# Contents

## General notation

- Single click (assume left button unless otherwise stated)
- Double left click
- > Followed by
- Type

## 1 Opening StatsDirect

To get into StatsDirect:

Either

- Start > All Programs > StatsDirect > 


Or

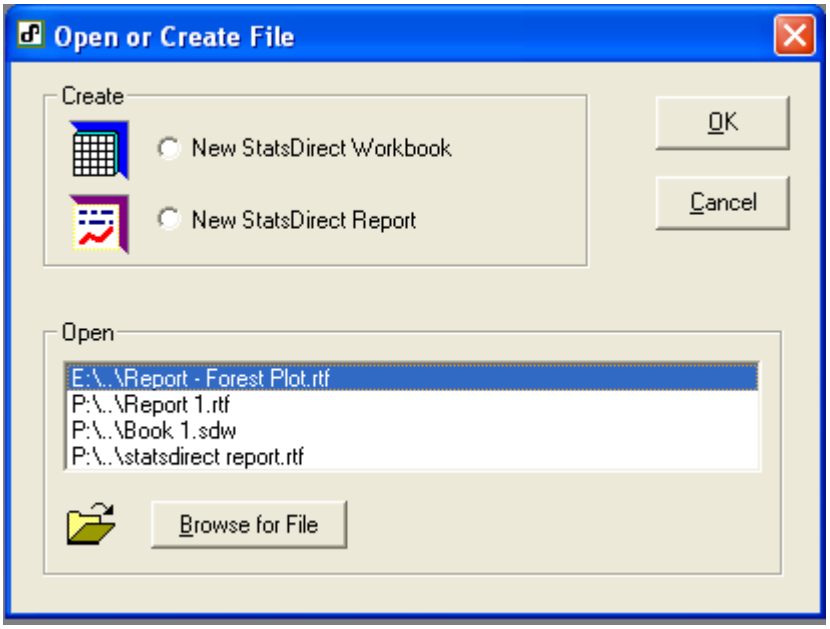
Double click on the StatsDirect icon 

on your desktop.

## 2 Layout of StatsDirect

### 2.1 Screens

StatsDirect has two windows, the *Workbook* and the *Report*. When you open StatsDirect a window similar to the following will appear:

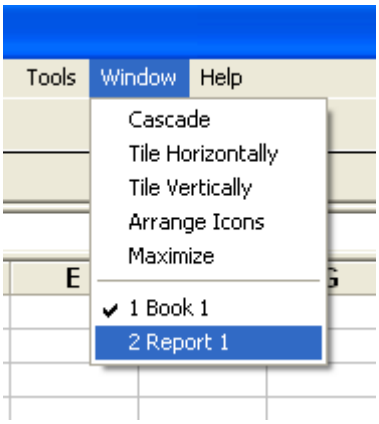


Here you can choose to open a new workbook or report (under *Create*) or open an existing one (under *Open*).

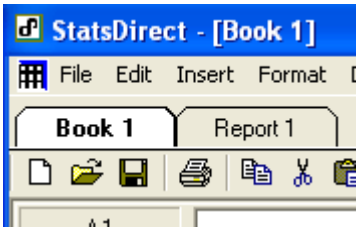
A *Workbook* looks like an *Excel* spreadsheet. Data entry, manipulation, plotting graphs and statistical analyses are all done from the *Workbook*. A *Report* is like a *Word* document and contains a record of all the output created as a result of a procedure you have carried out.
You can have more than one report and workbook open at any time.

To move from one screen to another either:

1) Use the Window menu and highlighting the window that you want to move to. For example, if you are in *Book 1* and wish to move to *Report 1*:




2) Use the tabs at the top of the screen:




## 2.2 Saving files

Because the two screens contain two different files, they are saved separately. Depending on the active screen, go to *File* menu then click *Save Workbook As…* or *Save Report As…*, as applicable:

