An Introduction to SPSS 22

based on a version written by

Sidney Tyrrell

for the Engineering and Computing Faculty

Updated to Version 22 for Sigma by Mark Hodds

Reformatted with appendices by John Addy & Tim Sparks

Contents

1 An Overview

1.1 Introduction

These notes aim to introduce you to SPSS22, and give outline guidance for performing simple statistical analysis. SPSS22 comes with an extensive Help menu, built-in tutorials, and a Statistics Coach, all of which are explored briefly, so that we are hopeful that by the time you have worked through them you will feel both confident and competent in using the package.

As well as being available centrally on University PCs, the University license enables you to put a copy on your personal PC. For more information about this contact IT support located on the ground floor of the Library.

The notes refer to the **pulse.sav** data set, which is on the **W** drive.

W:\EC\STUDENT\SIGMA Statistics Workshops\Workshop 4

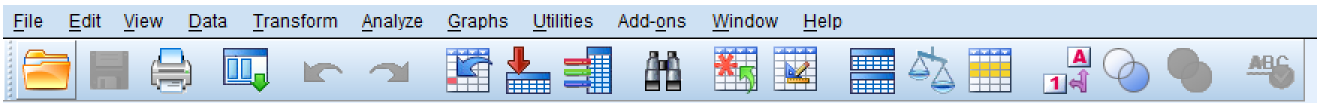
We start with an overview of SPSS22. It is then important to know how to put data in and after that how to get information out.

1.2 Getting Started

Log on to the Network

Use **Start> All Programs** > **IBM SPSS Statistics** > **IBM SPSS Statistics 22** (referred to as SPSS from now on)

- You will get a Dialog box which you can cancel the first time you enter SPSS.
- SPSS is like a spread-sheet **but it does not** update calculations, tables or charts, if you change the data.
- At the top of the screen are a series of menus which can be used to instruct SPSS to do something:

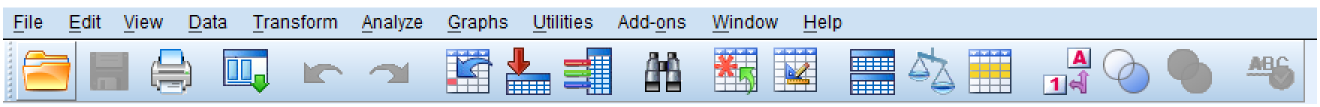


SPSS uses 2 windows: the Data Editor, which is what you are looking at and which has 2 tabs at the bottom, and the output window (Viewer).

The Viewer is not visible yet, but opens automatically as soon as you open a file or run a command that produces output, such as statistics, tables and charts.

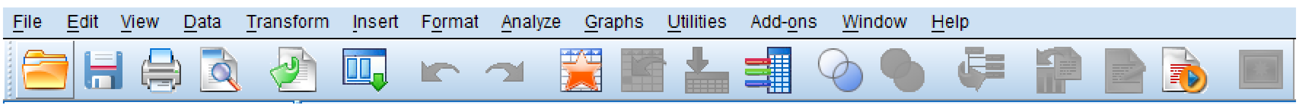
The toolbars are the same in each window but the icons are different. To switch between the two windows use the tabs at the bottom of the screen, or the **Window** command at the top.

The Data Editor window:



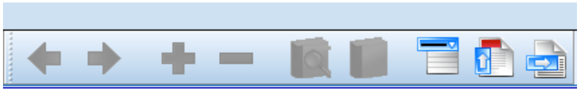
Open Data File | Save | Print | Recall recent dialog boxes | Undo | Redo | Go to case | Go to variable | Variables | Find | Insert Case | Insert Variable | Split | Weight Cases | Select Cases | Show labels | Use Sets | Show All | Spell check

The Output window:



Open Output File | Save | Print | Print Preview | Export | Recall recent dialog boxes | Undo | Redo | Go To Data | Go to case | Go to variable | Variables | Use Variable Sets | Show All | Select Last Output | Associate Autoscript | Create/ Edit Autoscript | Run Script | Designate Window

Cont.



Promote Demote Expand Collapse Show Hide Insert New New
selected items Heading Title Text

1.3 Opening a file

The data we shall use comes from a class of 92 students. Students were asked to take their pulse for a minute; these pulse rates are in the first column. Each student was then asked to toss a coin; if it came up heads they had to run on the spot for a minute, while the others sat down. At the end of the minute the whole class took their pulse rates again, and those figures are the ones in the second column. The students also supplied additional information about themselves.

The data file is called **pulse.sav** and is located on the **W drive** under

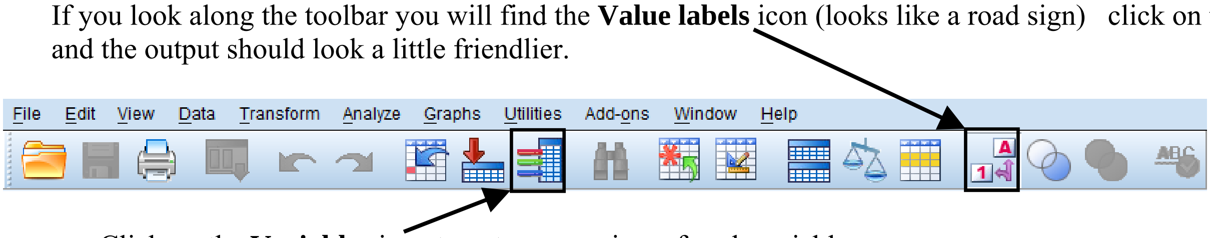
W:\EC\STUDENT\SIGMA Statistics Workshops\Workshop 4

To open the file use **File > Open > Data**

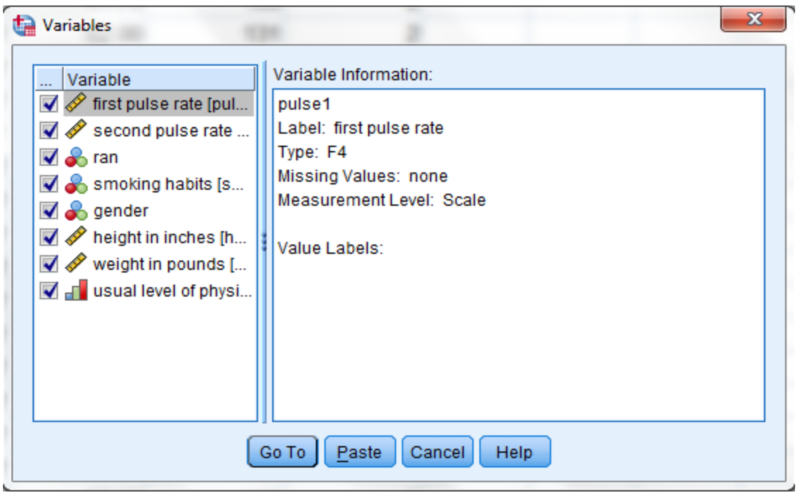
- Find the W drive then **W:\EC\STUDENT\SIGMA Statistics Workshops\Workshop 4**
- Double click on the appropriate directories to open each
- Double click on the file **pulse.sav** (if this doesn't work from the W drive you may need to move this file somewhere else, e.g. to the desktop, and try again)

At first you will probably be faced by a mass of seemingly meaningless numbers.

If you look along the toolbar you will find the **Value labels** icon (looks like a road sign) click on this and the output should look a little friendlier.



Click on the **Variables** icon to get an overview of each variable.



Now you can see each variable and its corresponding information.

1.4 Frequencies

All that data looks a bit overwhelming so we need to get a handle on it and pick out the main messages.

First of all how many men and women are there in this group?

For a simple count, and for percentages use

Analyze > Descriptive Statistics > Frequencies

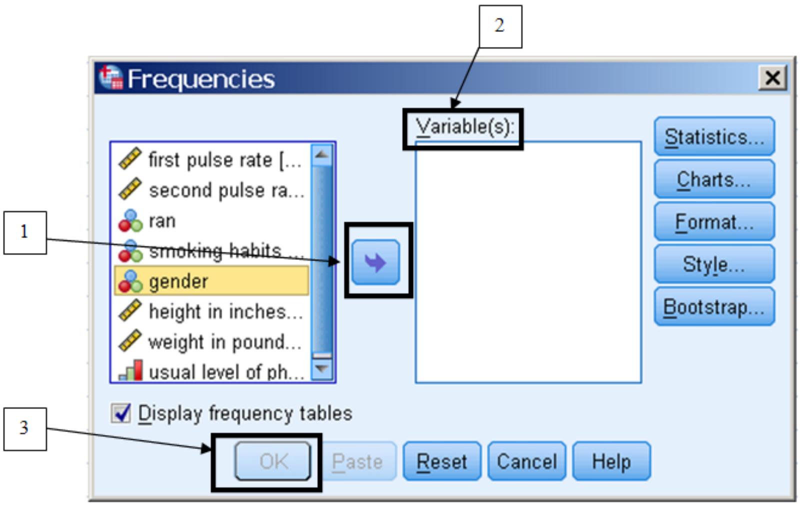
SPSS uses **Dialog boxes** for the selection of variables and options.

The source list contains the list of variables, with icons (see page 39) indicating data types.

- Your dialog box may have listed both the variable **labels** such as ‘smoking habits’, as well as the variable **names**, e.g. ***smokes***. It is more helpful in analysis to see these labels.
- If they are not shown use **Edit** > **Options**
- Select the **General** tab and look for **Variable Lists Display Labels**
- Unfortunately it doesn’t take effect until you next load a file.

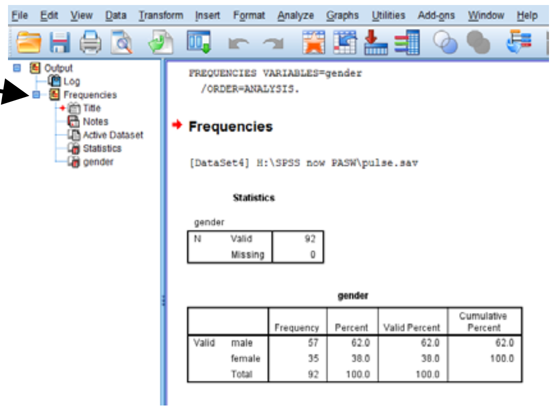
Steps

1. Use the **arrow button** to move a variable to the **Variable(s)** box on the right.
2. Place ***gender*** in the **Variables(s):** box,
3. Then click “ok”.



The resulting output introduces us to the Viewer window, and shows that 57 students, or 62%, were men.

If you don't need it all for the moment, you can hide it by clicking on the **minus** sign that appears in the left hand frame.



To hide one item, click on the minus sign. This is useful if you only want to print a small selection, as only what is shown is printed.

To change the order in which the items are displayed, drag and drop in the left hand pane - try it.

To delete, click on an item and press delete.

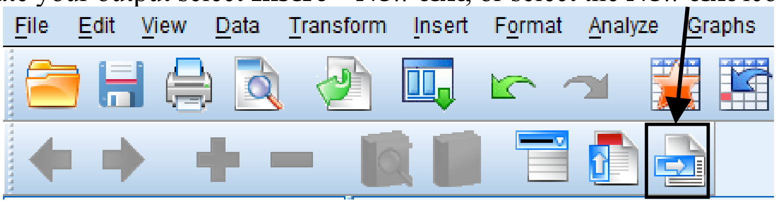
The left-hand pane contains the outline view. To go directly to an item click on it, this is very useful when you have masses of output.

Exercise

Using **Analyze** > **Descriptive Statistics** > **Frequencies** find how many different physical activity levels there are? …….

1.5 Annotating output

Never do any analysis without thinking beforehand what you want, and how to interpret it.
To annotate your output select **Insert** > **New text**, or select the **New text** icon.

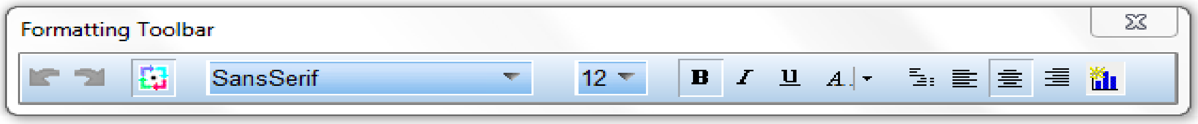


Back to the output itself, this **can be edited**. SPSS tends to give Output that is **unnecessary**, and we always **delete unnecessary** Output.

Tip: Always delete unnecessary Output, and annotate the rest as you go.
Click on all the text at the top of the screen and press Delete on your keyboard.

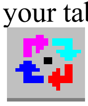
Double click on the table to bring up the Formatting Toolbar.

gender					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	male	57	62.0	62.0	62.0
	female	35	38.0	38.0	100.0
	Total	92	100.0	100.0	

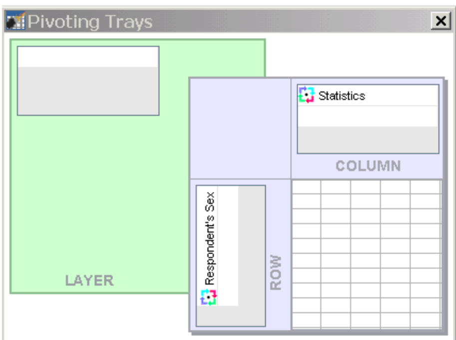


If it does not appear use **View** > **Toolbar**

- Click on any text to change its format and use the Formatting Toolbar to do so.
- Double click to rewrite the text itself.
- When you have finished close the Editing window by clicking on the “x”
- The Formatting Toolbar also gives Pivoting Control (!).
- Pivoting control is a useful device, which enables you to change the look of your tables.
- Click on the icon to bring up the Pivoting Tray, if it is not already shown.



Clicking on each of the icons at the edges tells you what they represent.
Here the columns are Statistics, and the Rows are Respondent’s Sex.
Drag the Statistics icon on to the Row bar so that the 2 are side by side, to see how the table changes; drag it back before proceeding.



You can copy Output into Word by clicking on it and using **Edit > Copy**, In Word use **Edit > Paste**. You can also export it to Word (see page 9)

1.6 The SPSS Tutorial

The SPSS Tutorial is an extremely useful feature

Click on **Help** > **Tutorial**
Click on the **Introduction** book and take it from there.

Finally, in this overview, visit **The Statistics Coach**. Click on **Help** > **Statistics Coach**

As an example, follow the default settings, and click **Next** each time.

