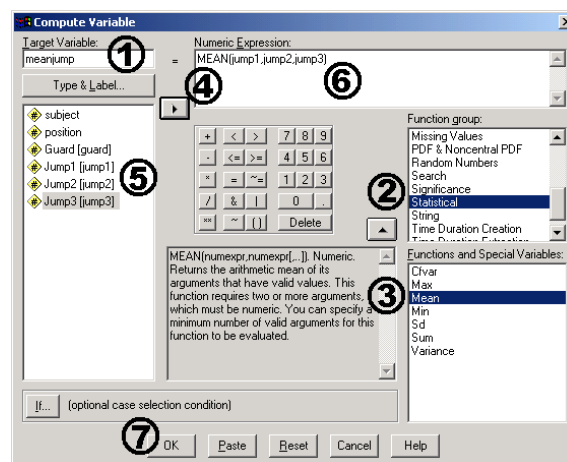
Appendix 4 Built in functions.

Open the file called “*three jumps.sav*”

We are going to calculate the mean jump height for each subject.

It is worth noting that in MS Excel this can be done easily by copying formulae, the concept is different in SPSS but no less effective, if you are an expert in MS Excel you could do calculations in that package and transfer the data over if that seems easier.

To calculate the mean height for each subject choose Transform then Compute from the menus. The dialog illustrated here should appear.



1 Type in a name for the target variable, you might also give it a label, but this can be done later in the variable view screen.

2 Either choose “All” in the function group if you aren’t sure of the subheading or, in this case, scroll down and choose “Statistical” to see a shorter list of stats functions.

3 Choose the function you want, for us it’s the Mean. When you pick it the system offers some help on that function, read it to check it’s going to do what you want.

4 You can double click the function or hit the up-arrow button to transfer it to the “Numeric Expression” box. Instead of putting actual numbers in the function we can just put in the variable names.

5 To get a name in the box, select the place in the formula you want it to go to then select it in the variables box and use the transfer arrow. If part of the formula is selected then the variable name will replace it. The best way to see how it works is to play, don’t forget to check the results afterwards!

6 Get the formula to look like the one here, each variable name is separated by a comma, no question marks, they’re all replaced by variables.

7 The final result in our case should read *MEAN(jump1,jump2,jump3)* when it does click OK.

8 Check The Results!



References.

ALTMAN, D. G. and BLAND, J. M., (1994), Statistics Notes: Diagnostic tests 1: sensitivity and specificity, *BMJ*;308;1552-

ALTMAN, D. G. and BLAND, J. M., (1994), Statistics Notes: Diagnostic tests 2: sensitivity and specificity, *BMJ*; 309;102-

ALTMAN, D. G., (1995) Practical Statistics for Medical Research, Chapman and Hall

BOYD, N. F., WOLFSON, C, MOSKOWITZ, M., et al. (1982) Observer variation in the interpretation of xeromammograms, *Jrnl Nat. Cancer Inst.*, 68, 357-63,

BRACE Nicola, KEMP Richard & SNEGLAR Rosemary (2000) SPSS for Psychologists. Macmillan press.

BUCKINGHAM Alan, SAUNDERS Peter, (2004), The Survey Methods Workbook, Polity Press

CLEGG Frances, (1995), Simple Statistics, Cambridge University Press.

DIENES Zoltan, (2001), *One-way Repeated-Measures ANOVA*, [online] Last accessed on 2 October 2007 at ;
URL:www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/RM%20II%20SPSS%20repeated%20measures.doc

FIELD Andy, (2006), Discovering Statistics using SPSS, Sage.

KINNEAR Paul R. and GRAY Colin D., (1997) SPSS for Windows made simple. Psychology Press.

Lenth, R. V. (2006). Java Applets for Power and Sample Size [Computer software]. Last accessed on the 2nd of November 2007 at URL: <http://www.stat.uiowa.edu/~rlenth/Power>.

PALLANT Julie, (2007) SPSS survival manual: a step by step guide to data analysis using SPSS for Windows, Open University Press

ROWNTREE Derek (1981) Statistics Without Tears. Penguin Books ltd.

TexaSoft, (1996-2001) *Interpreting Statistical p-values*. [online]. Last accessed on 11 June 2002 at URL: <http://www.texasoft.com/pvalue.html>

UCLA Academic Technology Services, SPSS FAQ: What does Cronbach's alpha mean? last accessed 10/11/2006 at; <http://www.ats.ucla.edu/STAT/SPSS/faq/alpha.html>

University of Western Ontario, Faculty of Social Science. [online]. Last accessed on 11 June 2002 at;
<http://www.ssc.uwo.ca/sscl/statsexamples/sas/anova/singlefactorproblem.html>



WUENSCH Karl L. , (2002), Dr. Karl L. Wuensch's SPSS-Data Page, last accessed 27/9/2006 at <http://core.ecu.edu/psyc/wuenschk/SPSS/SPSS-Data.htm>