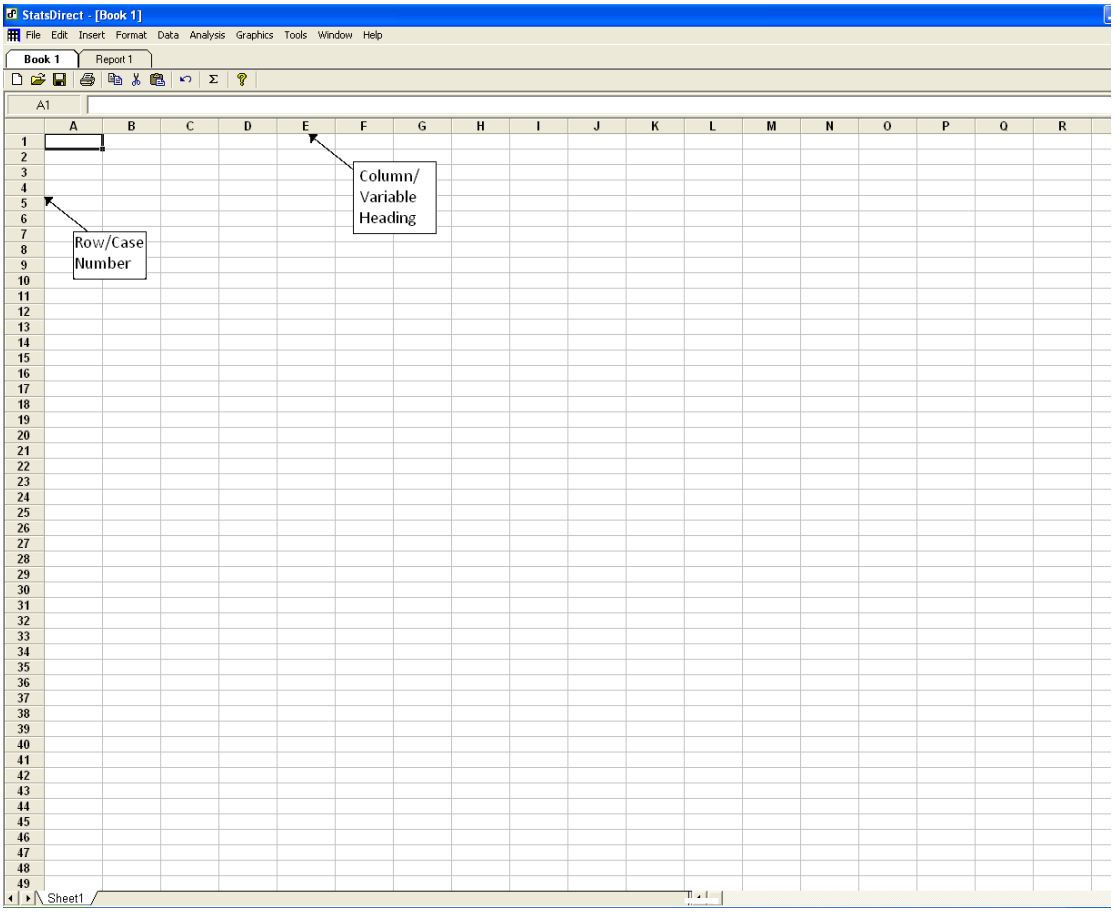
- *Workbook* files are saved with a ".sdw" file extension/ending
- *Report* files are saved with a ".rtf" file extension/ending. This means that they can be re-opened in *Word* or another word processor.

# 3 Creating a data file

## 3.1 General

### 3.1.1 Layout

A new workbook will look like this:




A typical way to enter data is to have one line (row) per subject and one column per variable.

### 3.1.2 Selecting columns

To select/highlight a column:

- Grey column heading

### 3.1.3 Selecting rows

To select/highlight a row:

- Grey row number

3.1.4 Selecting the dataset

To select/highlight all the worksheet:

- Grey square in the top left hand corner above row 1 and to the left of column A

## 3.2 Setting up the data file

3.2.1 Naming columns

- Grey column heading
- Variable name in the window that appears.

NB. If you are familiar with Excel please note that although the window looks like Excel, you should not enter column headings in row 1.