Summarize, describe or present data	**Next >**
Data in categories	**Next >**
Tables and Numbers	**Next >**
Counts or percentages by category	**Finish** > **OK**

2 Entering Data

2.1 *Opening an existing SPSS file*

We have already covered this earlier,
To open an existing file we use:

File > Open > Data in **the W drive,**
W:\EC\STUDENT\SIGMA Statistics Workshops\Workshop 4\pulse.sav

2.2 *Entering data directly*

Entering numbers and text.
The Data Editor Window looks suspiciously like a spreadsheet, and numbers and text can be entered directly.

Be warned, though it looks like a spreadsheet it does not behave like one. Your charts and output will not automatically update if you should change the original data, and you cannot enter formulae directly into a cell, though you can do calculations using a different facility.

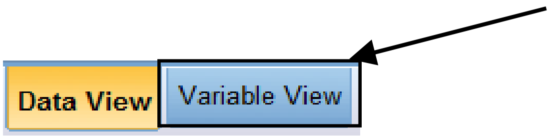
Try entering some numbers in the column to the right of activity.
Type what you want in each cell; press the return key or a cursor key, if you make a mistake retype the entry.

Now try to put some text into the same column.
Can you? You can type it in but when you press Enter it disappears.
This is because SPSS has identified the column as a numeric one and won't allow any text; see 'Type' column where the type of variable (numerical, string etc.) is defined.
Put some names of countries in the next column to the right including The United States of America and Australia. What happens? Most probably it is cut short. To solve this change the 'Width' specification in the corresponding column.
Try entering numbers in this column – you can but you will not be able to do any calculations with them as SPSS thinks they are text. (Notice that they are 'left-aligned' rather than 'right-aligned')

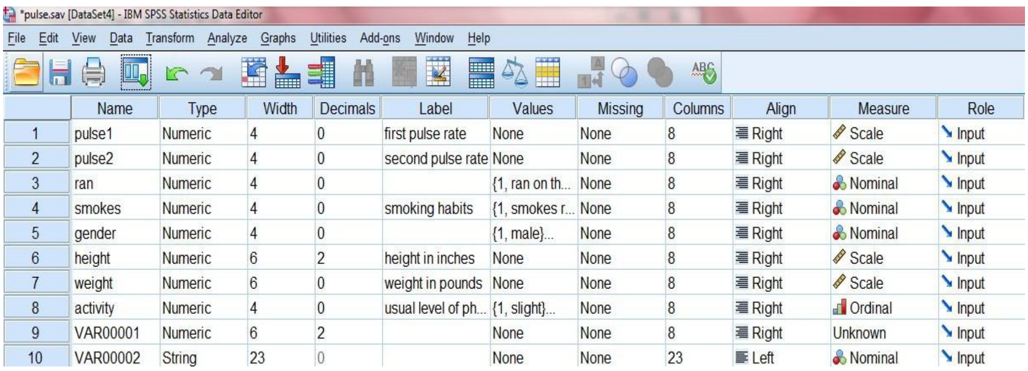
Your new variables have been given the names VAR00001 and VAR00002 which we will now change.

2.3 *Defining Variables*

At the foot of the screen are two tabs. Click on **Variable View** to get the following screen.



Overtype VAR00001 and VAR00002 with the names of your new variables: **xyz1** and **xyz2** if you can't think of anything!



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	pulse1	Numeric	4	0	first pulse rate	None	None	8	Right	Scale	Input
2	pulse2	Numeric	4	0	second pulse rate	None	None	8	Right	Scale	Input
3	ran	Numeric	4	0		{1, ran on th...	None	8	Right	Nominal	Input
4	smokes	Numeric	4	0	smoking habits	{1, smokes r...	None	8	Right	Nominal	Input
5	gender	Numeric	4	0		{1, male}...	None	8	Right	Nominal	Input
6	height	Numeric	6	2	height in inches	None	None	8	Right	Scale	Input
7	weight	Numeric	6	0	weight in pounds	None	None	8	Right	Scale	Input
8	activity	Numeric	4	0	usual level of ph...	{1, slight}...	None	8	Right	Ordinal	Input
9	VAR00001	Numeric	6	2		None	None	8	Right	Unknown	Input
10	VAR00002	String	23	0		None	None	23	Left	Nominal	Input

Click in the cell under Type and click on the small box with 3 dots in it to bring up a **Variable Type** box which you can use to define your variable; in a similar way you can control the number of decimal places shown, column width etc. by typing in a number directly, or using the little 'up-and-down' arrows.

Tip: It is better to enter most data as numerical codes and then provide labels explaining what the codes represent. Adding Variable labels is explained in section 2.5 (below).

2.4 Entering data via a spreadsheet

Excel spreadsheets can be opened in SPSS. You can also simply copy and paste the data cells from Excel into SPSS but you will have to label the columns.

The pulse data has been saved as an Excel spreadsheet. To open it use **File > Open > Data**
Ask the dialogue box to display **All files** and not just the SPSS ones.
Find the file **pulse4.xls** in:

W:\EC\STUDENT\SIGMA Statistics Workshops\Workshop 4

SPSS will recognise the format and automatically give a dialogue box.

Tick **Read variable names**.

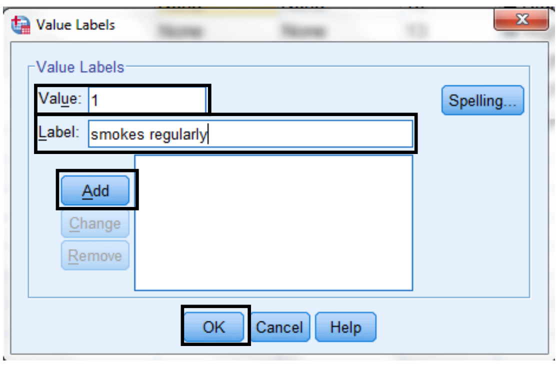
Click **OK**.

Exercise

Using **Analyze > Descriptive Statistics > Frequencies** find out how many smokers there are? This time there are no Value Labels, which makes the output more difficult to read.

2.5 Adding variable and value labels

Move to the Variables view of the Data Editor. (Double click at the top of the smokes column or click on the tab at the foot of the screen).



Click on the cell under the Label column and type in a suitable label e.g. smoking habits. Click in the cell under the Values column and bring up the Define Variable box.

Enter a value in the **Value** box, here it is 1

Type an appropriate label in the **Label** box, e.g. ***smokes regularly***
Click on **Add**

Enter the value 2, and a label, e.g. ***non-smoker***
Click on **Add**
When all the values have been entered click on **OK**

A very useful tip: If you are entering lots of identical value labels for different variables, e.g. 0 = No and 1 = Yes, enter them for one variable.

Then right click on the cell
Select Copy
Go to a new variable(s) and use Paste under the Value column. This is a huge time saver!

To return to the data click on the **Data View tab** at the bottom of the screen.

Variable names can now be 64bytes long but start with a letter. They cannot contain spaces or the %

sign but can contain numerals.

Keep them short!

3 Editing and Handling data

3.1 Editing entries

Reopen the file **pulse.sav** found in

W:\EC\STUDENT\SIGMA Statistics Workshops\Workshop 4

Try each of the following (It doesn't matter if you change the data, as it is only for demonstration).

- ***Correcting entries***
  Any entry can be over-typed.
  Click on the cell, type in the correct entry and press Enter.
- ***Deleting entries***
  To delete an entry for a cell, click in the cell and press delete.
  Complete columns and rows can be deleted by clicking on the grey cell at the top or side and pressing the Delete key on the keyboard.

  Remember the useful Undo icon! 
- ***Copying cells***

  Cells, columns and rows can be copied by first highlighting them and then using the **Edit Copy** menu.

3.2 Inserting new data

- ***Inserting a variable (a column)***
  Click on the top of the column to the right of where you want the new column to appear. Use **Edit > Insert > Variable**
  or **Right Click** at the top of the column to the right of where you want the new column to appear, and use Insert Variable
- ***Inserting a case (a row)***
  Click at the side of the row below where you want the new row to appear. Use **Edit > Insert** Cases.
  Or **Right Click** at the side of the row below where you want the new row to appear, and use **Insert Cases**.
- ***Moving data***
  You can drag and drop columns to wherever you like – highlight them first.

3.3 Sorting data

SPSS can sort the data, e.g. by initial pulse rate using **Data > Sort Cases**.

In the dialog box highlight the '**first pulse rate**' variable.
Click on the arrow to transfer it to the **Sort by** box
then **OK**

3.4 Saving and output

Data and output have to be saved separately, using **File > Save**

Charts are saved as part of the Output in a **.spv** file; data is saved as a **.sav** file. Once you have saved your data it can be exported in another format.

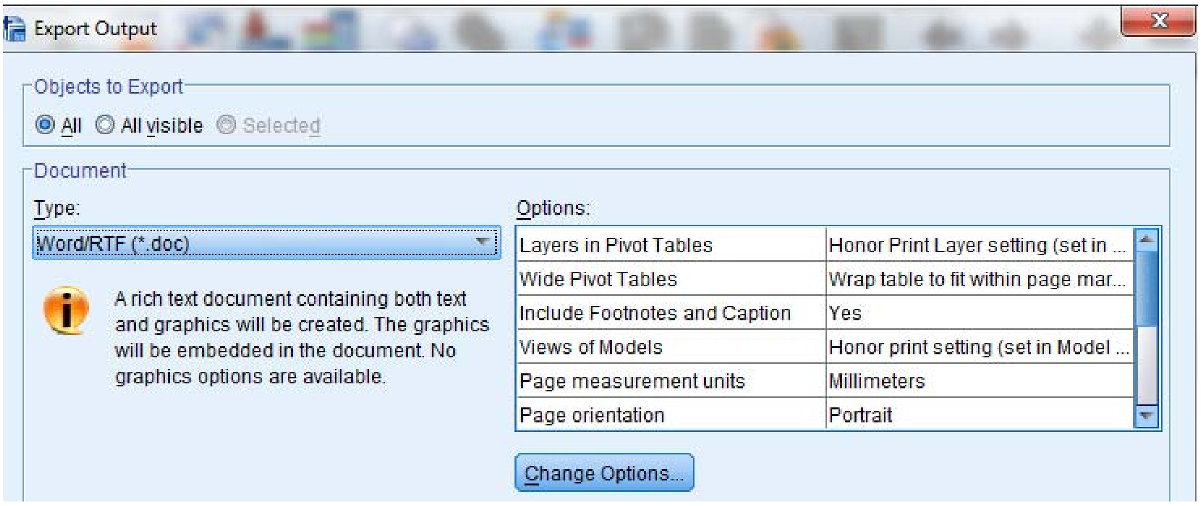
Output can also be saved as a **Word RTF** (**Rich Text File**) which contains graphics. Use **File > Export** and choose Word/RTF from the drop down box

It is an excellent rule to save frequently.

3.5 Exporting your output to Word

You can copy Output into Word by clicking on it and using **Edit** > **Copy**, in Word use **Edit** > **Paste**.

Output can be exported as a Word RTF file or Text file
Use **File** > **Export** and select the appropriate entry under Type.



Copying tables and charts into Word

In the Viewer window click on what you want to transfer to Word, either a table or chart.
Use **Edit** > **Copy** and in Word use **Edit** > **Paste.**

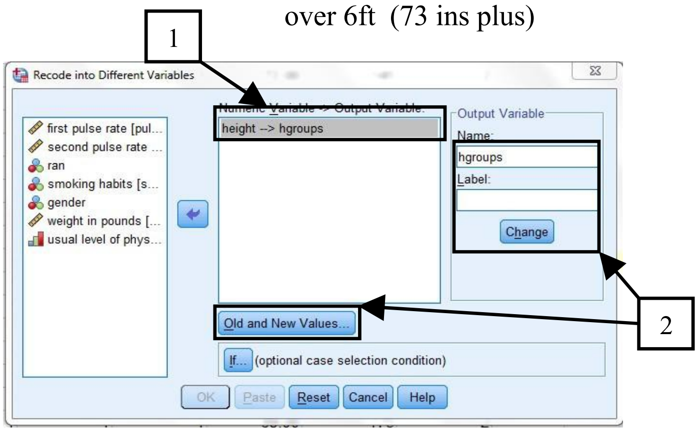
3.6 Printing from SPSS

You can print directly from the Viewer window, using **File** > **Print** but **use Print Preview first** to make sure you have what you want. To print just one specific thing click on it first to select it.
Output that you don't want can be hidden by clicking on the icons in the left hand pane

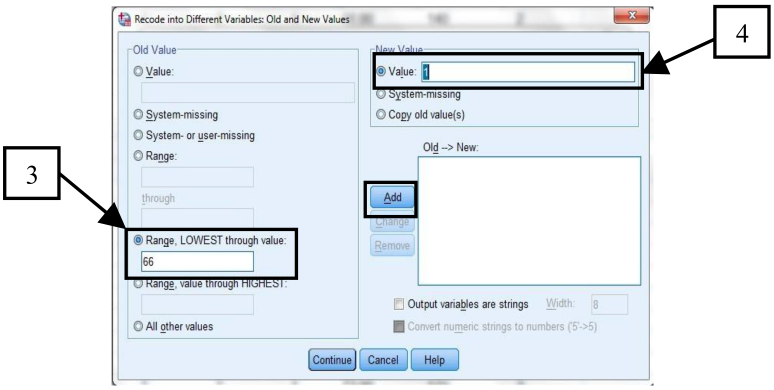
3.7 Recoding into groups

You will find it very useful to be able to recode data.
The pulse data includes the heights of students and it might be useful to regroup the data into subgroups and give each group a numerical code.
Ask for a frequency table first to see where it might be appropriate to divide the data.
The heights are given in inches (12 inches to a foot) and as an example I suggest recoding the students into 3 groups:

5ft 6inches (66 ins) and under	Group 1
5ft 7ins to 6ft (67 to 72 ins)	Group 2
over 6ft (73 ins plus)	Group 3



1. Use **Transform** > **Recode Into Different Variables**
   Place ***height*** in the large box. Name the new variable ***hgroups*** in the right hand box.
2. Click on **Change.** Type height groups in the **Label.** Click on **Old and New Values**



3. To get the next dialogue box: Click on **Range, LOWEST through value** and place **66** in the box.
4. Type 1 in the **New Value** box on the right hand side, as shown
   Click **Add**.