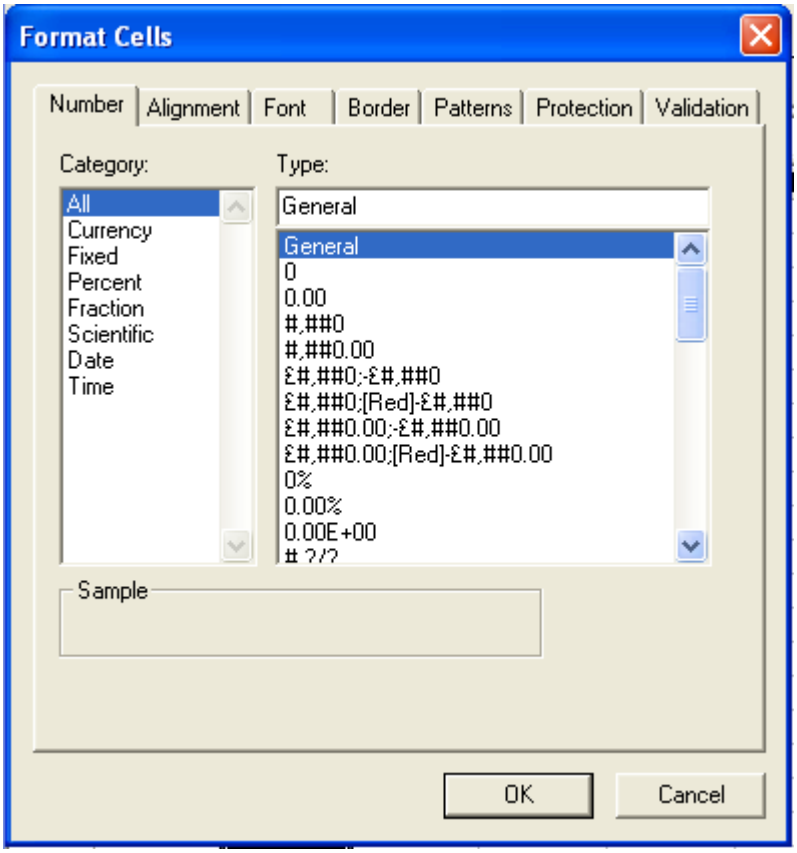
3.2.2 Changing the width of a column

- Use the mouse to move the cursor to the edge of right hand side of the column heading so that the  cursor changes to a  .
- Drag until you get the width you want.

3.2.3 Changing the number of decimal places of a numeric variable

- Highlight the column you wish to change.
- *Format* > *Cells…*
- *Number* tab.
- *All* under *Category:*
- Change *Type:* to *0* for 0 decimal places, *0.0* for one decimal place, *0.00* for 2 decimal places, and so on. If the format is not already there in the menu underneath you willl have to type it in.

To display decimal places the window should look as below:




### 3.2.4 Entering a date/time

- Highlight the column you want to enter a date or time into
- *Format menu > Cells...*
- *Number* tab
- *Date* under *Category:*
- Under *Type*: select the format you want.

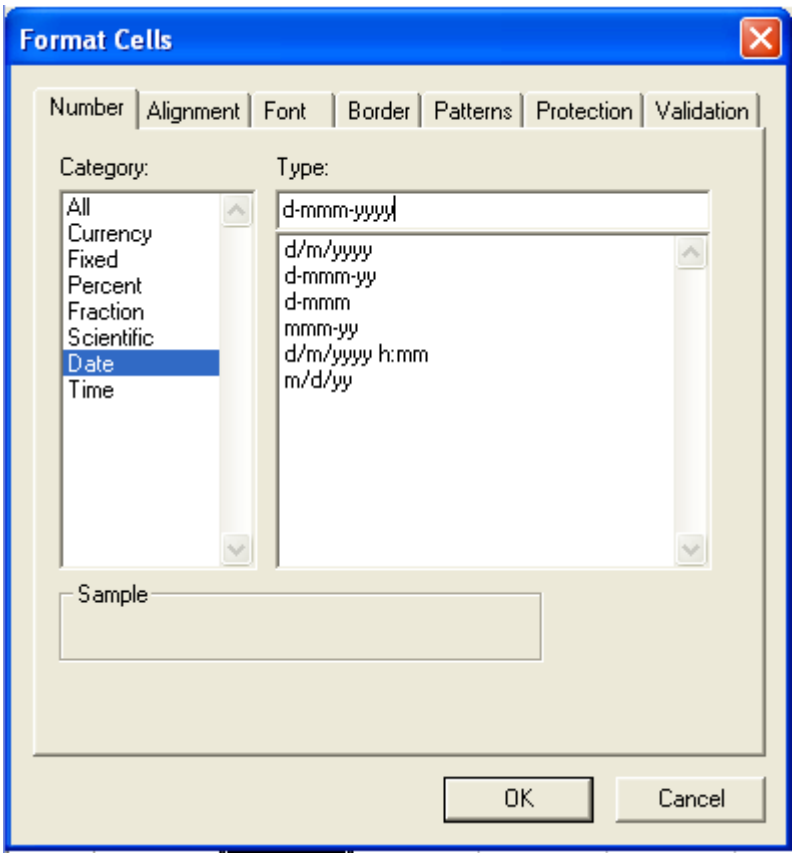
'd' stands for day, 'm' for month (displayed numerically), 'mmm' for month displayed as a 3 letter abbreviation, 'yy' for two digit year and 'yyyy' four digit year. It is good practice to use the full four digit year 'yyyy'.

If necessary you can create your own format.

For example, to set a date in the format '12-Mar-1997'

⌨ “d-mmm-yyyy” in *Type:*
🖱 OK.

The window should look as below:




**Warning: Be careful when entering dates as numbers. The best method is to create format ‘d-mmm-yyyy’ and type in “2-aug-1976”; if you type “2-8-1976” it may change the order to the american way of dating, i.e. ‘8-Feb-1976’.**

### 3.2.5 Entering a categorical variable

It is good practice to enter as much data as possible as a number even if it is a code for something else. This is because it is so easy to make data entry errors when typing in words. (Entering month as a word is seen an exception for the reason given above).

For example, for a variable indicating the sex of a patient, enter ‘0’ to indicate ‘male’ and ‘1’ for ‘female’.

Once you have entered your data, create a copy of the variable (see section 4.1) and use *Search & Replace* in the copied column to replace each number with the appropriate word (see section 5.3).

### 3.2.6 Missing values

If a response is missing, rather than leaving a cell blank, it is good practice to enter another value so that you know that it is actually missing rather than a data entry error. The best missing value code to use in StatsDirect is a full stop "."

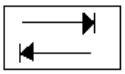
## 3.3 Entering data

- Cell you wish to start entering data into.
- The desired value for that cell.

It's helpful to use the numbers on the right hand side of your keyboard; if the *Num Lock* light 

is not lit on your keyboard, press the 

key.

Press ENTER to move down to the cell below or the TAB key  to move across to the cell on the right. The arrow keys also work.

To correct a mis-entered value, click the cell, the entry that is currently in the cell will be displayed in the white box above the grey column headings.

To correct a mis-entered value, click in the cell. The entry that is currently in the cell will be displayed in the white box above the grey column headings. You may either amend the value in this white box, or enter the new value directly into the cell itself; the old value will automatically be deleted.

# 4 Data manipulation

## 4.1 Editing

### 4.1.1 *Cut*

Deletes a column, row or cell(s), making the contents available to paste

- Select a column, row or cell(s)
- *Edit* > *Cut* (or right-click *'Cut'*)

### 4.1.2 *Copy*

Copies a column, row or cell(s)

- Select a column, row or cell(s)
- *Edit* > *Copy* (or right-click *'Copy'*)

### *4.1.3 Paste*

Pastes a selection that has been copied or cut (will start in top left-hand corner of selection)

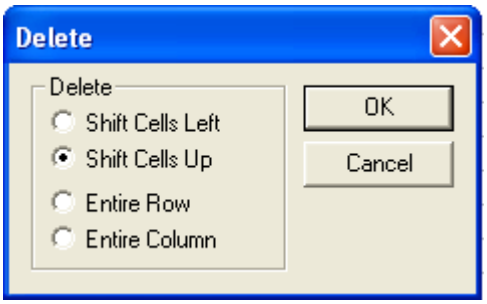
- Select (empty) column, row or cell(s). (If not empty, will replace current values).
- *Edit* > *Paste* (or right-click *'Paste'*)

### 4.1.1 Delete column, row or cell(s)

Deletes a column, row or cell(s) – no copy made

- Select column, row or cell(s)
- *Edit > Delete Column* **or** *Edit > Delete Row* **or** *Edit > Delete Special* (as applicable)

If deleting cells (*Delete Special*) the following window will appear:




To replace deleted cells with cells from the left

- *Shift Cells Left*

To replace with those from below

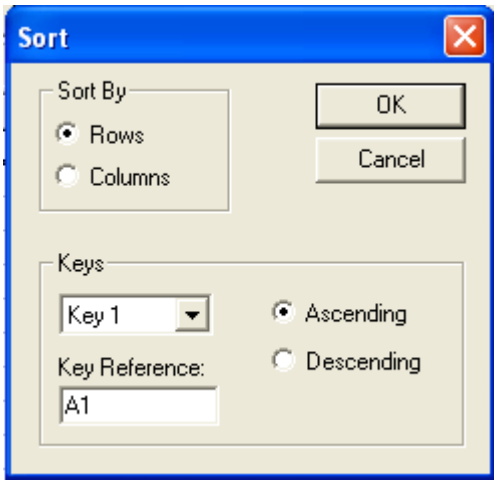
*Shift Cells Up*

### 4.2 Sorting

The dataset can be sorted according to the order of one or more variables. *Ascending* sorts from smallest to largest and *Descending* from largest to smallest.