Recode the other groups as follows:

Group 2 67 inches to 72 inches (5ft 7 ins to 6 ft)
For **Old Value** use **Range 67 through 72** and for the **New Value 2**

Group 3 over 72 inches, i.e. 73ins plus (over 6ft)
For **Old Value** use **Range 73 through highest** and for the **New Value 3**

Complete the recoding then use **Continue** > **OK**

Exercise
1. Provide **labels** for the new variable **hgroups** to explain what the numbers represent.
2. Change the display for **hgroups** to zero decimal places

3.8 Doing calculations on variables

The Pulse dataset is American and at present the heights are in inches and the weights are in pounds.

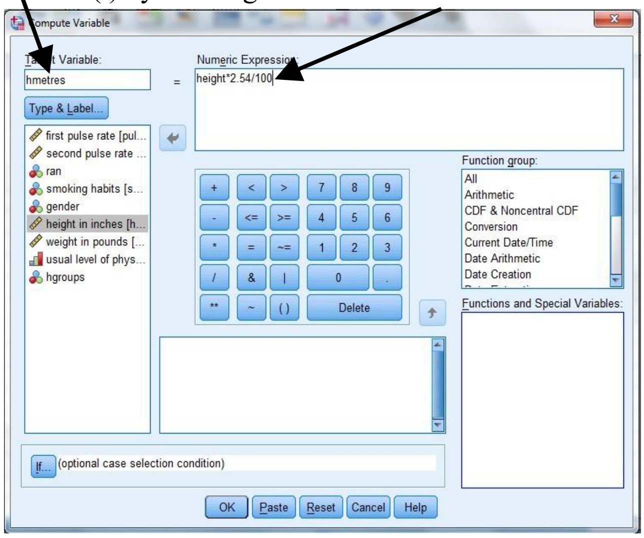
Although we can't enter a formula into a cell of the Data Editor to convert these values we can **transform** the measurements into meters and kilograms respectively using the **Compute** function.

Use **Transform > Compute Variable**

The example shows what has to be filled in for recalculating the heights in meters.

A new variable **hmetres** has been created.

The height, in inches has been multiplied (*) by 2.54 to give cm and divided (/) by 100 to give metres



Exercise

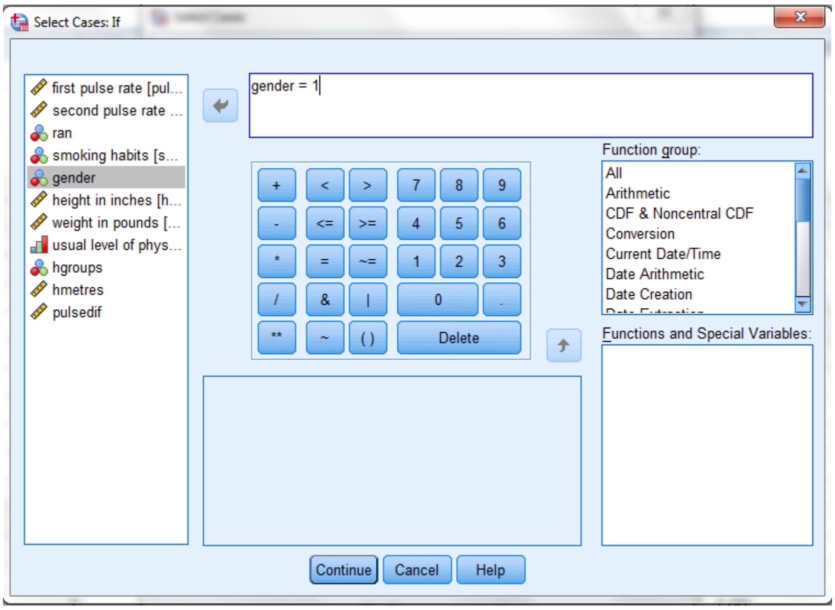
1. Transform the weights into kg. There are 2.2lb per kg so you will need to divide by 2.2.

2. Create a new variable that shows the difference between the second and first pulse rates Call this pulsdif. You will be using this variable later so at this point save your data to your H drive using **File** > **Save** from the menu bar.

3.9 Selecting a subset

During your investigations you may want to look only at the data for the males, or females, say.

SPSS enables us to select just these cases using
Data > **Select Cases** > **If condition is satisfied** (click in the circle next to this)
Click on the **If …** button under **If condition is satisfied** (the **If button** will not be available if you have not clicked in the circle). Enter the appropriate condition, e.g. the example shows what has to be filled in for selecting males.

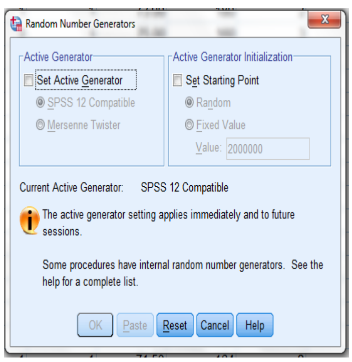


Continue select **Filter out unselected cases OK (Tip: Do not delete the other cases as they will be lost for good).**

If you scroll down the data sheet you will notice that the females are crossed out on the left, and are now ignored in any operation. Try a frequency of genders and see what you get.

To **restore all the data** use **Data** > **Select Cases** > **All cases OK**
Be warned: this is all too easily overlooked when you have been working on only part of the data, and then decide to analyse what you think is the complete data set.

3.10 Selecting a random sample



This is a two stage process: first we set the starting point and type of random number generation. Then we select the actual sample.

To select the starting point, use **Transform** > **Random Number Generators**

Select Set Starting Point
Choosing Random allows a different start point for the random selection each time you enter SPSS
Entering a Fixed Value (which can be any number) allows a random selection to be reproduced. Try them both in the next example and see what happens. Click OK following your selection.
If you do not set a starting point you will get the same random selection each time you enter SPSS.

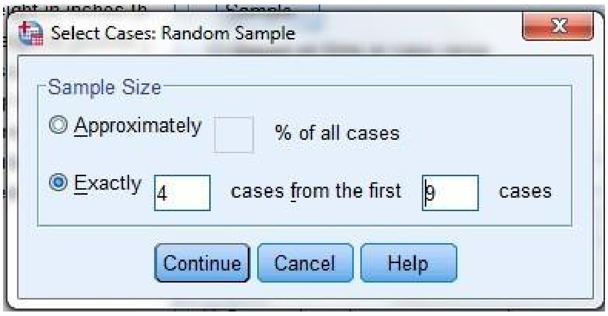
To select the actual sample use:

Data > Select Cases > Random Sample of Cases

Click on the **Sample** button

Fill out the dialogue box appropriately.

Suppose you wanted to select a random sample of 4 from the first 9 cases, the box would be set out as follows:



3.11 Merging files

1. ***Adding Variables to an existing file***

Open the file **File 1 for merge.sav** and **File 2 for merge.sav** from

W:\EC\STUDENT\SIGMA Statistics Workshops\Workshop 4

There are 10 cases, each uniquely identified by an id number. Open the file.

Always open using File Open – do not double click from Windows Explorer as this will open another running of SPSS.

Notice that this has the same cases (same id numbers) but with a new variable Region
We shall now merge the two files so that we have all the variables in one.

In order to do this you must have a key variable which identifies each case, and you must have sorted the files so that the key variable is in the same order in each.
In the files you have opened the key variable is in the same place.

Make sure File 1 for **merge.sav** is the active file (ie make sure you have this file displayed) Use **Data > Merge Files > Add Variables**.
Choose the file **File 2 for merge.sav** from the list under **An open dataset** and click **Continue**

2. ***Adding cases to an existing file***

Open the file Adding cases 1.sav from

W:\EC\STUDENT\SIGMA Statistics Workshops\Workshop 4

This contains 10 cases to which we shall add another 10 from **Adding cases 2.sav**
Use **Data > Merge Files > Add Cases**
Select **An external SPSS data file** and click the **Browse** button, then select **Adding cases 2.sav**
Click on **Open**
There should be no unpaired variables
Click on **OK**
You should now have a file with 20 cases.

4 Descriptive Statistics

4.1 A 'very rough guide' as to what is appropriate to use when

The Analyze function in SPSS enables us to summarise our data in a number of ways.
The confusion is what to use when, especially as there is often more than one way of doing things in SPSS.

This section provides a guide to what to use, a brief look at the functions in turn and some exercises for you to try out using the pulse data.

What follows is a brief description of the following functions, with exercises: Frequencies, Descriptives, Explore, Crosstabs and a brief look at other Tables

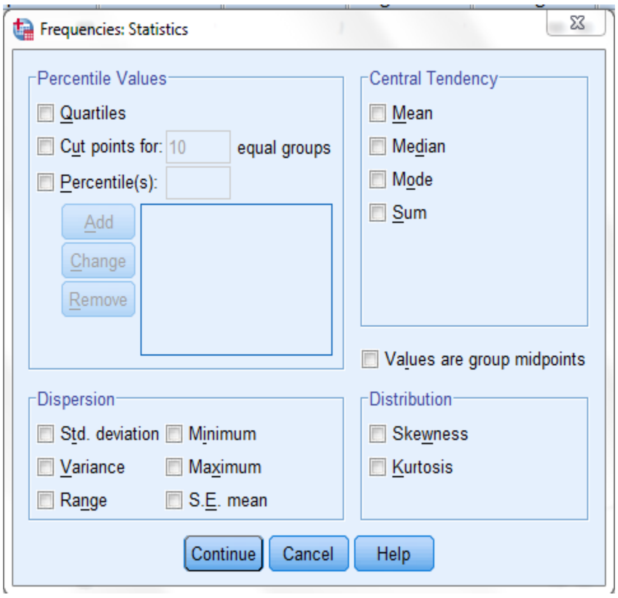
Task	SPSS function	Comments
Counts	Frequencies (offers charts too) Crosstabs	Use %'s as well as counts. %'s are used for comparisons. Round %'s to the nearest whole number in reports.
Averages and Measures of spread	Frequencies with the Statistics option Descriptives Explore (offers charts too)	Make sure you use a sensible measure, e.g. the mean gender is meaningless.
Comparing sets of data	Explore (offers charts too) Crosstabs Basic Tables General Tables (use Analyze > Custom Tables)	Beware of using boxplots for inappropriate data, e.g. nominal. Crosstab tables can look untidy, so think carefully about the number of levels and the information required in them. Basic Tables are plain and simple. General Tables offers more options than Basic Tables, if needed. Extra information can be edited out. All tables can be modified
Looking for relationships	Tables: Crosstabs Scatterplots (see Scatter/Dot in the Graphs menu)	Plots and tables give a visual impression of possible relationships: the eyeball test. You may then need to follow this up with the appropriate statistical test.

4.2 Frequencies

Analyze > Descriptive Statistics > Frequencies

Frequencies are used when you want to know how many of something you have. However, additional statistics available via the Statistics button makes it far more useful than just counting. **This is the best function for overall summaries**

The **Charts** button is particularly useful; automatically producing charts of your data. The **Statistics** button brings up the following dialogue box:



These statistics would be helpful for height but don't be tempted to use them for gender!

Exercise

Use Frequencies to find the percentage of students who took moderate exercise, and draw a bar chart to represent this.

4.3 Multiple response variables

When you write a questionnaire you often include a question where the respondent can tick more than one response.

Open the Data file **pets.Sav** from

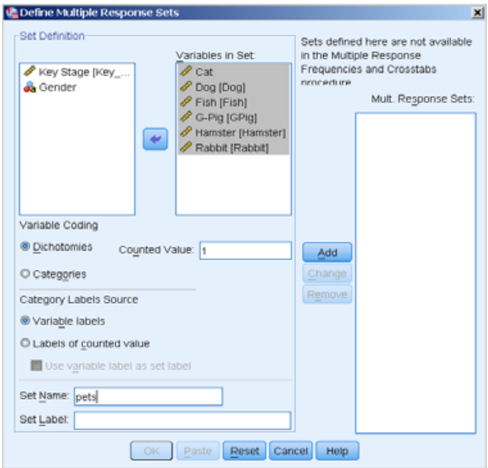
W:\EC\STUDENT\SIGMA Statistics Workshops\Workshop 4

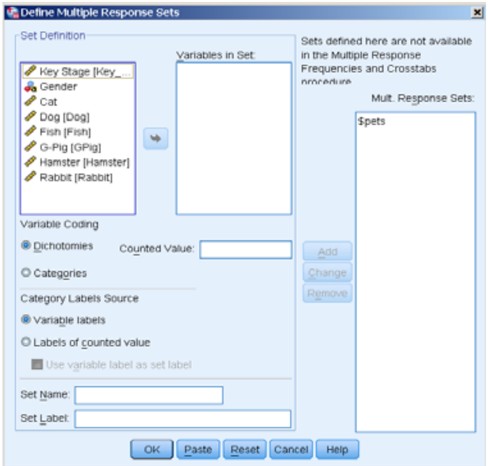
This contains data about 73 school children and their pets.

Using frequencies we could obtain a separate table for each pet but SPSS can combine these multiple responses into one table for you.
Use

Analyse > Tables > Multiple Response Sets.

First we need to define our Multiple Response set.
Fill out the dialog box as shown, with the various pets in the Variables box
Dichotomies Counted Value: 1 (because there is a 1 in the column when a pupil has that pet)
Name: **pets**
Click on **Add** then **OK**





NB The new variable is called $pets to indicate a multiple response set. You get this output:

Multiple Response Sets

Name	Coded As	Counted Value	Data Type	Elementary Variables
$pets	Dichotomies	1	Numeric	Cat Dog Fish G-Pig Hamster Rabbit

Now use **Analyze > Tables > Custom Tables**

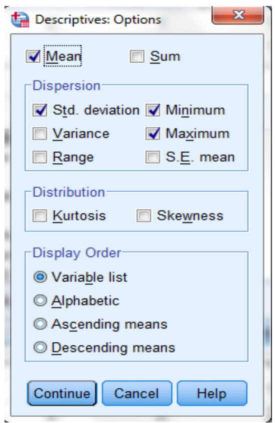
Your variable pets should now appear in the Table dialog box. Place it in the **Rows** and Click **OK**. For percentages use the **Statistics** button.

4.4 Descriptives

For the rest of the exercise return to the **pulse.sav** data.

Analyze > Descriptive Statistics > Descriptives

Descriptives offers much less than Frequencies - only giving a mean for averages, and the standard deviation and range for spread. “Options” brings up the following dialogue box:



Exercise

Find the mean and standard deviation of the first pulse rate (pulse1).