- Highlight the dataset (see section 3.1.4)
- *Data > Sort > Manipulate Worksheet*

The following window will appear:

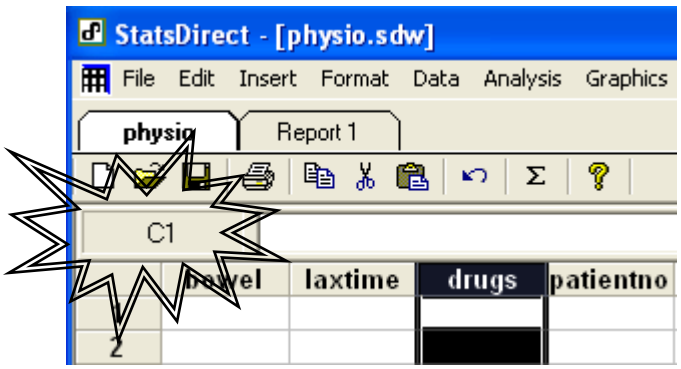



- Leave *Sort By* as *Rows*

*Keys* tells StatsDirect which variable you wish to sort the dataset by. *Key 1* is the first variable you wish to sort by. *Key Reference* is based on the original label that the column had. For example, if you wish to sort by 'Group' in the 3rd column, you would set *Key 1* to be *C1* as above.

If you wish to sort the data by more than one variable, this can be done by putting the *Key Reference* for the second variable in *Key 2*, the third in *Key 3* and so on.

Note: Before you sort, if you click in the column you want to sort by, the column reference is displayed in the top left hand corner:



# 5 Creating new variables from existing ones

## 5.1 Applying functions to numbers and dates

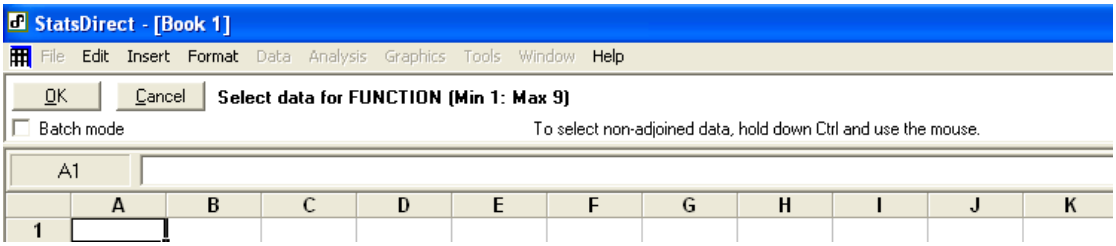
You can use variables already entered to create a new one.

> *Note: If you are familiar with Excel – you can put a formula straight into a cell as you would in Excel.*

Otherwise:

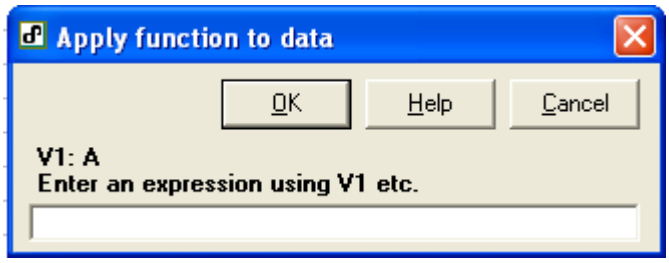
- *Data > Apply Function*

The top of the screen will change to say:




- Select the columns you want to use by clicking on the top of the column. To select more the one hold down the *Ctrl* key inbetween clicking.
- *OK*

A screen similar to below should appear:



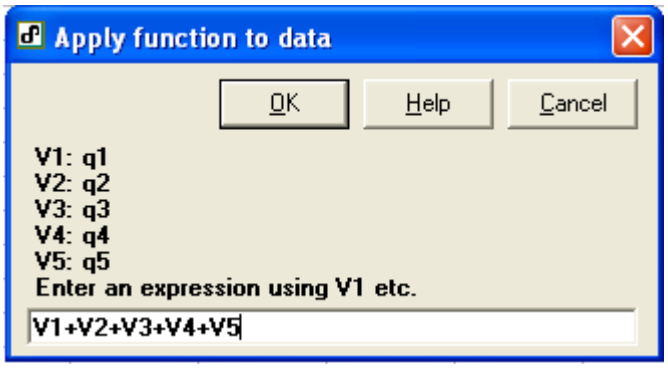

- Enter the calculation for the variable you wish to produce.
- In the new screen that appears, select the workbook you want the new variable to be put in (probably the one you're already in).
- *Select*
- In the next screen, type in the desired name for the new variable.
- *OK*

Example 1: Creating a variable that contains the total number of positive responses in a five question questionnaire, for each case.