Answer

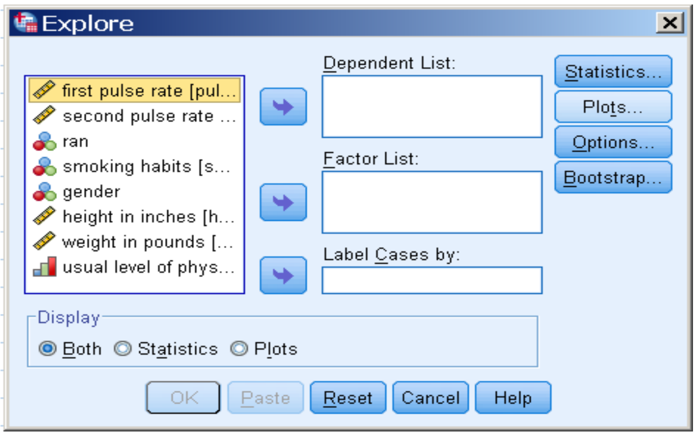
(mean: 72.87 beats per minute; standard deviation: 11.01 beats per minute)

4.5 Explore

Analyze > Descriptive Statistics > Explore

This is an extremely useful command when you need to compare two sets of data, e.g. heights of males and females. It explores the differences.

The example shows the dialogue box set up to compare the first pulse rate of men and women. SPSS has been asked to display both statistics and charts (boxplots, stem and leaf plots, histograms) - again a very useful automatic facility.



Exercise

Find the mean first pulse rate of men and women.

Draw the boxplots and note the differences.

Answer

(female mean: 76.9 beats per minute; male mean: 70.4 beats per minute.)

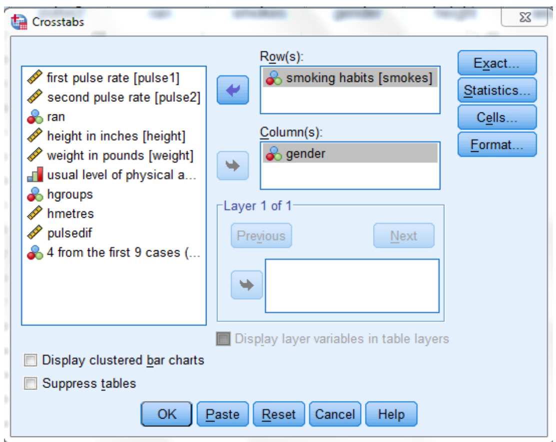
4.6 Crosstabs

Analyze > Descriptive Statistics > Crosstabs

Crosstabs produces slightly complex tables, but these can be edited to look tidier.

It has the useful additional facility of doing a Chi-Squared test (and others) if asked - use the **Statistics** button. The **Cells** button enables one to choose **Column %'s**, **Row %'s** and **Total %'s**, but it is advisable to ask for only one at a time, for clarity.

The example shows the dialog box set up to produce a table of smoking habits by gender.



Exercise

Produce a table of smoking habits by gender.

NB Producing a table with a variable taking many different values, e.g. height, is not a good idea.

What percentage of students are female and smoke?
What percentage of smokers are female?
What percentage of females smoke?

Tables can be tricky!

Look at these 2 tables and answer the following questions:

What % of females thought the prospects for the police will be better in future? Of those who thought the prospects better in future, what % was female?

prospects_police	female	male	Total		prospects_police	female	male	Total
better	58%	42%	100%		better	34%	28%	31%
same	51%	49%	100%		same	47%	52%	49%
worse	52%	48%	100%		worse	19%	20%	19%
Grand Total	53%	47%	100%		Grand Total	100%	100%	100%

The answers are 34% and 58%, respectively, and you may well have got it the wrong way round. This is often the biggest problem people have – wrongly interpreting percentages's in tables.

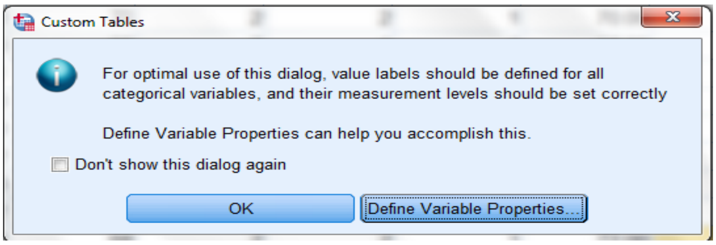
The tip is to do both column and row %'s and have them in front of you so that you can see the difference.

Crosstabs will probably produce adequate tables for all your needs, but there are other Tables functions in SPSS. Below shows an alternative way of achieving tables.

4.7 Basic tables

Analyze > Tables > Custom Tables

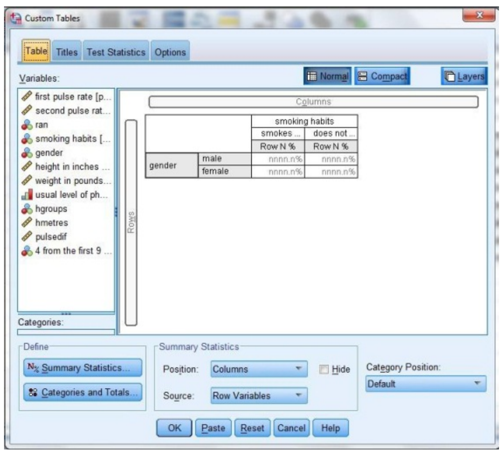
Tables are handy for summarising bivariate data in an easy-to-understand form.
The example uses Basic Tables and shows the dialog box set up to produce a similar table to the one you have produced in .



This dialog box reminds you to have the level of measurement defined for your variables – Nominal, Ordinal or Scale.

See the **Variable View** in the **SPSS Data Editor** under the **Measure** column, or use the button in the dialog box.

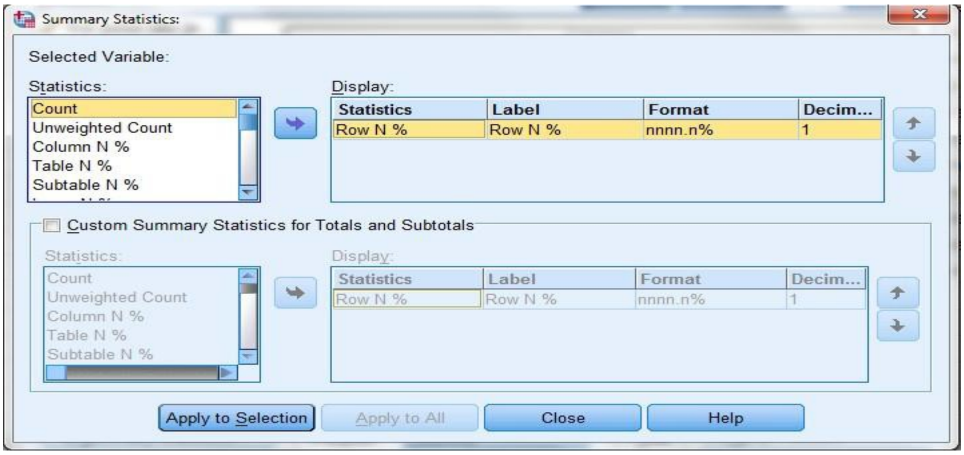
The pulse data are already labeled, so press OK, although you might want to remind yourself of the measurement types.



Exercise

1. Produce this Table and compare the differences between it and the earlier table produced by Crosstabs.

   The default option is to display counts. To display percentages, select the cell containing gender and then select the **Summary Statistics** button to get the dialog box below. Remove (or leave if required) **Count**, select **Row N%** and **Apply to Selection**.



2. Explore what the Categories and Totals button allows you to add to your table.

   You could investigate the Help section in Tutorials for Tables for more information.

5 Charts

Open the file **pulse.sav** data file from

W:\EC\STUDENT\SIGMA Statistics Workshops\Workshop 4

5.1 Introduction

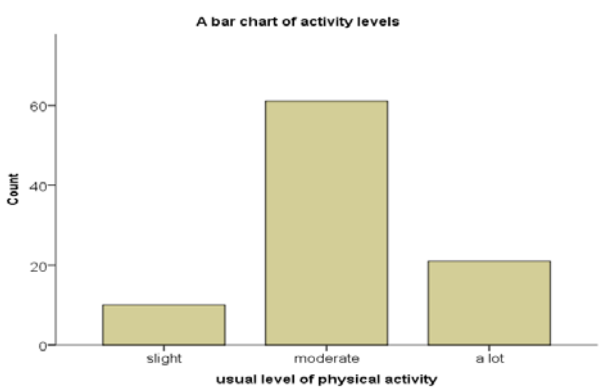
SPSS provides a wide variety of charts to choose from including bar charts, histograms, pie charts, scatterplots, and boxplots. These are accessed via **Graphs > Chart Builder**.

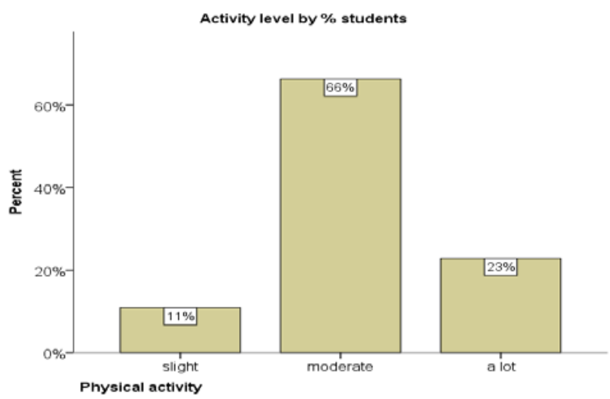
Charts should convey a message; they should help the reader to understand the data and not get confused.

Try to use as **little** 'ink' as possible – cluttered charts are not easy to understand.

Drawing appropriate charts is not as easy as it looks, so if you feel daunted use the **Charts** options under **Frequencies** and **Explore** – they will do most of the thinking for you.

Here are 2 examples:





Think about any the differences and decide which are more helpful.

In general there are 2 simple rules which will help:

- Decide what your message is and find a chart that conveys it clearly.
- Label everything, but don't swamp the chart with words – adjust the font size.

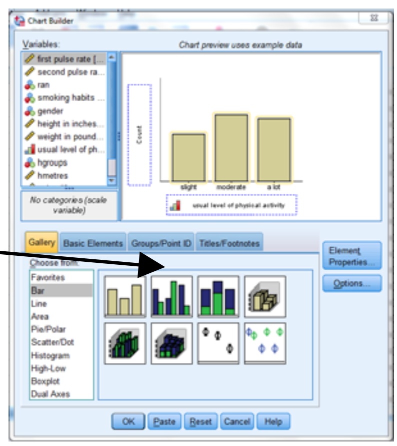
5.2 Bar charts

- ***Simple bar chart***

  Use the **pulse.sav** data file from

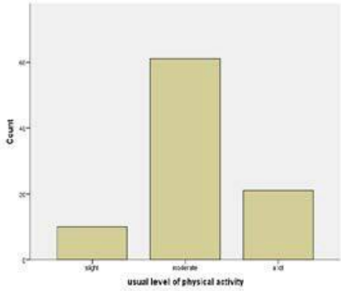
  W:\EC\STUDENT\SIGMA Statistics Workshops\Workshop 4

  Start by drawing a bar chart of the activity levels.



1. Use **Graphs > Chart Builder**
2. If you get a warning dialogue box just click **OK**.
3. Drag the left hand Bar Chart from the gallery into the main window
4. Drag u**sual level of physical activity** into the x axis box.
5. Under **Titles/Footnotes** click **Title 1** and enter **usual level of physical activity**
6. Click on **Apply**
7. Click on **OK**

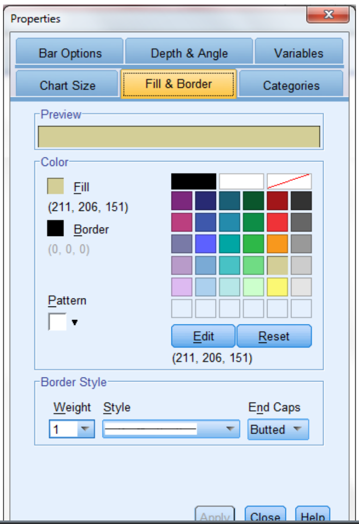
You should get this:



1. To edit the chart, double click on it, and the Chart Editor appears. Depending where you double click on the chart a **Properties** box should appear with different tabs.

2. Click on a bar; under **Fill & Border** change the colour. Click **Apply**.

3. To change the colour of a single bar click once on one bar, it alone will be selected

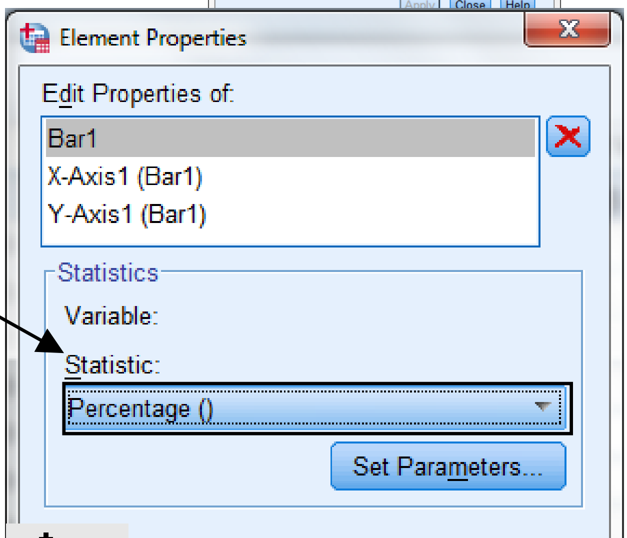


4. Apply colour as before.

5. Under **Depth & Angle** do NOT be tempted to apply shadow or 3-D, as this yields mathematically-poor representations of the data!

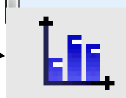
- ***Percentage bar chart***

1. Use **Graphs > Chart Builder**
2. In the **Element Properties** box (only the top half is shown), select **Percentage()**
3. **Apply**
4. **OK**

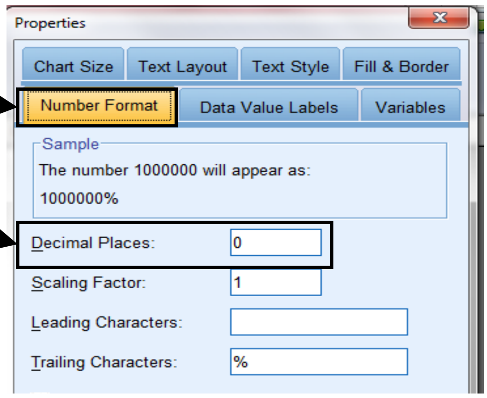


5. Double click to edit this chart

6. Click once on a bar to select them all, then on the **Show Data Labels** icon



You can amend the format of numbers by selecting the **Number Format** tab.