For each question (q1, q2, q3, q4, q5) a positive response was stored as 1 and a negative as 0.

- *Data > Apply Function*
- Select *q1* to *q5 > OK*
- "v1+v2+v3+v4+v5" into the *Enter an expression using V1 etc.* box

The *Apply function to data* box should look as follows:



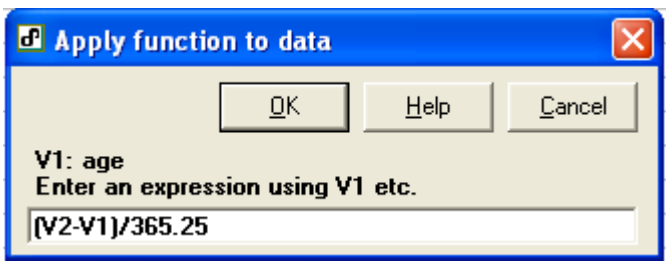

- Worksheet your dataset is in > *Select*
- "total" as new column heading.
- *OK*

Example 2: Calculating age

Date of birth (dob) and date of operation (dofop) are already on the database, wish to calculate age at operation:

- *Data > Apply Function*
- Select *dob* and *dofop*
- "(V2-V1)/365.25"

the *Apply function to data* box should look as follows:




- Worksheet your dataset is in > *Select*
- "age" as new column heading

- *OK*

### 5.2 Recoding categories

The function that we use to do this is meant for coding text to a number, but you can also use it to recode numbers.

- *Data > Text to Numbers*
- Column you wish to recode > *OK*

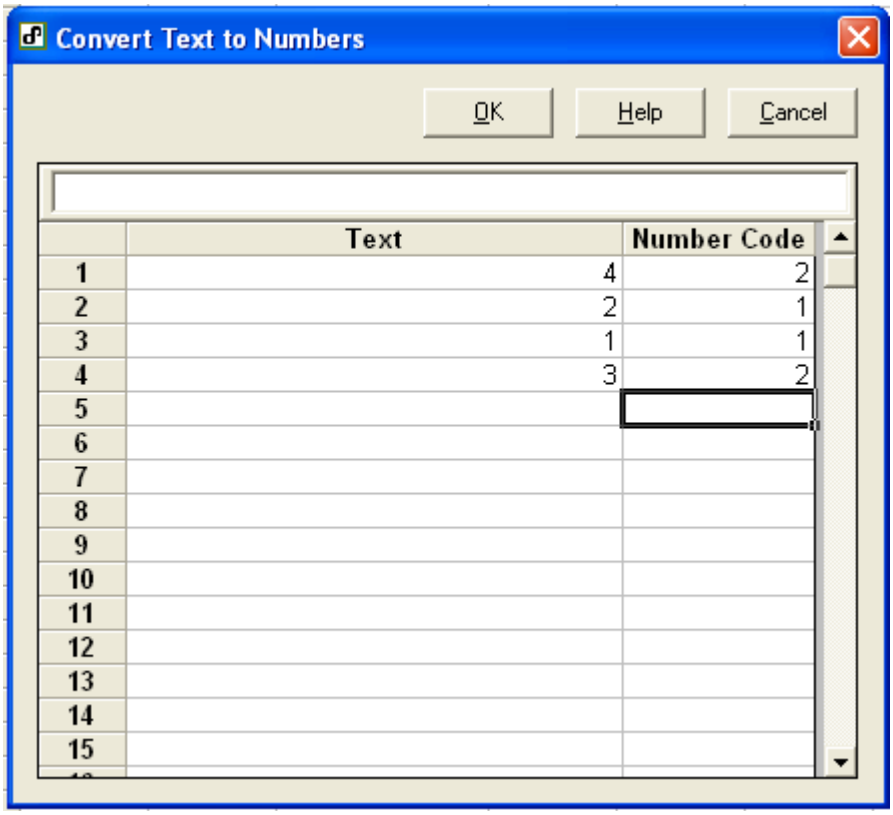
A window will appear asking "Does the top row of your selection contain titles?"

- *No* (you should have titles in the header)

Another window will appear asking "Do you want to specify a numeric code for each different text string?"

- *Yes*

Another window will appear (similar to below). The first column has the codes for the existing variable and in the second column you can tell StatsDirect how you know want them to be coded in the new variable. In the example below there are 4 existing categories (coded '1', '2', '3' and '4'), but the new variable will group categories '1'&'2' together and categories '3'&'4' together.

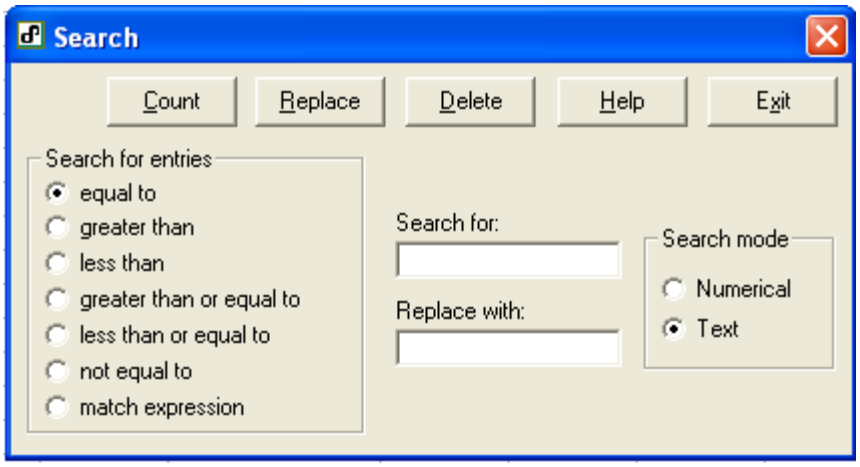



- *OK*
- Worksheet that you are already in > *Select*

### 5.3 Categorising numbers

- Copy the variable to be categorised into a new column.
- *Data* > *Search & Replace*
- Highlight column just pasted > *OK*

The following window will appear




- Under *Search* mode select *Numerical.*
- Use the options under *Search for entries* and *Search for:* box to specify the criteria for replacement of a particular category.
- In the *Replace with:* box type the expression you wish any values satisfying the criteria to become.
- *Replace*

The window will remain open so that you can repeat this method for other categories if necessary.

**Warning**: Categorise the lowest values first and beware of replacing the code for the category you have just created when categorising the rest of the values.

Example: Creating a new variable to distinguish under 50 from over 50 year olds.

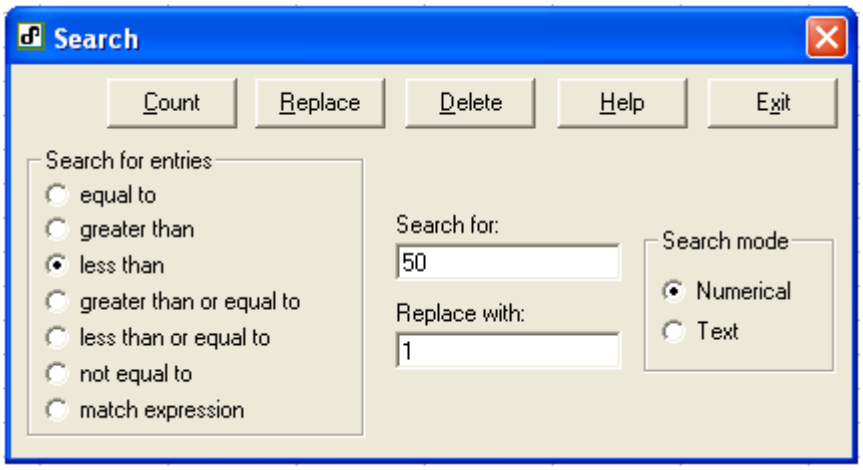
- Copy the column with age in and paste in new column as below:

| o | age | age |
|---|---|---|
| 1 | 54 | 54 |
| 2 | 42 | 42 |
| 3 | 65 | 65 |
| 4 | 31 | 31 |
| 5 | 49 | 49 |
| 6 | 54 | 54 |
| 7 | 36 | 36 |
| 8 | 59 | 59 |
| 9 | 31 | 31 |
| 0 | 53 | 53 |

- *Data > Search & Replace*
- Highlight the copy of the column
- *OK*

- *less than* under *Search for entries*
- "50" in the *Search for:* box
- "1" in the *Replace with:* box