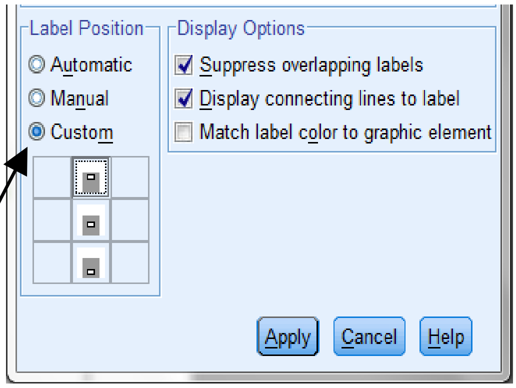
1. Always show %'s as whole numbers.
2. Type 0 in the **Decimal Places** box, **Apply**

3. Use the **Text Style** tab to increase the font size, try 12, **Apply**

To change the %'s on the axis to whole numbers:

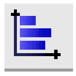
1. Click on the y axis to select it
2. Double click to bring up the properties box
3. Select the **Number Format** tab
4. Type 0 into the **Decimal Places** box

Apply



To change the position of the labels in the bars:

1. Click on the % boxes to select
2. Click on **Data Value Labels** in the **Properties** box-lengths
3. Select **Custom** and make your choice

To transpose (swap axes) the chart use the **Transpose** icon 

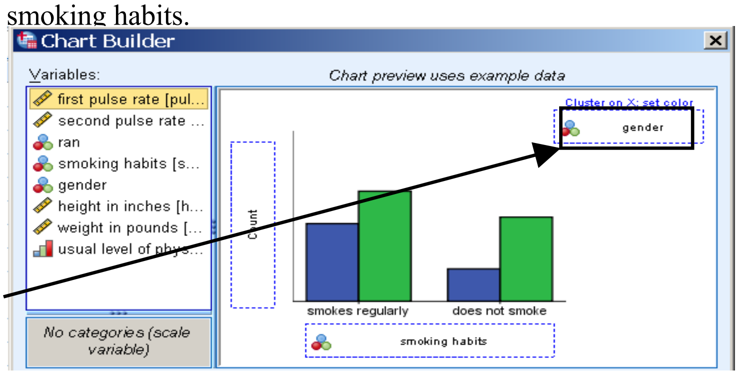
The chart can be copied from Output into a Word document using **Edit > Copy Chart**, when in Word use **Edit > Paste**.

- ***A clustered bar chart***

A clustered bar chart is good for comparisons.

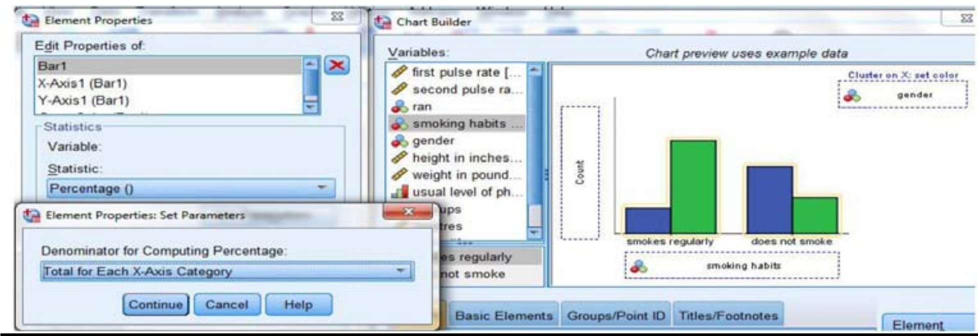
Here we shall compare male and female smoking habits.

- Use **Graphs > Chart Builder**
- **Reset**
  Drag the Clustered Bar Chart (second bar chart option) into the Gallery
- Drag smoking habits into the x axis box
- Drag gender in to the **Cluster on x box** in the top right of the Gallery window.



You can change the order of 'smokes regularly' and 'does not smoke' by clicking in the Element Properties box on Y-Axis1 (Bar1)

- ***Percentage clustered bar chart***



For a percentage chart use the Element Properties Box with

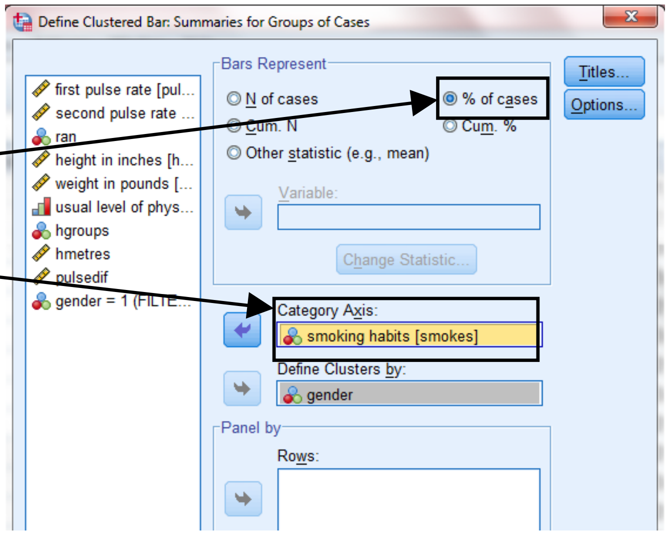
1. Bar 1 highlighted
2. Choose **Percentage()** from the **Statistic** box
3. Click on **Set Parameters**
4. Choose Total for Each x-Axis Category
5. **Continue** > **Apply** > **OK**

Warning: if you apply labels to the bars they will give the wrong %'s

Percentage Clustered Bar Chart using Legacy Dialogs
With correct labels!

For some reason %'s on charts in SPSS pose problems; here is another way of drawing the same chart but with correct labels. It uses the Legacy Dialogs option.

1. Use **Graphs** > **Legacy Dialogs** >**Bar…** > **Clustered**, using **Summaries for groups of cases** > **Define**
2. Use **%** of cases
3. Place **smoking habits** in the **Category Axis**
4. **Define Clusters by** gender
5. Click on **OK**

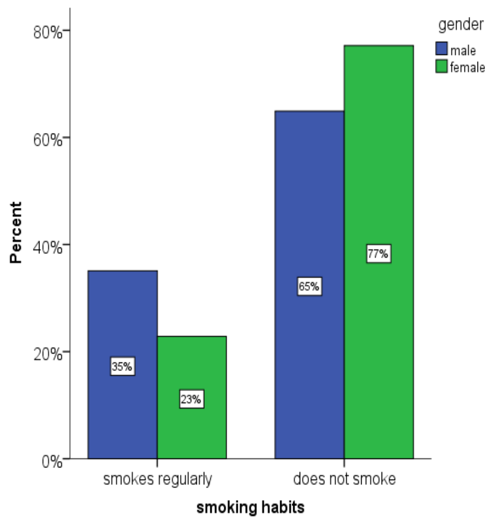


You should get a chart similar to the one below in blue and green, before editing.

If you are printing charts in black and white use strong contrasts of white and grey with some black.

Edit the chart to

1. Add the Bar Labels as before
2. Reduce the decimal places to 0
3. Increase the font size
4. Think of each colour as being a length of ribbon. All the ribbons are the same length and represent 100% of each category (males and females)
5. They are then cut up into the different sections
6. Here 77% of the women do not smoke
7. **It is NOT saying that 77% of the non-smokers are women**



- ***A stacked % bar chart***

  BEWARE: If you ask SPSS to add labels to this it will give you the wrong percentages. Create a table in cross tabs to find what the %'s should be and add the labels as text boxes.

  Use the Chart Builder, or

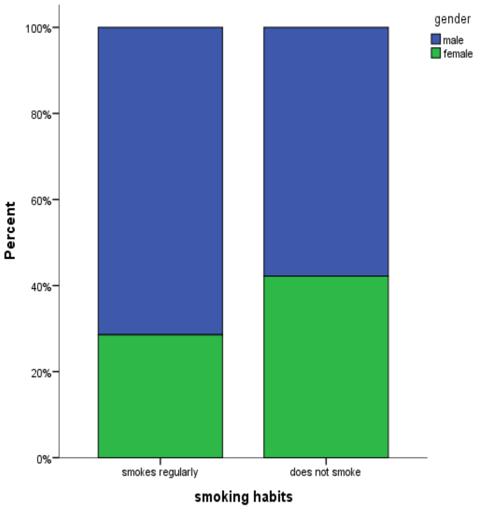
  Graphs > Legacy Dialog > Bar > Stacked > Summaries for groups of cases > Define

  Place smoking habits in the **Category Axis** and **Define Stacks by** gender

  Make sure you have number of cases **N of cases** selected > **OK**

  When your chart appears

  Click on **Options > Scale to 100%**



  Add text boxes for labels.

  What do you think about this display in terms of representing the data?

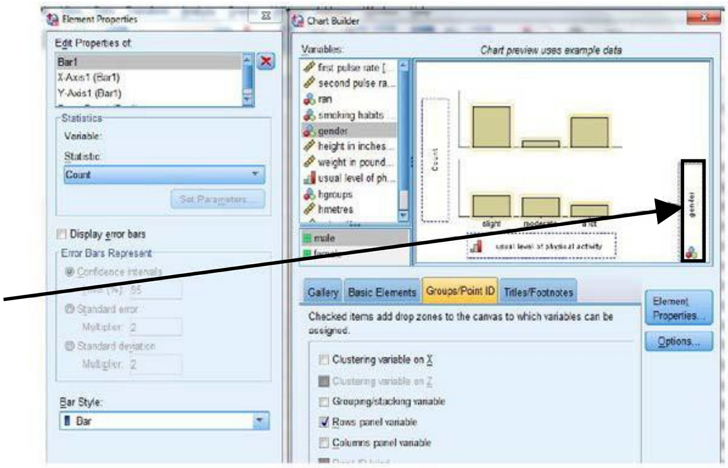
- ***Drawing a panel bar chart***

  This can be done using either the Chart Builder or Legacy Dialogs

  Panel plots are a style of plot in which subgroups of the data are plotted on separate axes alongside or above and below each other, with the scale on the axes kept common. These can be very useful plots for comparing different subgroups.

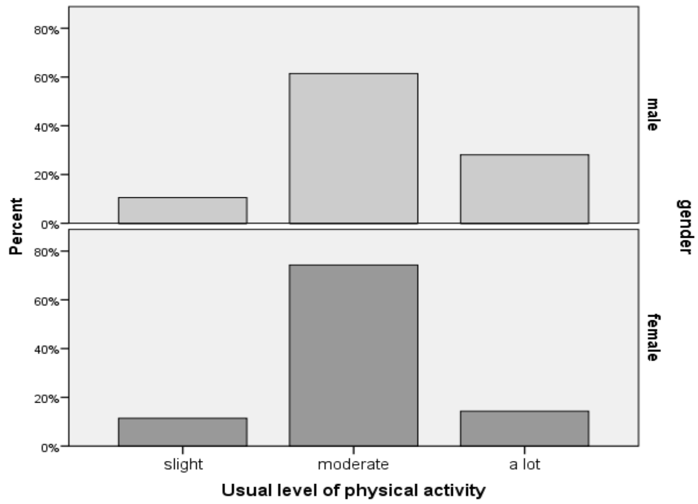
  To produce a panel bar chart of physical activity by gender

  - Use **Graphs > Chart Builder**
  - **Reset**
  - Drag the Bar Chart option from the Gallery into the window
  - Drag **usual level of physical activity** into the **x axis** box
  - Click on the **Groups/Point ID** tab and check **Rows panel variable**
  - Drag gender into the new Panel box
  - **OK**



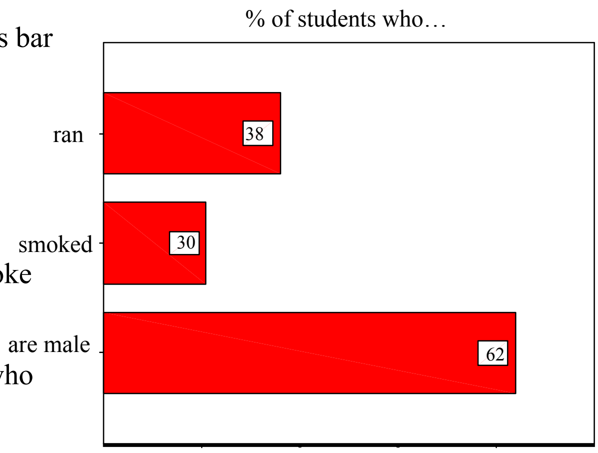
  With some editing, as before, you should be able to get the chart shown on the next page.

  Note: The panel option is available with many of the charts, and can be generated in columns if preferred.



5.3 Drawing a bar chart of more than one variable

Here's another type of bar chart. The difference in this bar chart is that each bar represents a different **variable**.



You have to be sensible about the variables you will compare.

We are limited with our data set, so a not very good example is the % of students who ran, the % who smoke and the % who are male.

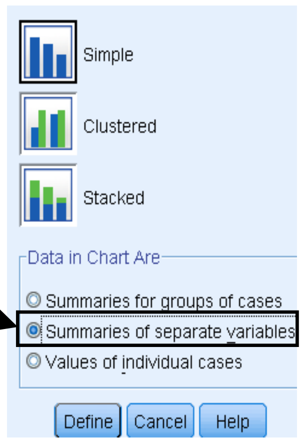
Those who ran are coded 1 in the ran column, those who smoke by a 1 and males by a 1.

To draw the chart use

Graphs >Legacy Dialogs > Bar … Simple

Summaries of separate variables

Define

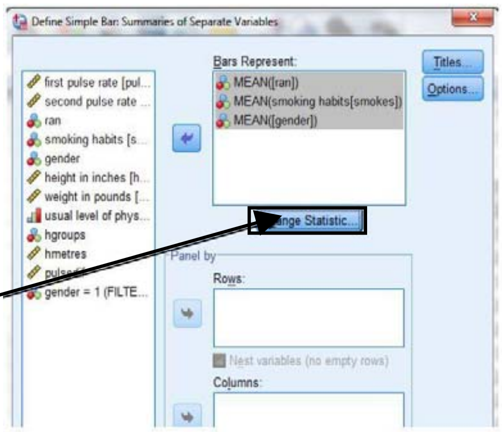


At the next dialogue box place each of the activities in the **Bars Represent** box.

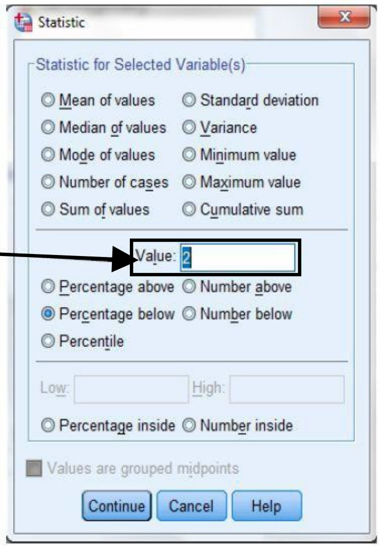
They will show **MEAN()** which we will need to change.

Highlight them all using 'Shift-click' by, or by holding down Ctrl and clicking on each.

Click on **Change Statistic**



We shall ask SPSS to calculate the % of the entries for each variable less than 2 (since 1 < 2)



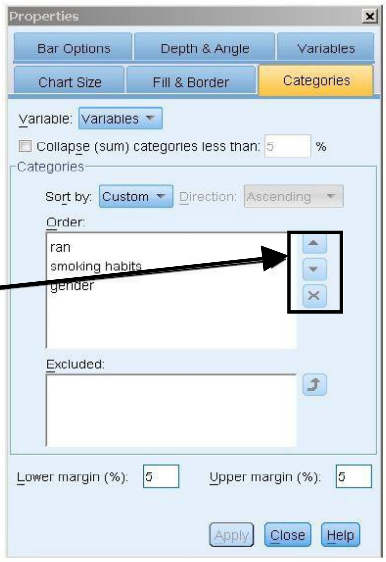
1. Select **Percentage below** and type 2 in the Value box

(If we had only wanted a count we would have asked for **Number below**)

2. **Continue** > **OK**

The resulting bar chart doesn't look quite like the one shown earlier.

By double clicking on it you can open up the **Chart Editor** window and make the necessary alterations.



You can change the order of the bars.

1. First write down the order you want.
2. Double click on a bar to open up the **Properties** dialog box, if not already visible
3. Select the **Categories** tab
4. Highlight the item you want to move under **Order**
5. Click the up or down arrow
6. **Apply**
7. **OK**

5.4 Drawing a pie chart

Pie charts are used to examine parts of a whole. As an example of drawing one in SPSS we shall draw a pie chart of activity level.

You can use **Analyze** > **Descriptive Statistics** > **Frequencies…** and click on the **Charts…** button to ask for a pie chart,

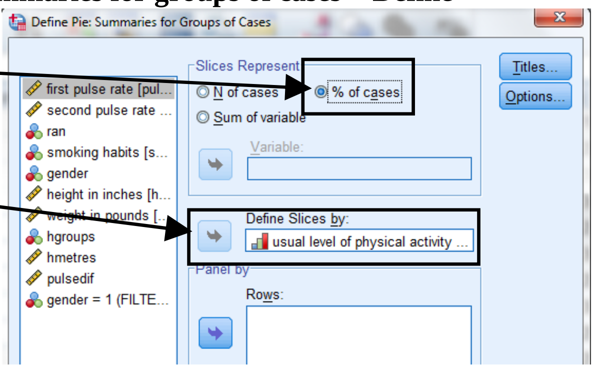
or use **Graphs** > **Chart Builder** selecting the Pie Option.

Or use **Graphs** > **Legacy Dialogs** > **Pie…** > **Summaries for groups of cases** > **Define**

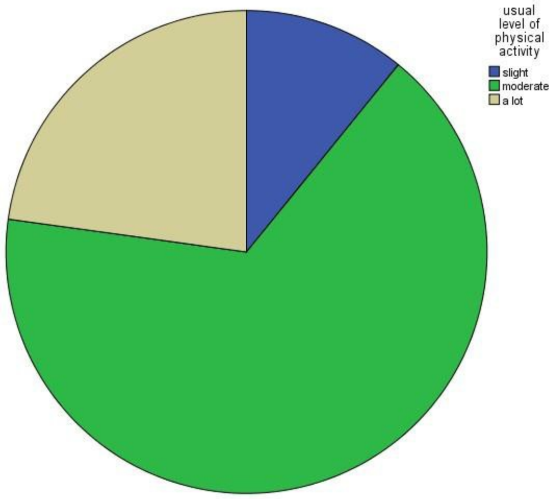
Use % of cases

Define Slices by: **usual level of physical activity**

Click on **OK**



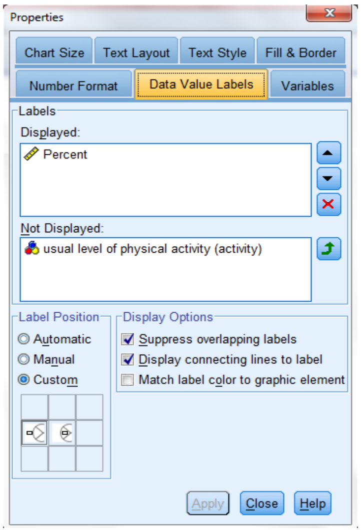
Hopefully you have a similar chart to this or in some other colours.
Open the Editing window. To add % to the labels use:

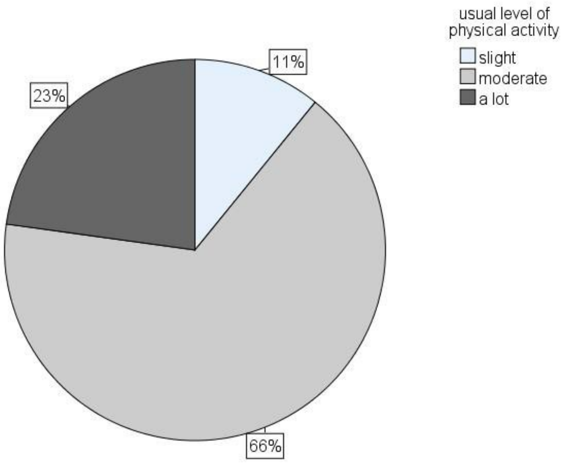


Hide data labels 

This brings up a dialog box but also automatically adds the percentages to the chart.

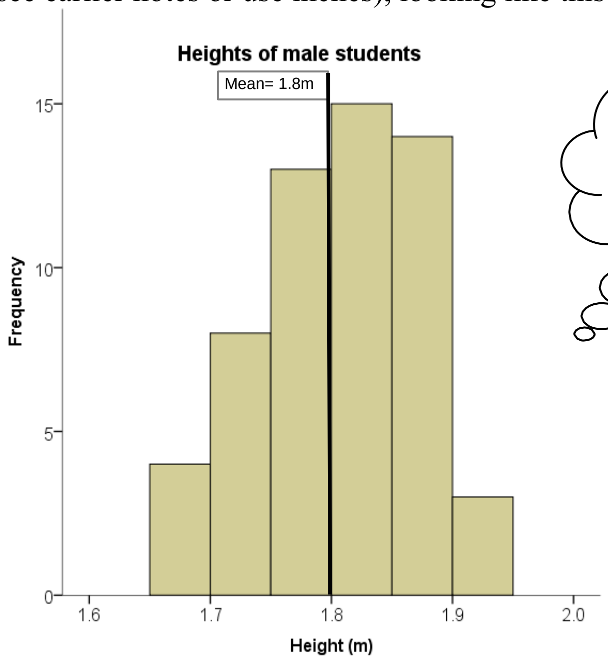
By choosing **the usual level of physical activity** and **Percent**, and selecting the position of the labels you should be able to get the Pie chart shown below.





5.5 Histogram

See if you can produce a histogram of the heights of **male** students, in metres (if you have not kept this variable, see earlier notes or use inches), looking like this:



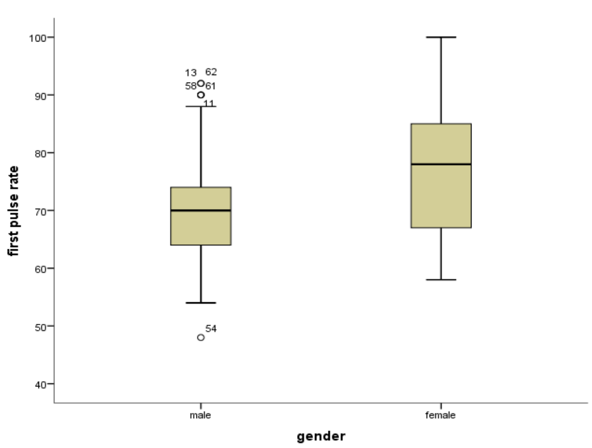
Don't forget to reselect **All** your data again to include the females again!

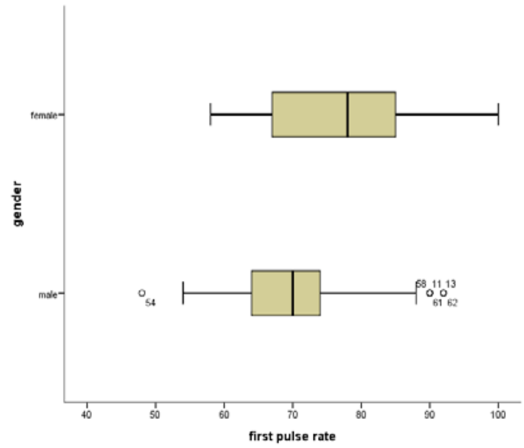
5.6 Boxplots

Boxplots are a useful way of comparing two or more data sets. They are, as the name implies, a box, whose length represents the interquartile range of the data.

The lower edge of the box is at the lower quartile of the data, and the upper edge at the upper quartile. A horizontal line indicates the median.

'Whiskers' are drawn to the minimum and maximum values within 1.5 box-lengths of each end of the box. Outliers are data points further than 1.5 box lengths from the box, but within 3 box lengths. They are indicated by **o**. Values further than 3 box-lengths from the box are indicated by * (not shown here).





The figures above compare the initial pulse rates by gender

Boxplots can also be horizontal

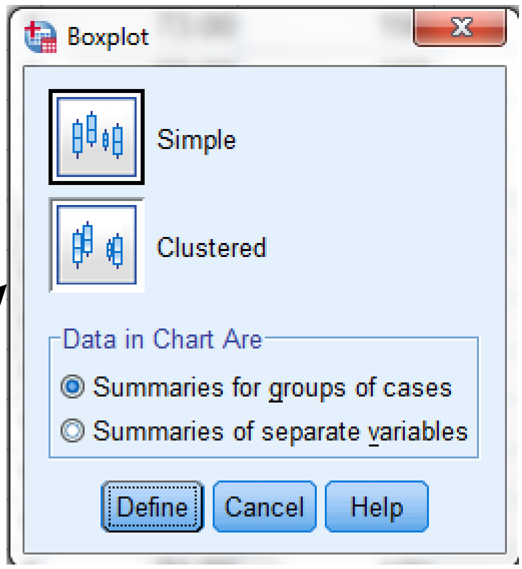
Drawing boxplots can get confusing.

It is easiest to use Explore under **Analyze > Descriptive Statistics**, but here is how to do it using the graphs menu with two examples to illustrate the differences in different types of boxplots.

It is also possible to draw boxplots using the **Chart Builder**, but vertical scaling starts at zero, which you may not want.

We shall draw the boxplots shown above. Use **Graphs > Legacy Dialogs > Boxplot** to obtain the dialogue box.

Select **Simple Summaries for groups of cases > Define**

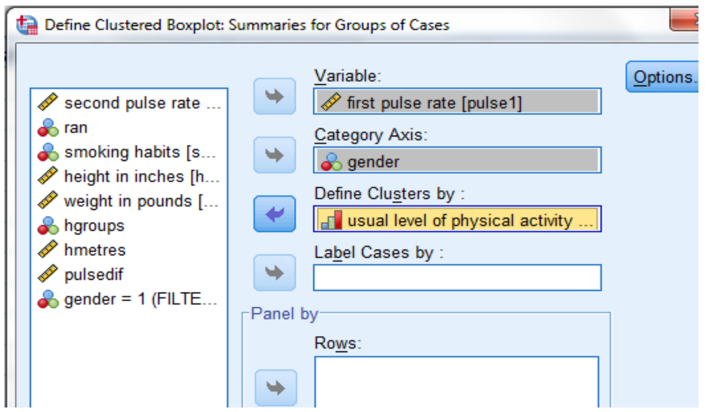


The second example is of a clustered box plot which will show the pulse rates, by gender, for each activity level.

- Use **Graphs > Legacy Dialogs > Boxplot > Clustered**
- **Summaries for groups of cases**
- **Define**

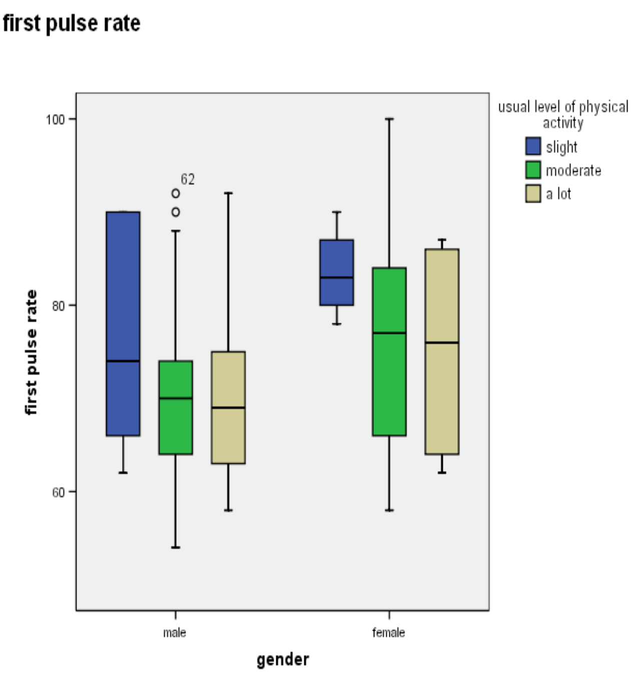
Complete the dialogue box as follows:

- **first pulse rate** in the (top) **Variable** box
- **gender** in the **Category Axis** box
- **usual level of physical activity** in the **Define Clusters by** box.
- **OK**



The result will be a colourful version of the following.

As remarked earlier, printing in black and white can lose the detail in coloured charts.
This is an example of that, e.g. it is difficult to find the median bar in some boxplots, so it is a good idea to change the shadings



There is only one way to master chart drawing in SPSS and that is by having plenty of practice – so over to you.

6 Statistical tests

6.1 Overview

All the following tests are from the **pulse.sav** file located:

W:\EC\STUDENT\SIGMA Statistics Workshops\Workshop 4

The following notes show you how to do a variety of statistical tests using SPSS. This is not an introduction to the theory of statistical testing. For the theory you will need to refer to an appropriate statistics text book, but an outline to the procedure is as follows:

- For any test the first requirement is that you know what you are testing: the hypothesis with reference to the population you are studying
- It is usual to test the Null Hypothesis which is a statement of no difference; no association
- Look at the data - what does the evidence of the sample suggest? Make a chart if possible
- Select an appropriate test
- Check that the requirements for that test have been satisfied; e.g. was the sample a random sample?
- Carry out the test and identify the P-value
- Is the P- value >= 0.05, or < 0.05?

Table of P-values and Significance

Probability	**P**	**Significance**	**Decision**
Less than 1 in 1000	< 0.001	**Significant** at 0.1% level	**Reject** null hypothesis
Less than 1 in 100	< 0.01	**Significant** at 1% level	**Reject** null hypothesis
Less than 5 in 100	<0.05	**Significant** at 5% level	**Reject** null hypothesis
More than or equal to 5 in 100	≥0.05	**Not significant**	**Don't reject** null hypothesis

- Decide if the evidence supports the null hypothesis.
- State the decision about the original hypothesis.

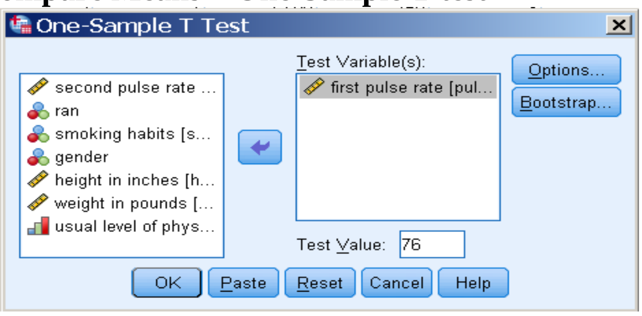
6.2 One-sample T test

The requirement for this test is that the sample has been randomly selected

This is an alternative method to using confidence intervals. Use this to test for a hypothesised value.

E.g. Test the hypothesis that the mean first pulse rate is 76 beats per minute.

Use **Analyze > Compare Means > One-Sample T test**



Place the **first pulse rate** in the **Test Variable** Box and **76** in the **Test Value** box. The output is:

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
first pulse rate	92	72.87	11.009	1.148

One-Sample Test

	Test Value = 76					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
first pulse rate	-2.727	91	.008	-3.130	-5.41	-.85

The significance value of **0.008** shows that there is a significant difference between the hypothesised value of 76 and the mean first pulse rate of the sample

6.3 Confidence intervals

The requirement for this test is that the sample has been randomly selected

Use **Analyze > Descriptive Statistic > Explore**

To find the confidence interval for the mean of a sample; use this to test for a hypothesised value. **E.g.** the following is the Output for the first pulse rate (pulse1) of all students:

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

The confidence interval would support any hypothesis which suggested that the population mean was between the Lower Bound of 70.59 and the Upper Bound of 75.15.

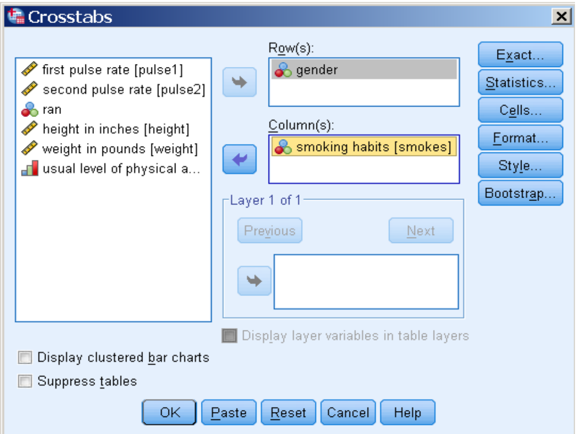
6.4 Chi-Squared Test for contingency tables

The requirements for this test are that the samples are random and at least 80% of the cells in the table should have expected counts of at least 5 and no cell should have an expected count less than 1

The question: Is there an association between smoking and gender?
The Research Hypothesis: There is an association between smoking and gender.
The Null Hypothesis: There is no association between smoking and gender.

Use **Analyze > Descriptive Statistics > Crosstabs**

Complete the dialogue box as shown



Use the **Statistics** button to click in **Chi-Squared** (top left box) **Continue** and the **Cells** button for **Counts: Observed, Expected > Continue > OK**

This should bring up the following Output. By looking at the table of expected and observed counts one can see that there is not much difference (the 'eyeball' test).

gender * smoking habits Crosstabulation

			smoking habits		
			smokes regularly	does not smoke	Total
gender	male	Count	20	37	57
		Expected Count	17.3	39.7	57.0
	female	Count	8	27	35
		Expected Count	10.7	24.3	35.0
Total		Count	28	64	92
		Expected Count	28.0	64.0	92.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.532[a]	1	.216		
Continuity Correction[b]	1.009	1	.315		
Likelihood Ratio	1.570	1	.210		
Fisher's Exact Test				.250	.158
Linear-by-Linear Association	1.515	1	.218		
N of Valid Cases	92				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 10.65.

b. Computed only for a 2x2 table

So it comes as no great surprise that the value of Chi-squared (**1.532**) is not significant because the P-value is **0.216**

The null hypothesis is not rejected.

The conclusion is that this sample offers no evidence for an association between smoking and gender.

In this particular case of 2 rows x 2 columns, it is usual to use the P-value from Fisher's Exact Test version of the chi-squared test results; hence, the p-value quoted would be 0.250 rather than 0.216. It is usually bigger than the basic chi-squared p-value.

6.5 Paired samples t-test

The requirement for this test is that the sample is randomly selected. There is no need for the underlying population to be normal provided the sample size is large, i.e. >30.

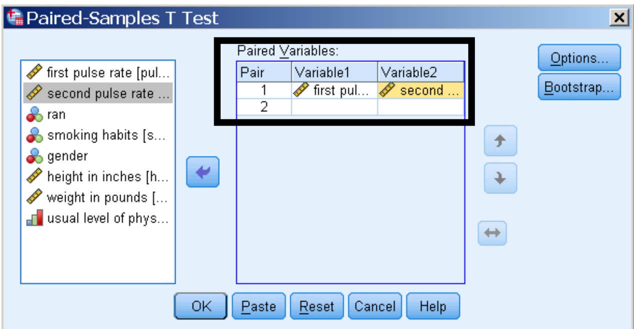
The question: Is there a difference in the population means of the first and second pulse rates of those students who ran on the spot for a minute?

The Research Hypothesis: There is a difference in the population means of the first and second pulse rates of those students who ran on the spot.

The Null Hypothesis: There is no difference in the population means of the first and second pulse rates of those students who ran on the spot.

We are comparing the differences between pairs of readings that are related: the 2 pulse readings from the same student.

Select only those students who ran. **Data** > **Select Cases** > **If ….** > **ran =1**



Use **Analyze** > **Compare Means** > **Paired-Samples T Test**

Complete the dialog box by clicking on **pulse1**, clicking on the arrow and then **pulse2** to place them in the variables box.

OK

You should obtain the following Output:

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	first pulse rate	73.60	35	11.436	1.933
	second pulse rate	92.51	35	18.943	3.202

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	first pulse rate & second pulse rate	35	.607	.000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	first pulse rate - second pulse rate	-18.914	15.050	2.544	-24.084	-13.745	-7.435	34	.000

By looking at the sample means one can see they are very different (the 'eyeball' test again) and indeed the P-value is <**0.001** showing that the t value is significant.

The null hypothesis is rejected.

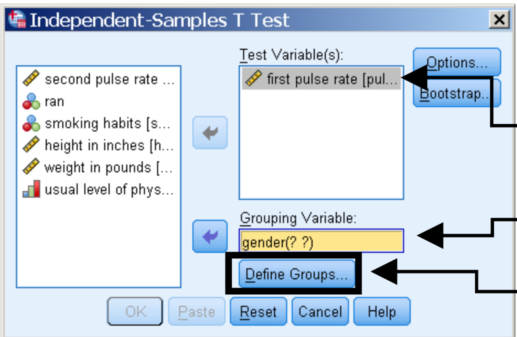
The conclusion is that this sample shows there is a statistically significant difference between the population means of the first and second pulse rates of those students who ran. And, furthermore, the mean pulse rate has increased by 18.9 (SD=15.1)

6.6 Independent samples t-test

The requirement for this test is that the samples are randomly selected. There is no need for the underlying population to be normal provided the sample sizes are large, i.e. >30

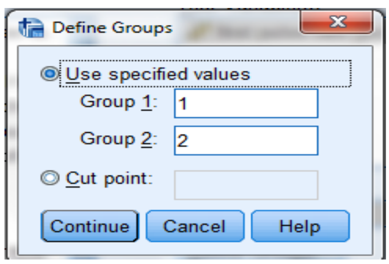
The question:	Is there a difference in the population means of the first pulse rates of males and females?
The Research Hypothesis:	There is a difference in the population means of the first pulse rates of males and females.
The Null Hypothesis:	There is no difference in the population means of the first pulse rates of males and females.

We are comparing the differences between pairs of readings that are not related - one set of pulse rates is from the males and the other is from the females.



Use Analyze > Compare Means > Independent-Samples t Test

1. Place **pulse1** in the Test Variable box
2. **gender** in the **Grouping Variable**
3. Click on **Define Groups**.
4. Fill out the box as shown.
5. The 1 and 2 are the codes for males and females. You should get the following Output (which is annoyingly wide).



Again the eyeball test, looking at the means, reveals a difference in the sample **means**.

Group Statistics

	gender	N	Mean	Std. Deviation	Std. Error Mean
first pulse rate	male	57	70.42	9.948	1.318
	female	35	76.86	11.617	1.964

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
first pulse rate	Equal variances assumed	2.284	.134	-2.825	90	.006	-6.436	2.278	-10.962	-1.910
	Equal variances not assumed			-2.722	63.675	.008	-6.436	2.365	-11.161	-1.712

Levene's test indicates, by the P-value, whether we should assume equal or unequal variances. If the **P-value is > 0.05** the evidence suggests that the variances can be assumed equal.

Here **p=0.134** so we use the Equal variances line for the t test for the means, which gives a low P-value.

We conclude that the samples show that there is a statistically significant difference between the population means of the pulse rates of male and females ($p < 0.05$).

You should then describe this difference in more detail.................... And think about whether it is an important practical difference in the context of your research.

6.7 One-way ANOVA

We are assuming here that we have independent simple random samples drawn Analysis of variance is a method for comparing the means of several populations. Simple random samples are drawn from each and are used to test the null hypothesis that the population means are all equal. ANOVA compares the variation among groups with the variation within groups.

The question: For men, is there a difference in the population means of the first pulse rate for each activity level?

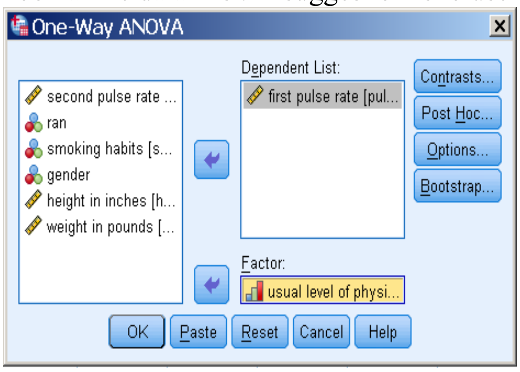
The Research Hypothesis: For men, there is a difference in the population means of the first pulse rate for each activity level.

The Null Hypothesis: For men there is no difference in the population means of the first pulse rate for each activity level.

Select only the men

Use **Data Select Cases If …. gender =1**

Look at the data first. A suggestion is to use Boxplots, to 'visualise' the pattern of data



Use **Analyze > Compare Means > One-Way ANOVA**

Fill out the dialog box as shown with the **first pulse rate** in the **Dependent List**, and usual level physical activity as the **Factor**.

Click on the **Options** button and select **Descriptive Statistics & Homogeneity of Varience Test**

The Output is shown below

Descriptives

first pulse rate

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
slight	6	71.67	16.513	6.741	54.34	89.00	48	90
moderate	35	70.17	9.310	1.574	66.97	73.37	54	92
a lot	16	70.50	9.018	2.255	65.69	75.31	58	92
Total	57	70.42	9.948	1.318	67.78	73.06	48	92

Test of Homogeneity of Variances

first pulse rate

Levene Statistic	df1	df2	Sig.
2.586	2	54	.085

ANOVA

first pulse rate

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	11.590	2	5.795	.057	.945
Within Groups	5530.305	54	102.413		
Total	5541.895	56			

The result of the Levene's test (p=0.085) suggests that we can assume equal variances, even though the 'slight' group is more variable than the other 2 groups – see table of Descriptives above, and compare SDs.

The p-value is 0.945 which is >0.05, so we conclude that there is no evidence to suggest that that the mean “first pulse rate” of the 3 exercise groups are not all the same.

Don’t forget to reselect **All** your data again!

6.8 Non-Parametric tests

A parameter is a number describing the population, e.g. the mean or standard deviation, as distinct from a statistic which is a number that can be calculated from the sample data without needing to know anything else about the population.

Many statistical tests are parametric tests and make the assumption that the populations involved have 'normal distribution'. These tests are very useful and robust but there are occasions when we would like to compare two samples which we cannot assume come from a 'normal' population, or where the measurements are on an ordinal scale as distinct from an interval one.

For such populations we use non-parametric tests. We can use these on 'normal' data too.

Note: If the values in the population have a skewed distribution or if the measurement scale is ordinal then it is better to use the median rather than the mean.

6.9 Wilcoxon and Mann Whitney tests

The Question: Is there a difference in the population mean/median of the two pulse rates of those students who ran on the spot for a minute?

The Research Hypothesis: There is a difference in the population mean/median of the two pulse rates of those students who ran on the spot.

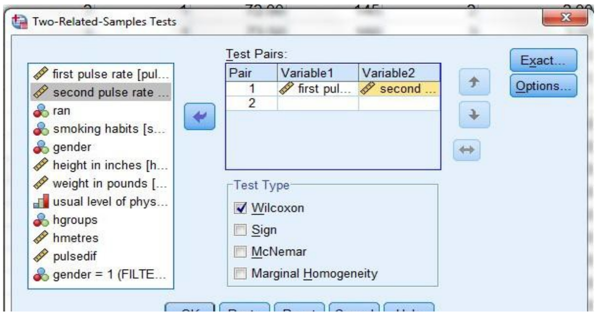
The Null Hypothesis: There is no difference in the population mean/median of the two pulse rates of those students who ran on the spot.

We are comparing the differences between pairs of readings that are related: the 2 pulse readings from the same student.

Select only those students who ran. **Data Select Cases If …. ran =1**

Use **Analyze > Nonparametric Tests > Legacy Dialogs > 2 Related Samples**

Complete the dialogue box
By placing both **pulse1** and **pulse2** in the Test Pairs box and ticking the Wilcoxon box.



Wilcoxon Signed Ranks Test.
The Negative Ranks refer to where **Pulse2** is less than **Pulse1**. The Positive Ranks are those where **Pulse2** is greater than **Pulse1**. Ties are where **Pulse2** equals **Pulse1**.

Ranks

		N	Mean Rank	Sum of Ranks
second pulse rate - first pulse rate	Negative Ranks	25[a]	24.04	601.00
	Positive Ranks	50[b]	44.98	2249.00
	Ties	17[c]		
	Total	92		

a. second pulse rate < first pulse rate
b. second pulse rate > first pulse rate
c. second pulse rate = first pulse rate

Test Statistics[a]

	second pulse rate - first pulse rate
Z	-4.368[b]
Asymp. Sig. (2-tailed)	.000

a. Wilcoxon Signed Ranks Test
b. Based on negative ranks.

The p-value is given as **0.000** which means it is less than **0.001**, indicating that we should reject the null hypothesis.

The conclusion is that there is a difference in the two pulse rates of those who ran.

Don't forget to reselect **All** your data again!

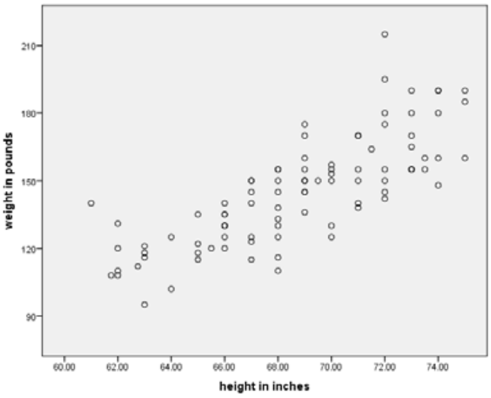
6.10 Pearson's correlation coefficient

The Question:	Is there a **LINEAR** relationship between the heights and weights of students?
The Research Hypothesis:	There is a **LINEAR** relationship between the heights and weights of students
The Null Hypothesis:	There is no **LINEAR** relationship between the heights and weights of students

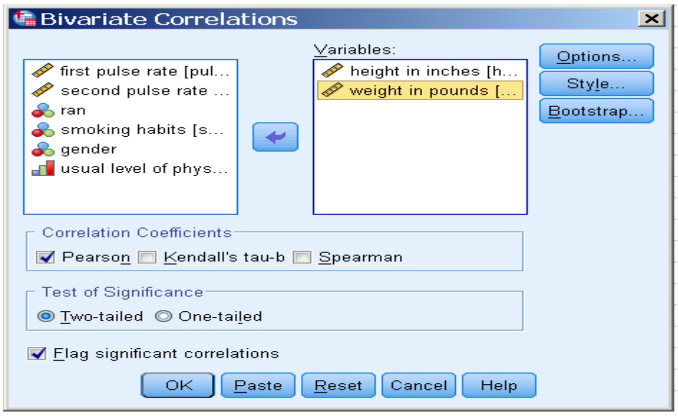
Firstly we look at the data, using a scatterplot.

Use **Graphs > Legacy Dialogs > Scatter/Dot…> Simple Scatter > Define**
Place weight on the Y axis and height on the X axis. Click **OK**.

There does appear to be a linear relationship, so it makes sense to find the correlation coefficient as a measure of the strength of this linear relationship, using **Analyze > Correlate > Bivariate**



Complete the dialogue box as shown with height and weight in the **Variables** box, tick the box for **Pearson**:



The Output should be:
The correlation coefficient between height and weight is **r = 0.79**

There are **92** in the sample. The p-value is **0.000 (<0.001)**

This indicates that if there is no linear association between height and weight there is a less than **0.1%** chance that a random sample of **92** would give a Correlation coefficient as extreme as **0.785**.

Correlations

		height in inches	weight in pounds
height in inches	Pearson Correlation	1	.785**
	Sig. (2-tailed)		.000
	N	92	92
weight in pounds	Pearson Correlation	.785**	1
	Sig. (2-tailed)	.000	
	N	92	92

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

In conclusion, there is evidence to suggest a strong linear relationship between students' height and weight. NB Could this be extrapolated to adults?

6.11 *Spearman's rank correlation coefficient*

To obtain Spearman's Rank Correlation Coefficient use the same procedure checking the square for Spearman in the dialog box.

Using the same data as before the output gives:

Correlations

			height in inches	weight in pounds
Spearman's rho	height in inches	Correlation Coefficient	1.000	.819**
		Sig. (2-tailed)	.	.000
		N	92	92
	weight in pounds	Correlation Coefficient	.819**	1.000
		Sig. (2-tailed)	.000	.
		N	92	92

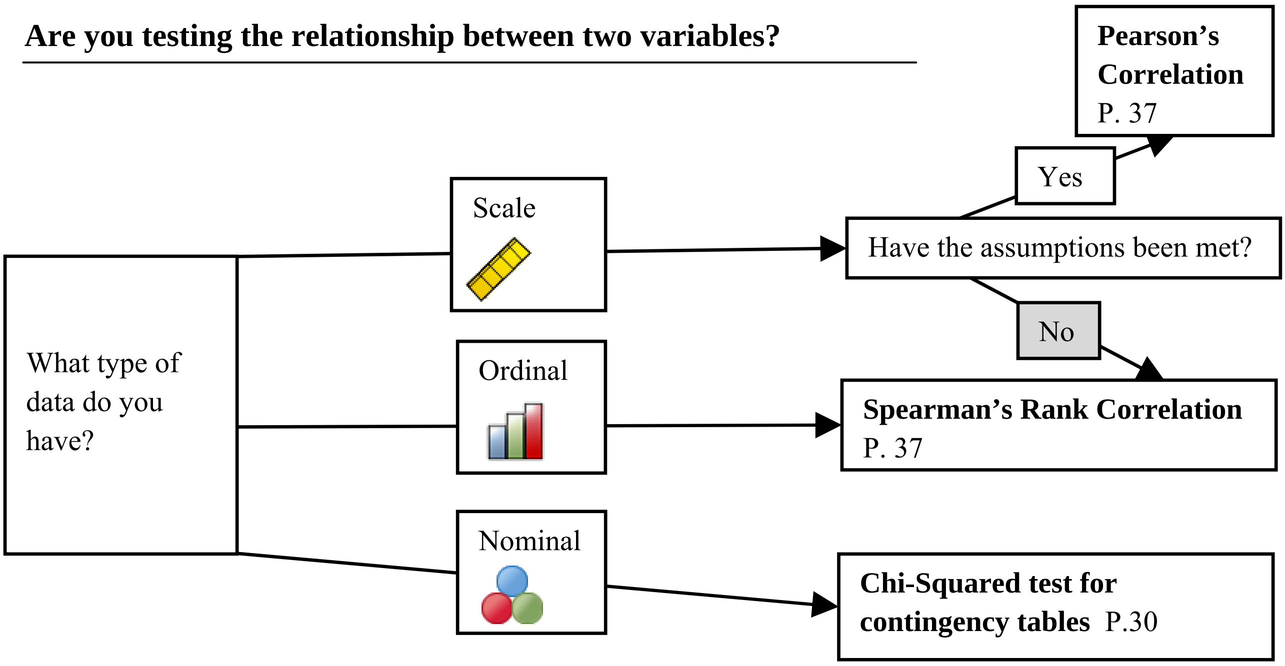
**. Correlation is significant at the 0.01 level (2-tailed).

The correlation coefficient is a slightly different value, **0.819**, having been computed differently, but the significance level (p-value) is again listed as **0.000**.

N.B. Remember that, when the significance, or p-value, is given as 0.000 it doesn't mean that it is zero, but we ususlly write it as P<0.001.

7 Flow chart for Statistical tests

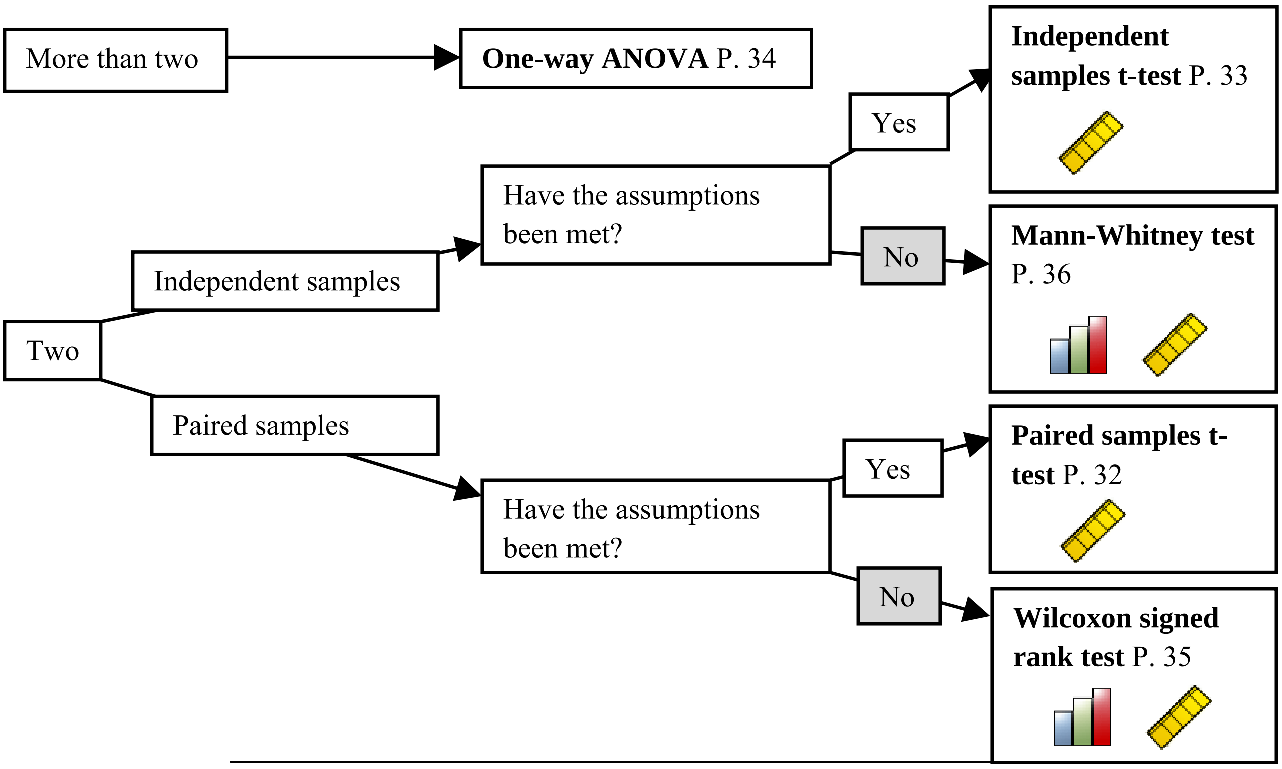
Are you testing the relationship between two variables?



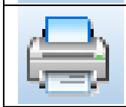
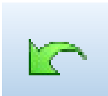
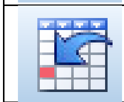
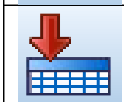
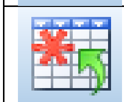
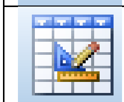
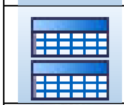
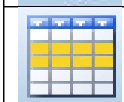
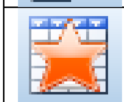
Do you want to compare one sample mean against a known value?

One sample t-test
P. 29

Do you want to compare the means from two samples or more than two samples?



8 Appendix 1: Functions

Symbol	Function
	Open a saved data file/saved output file
	Save the current data file/current output file
	Print the output/data file
	Toggle between recent dialog boxes
	Undoes the last operation
	Redoes the last operation
	Finds a specific case
	Finds a specific column
	Inserts a new case
	Inserts a new column
	Splitfile, splits the data into two different sets
	Weights the data file
	Select cases, select different subgroups of your data
	Value label, turn on the data labels
	Gold star, to go back to the data file from the output file
	Edit your output

9 Appendix 2: Recommendations for keying questionnaire data into Excel (then to be imported into SPSS)

Code these as 1 or 2

Example Questionnaire

Code these as 1, 2, 3 or 4

1. Gender: Male☐ Female ☐

2. Ethnicity: White British ☐ Black British ☐
   Asian British ☐ Other *(please specify)* ☐ ____________

3. How old are you? _______years

Code as separate columns, 1 if ticked, 0 otherwise

4. Which of the following do you like (tick **all** that apply)?
   Beer ☐ Chocolate ☐ Gin ☐
   Pizza ☐ Curry ☐

5. How far do you agree with the statement "Men shouldn't be allowed to teach children" (tick **one** only)?

   Strongly disagree☐ Disagree☐ Neutral ☐ Agree☐ Strongly agree☐

Code as 1, 2, 3 ,4 or 5

Example Excel file (for the above)

Only one cell for the heading

i.d.	Q1gender	Q2ethnicity	Q2other	Q3age	Q4a	Q4b	Q4c	Q4d	Q4e	Q5
1	1	1		54	1	1	0	1	1	1
2	1	2		48	1	1	1	0	1	2
3	2	1			0	1	0	0	0	4
4	2	1		62	1	1	0	1	0	3
5	1	4	Chinese	37	0	0	1	1	0	3

Leave blank if missing

•Where possible, try to enter all data as numbers, text values can be added in SPSS

•Have only one cell for column heading

•Have only one worksheet per Excel file

•Multiple answers require a separate column for each choice (e.g. Q4 above), enter as 1 if ticked, 0 otherwise

•Leave cells empty for missing data