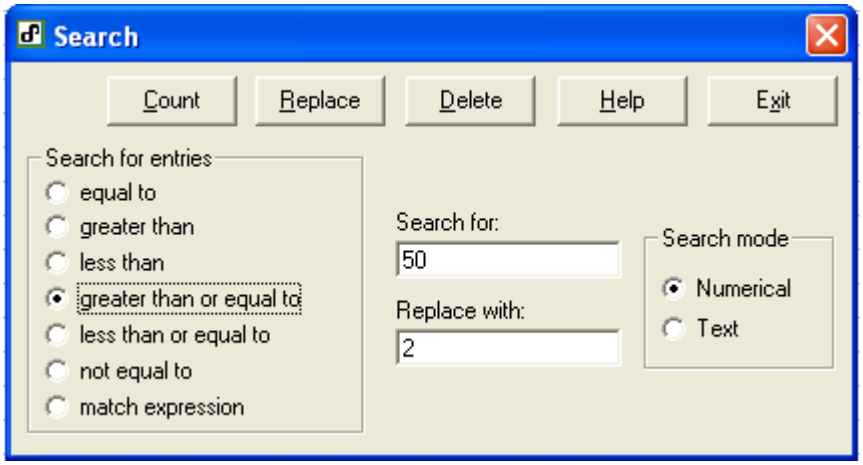
The window should be as below:




- *Replace*

- *greater than or equal to* under *Search for entries*
- "50" in the *Search for:* box
- "2" in the *Replace with:* box

The window should be as below:




*Replace > Exit*

The column should now look as below:

| age | age |
|---|---|
| 54 | 2 |
| 42 | 1 |
| 65 | 2 |
| 31 | 1 |
| 49 | 1 |
| 54 | 2 |
| 36 | 1 |
| 59 | 2 |
| 31 | 1 |
| 53 | 2 |

You should now rename the new column and create a column with labels for the values as described earlier (see section 3.2.5).

NB. If the values of the age variable 'greater than or equal to 50' had been categorised first these values would have become '1', so when categorising the values of the age variable 'less than 50', these '1's would also be included, and hence these values would have been replaced by '2's.

Note: If you'd prefer the categories to be presented in a different order to the one given, sort by that variable first (see section 4.2).

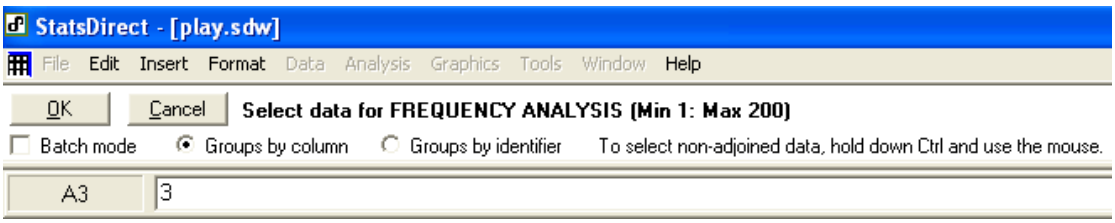
# 6 Frequencies and percentages

These commands should be used when you want to look at categorical data, i.e. counts of subjects. Section 6.1 is for when you're looking at a sample and want to see how many fall in each category, e.g. how many are male and how many female. Section 6.2 is for when you're looking at a sample of two or more groups and want to see how many fall in each category within these groups, e.g. how many are male and how many are female in group 1, and the same for group 2.

## 6.1 For all of the sample

- *Analysis > Frequencies*

The top of the screen will change to the following:




- Select *Groups by column* option
- Grey header of the measure you wish to summarise > *OK*
- Select the appropriate report in *Select report for output* > *Select*

Example: when summarising sex for all of the sample, the following output would be produced:

**Frequencies**

Frequency analysis for sex:

Total = 10

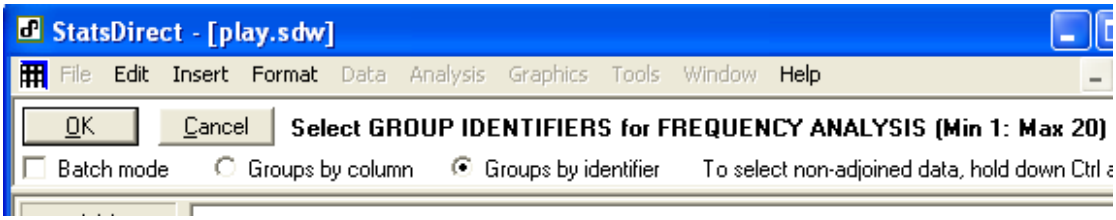
| Value | Frequency | Relative % | Cumulative | Cumulative Relative % |
|---|---|---|---|---|
| 0 | 5 | 50 | 5 | 50 |
| 1 | 5 | 50 | 10 | 100 |

## 6.2 Comparing groups

**Warning: this only works if your outcome data are entered as numeric codes**

- *Analysis > Frequencies*

- Check the *Groups by identifier* option
- Select the grey header of the grouping variable > *OK*




- Select the grey header of variable to summarise
- *OK*
- Select the appropriate report in *Select report for output*
- *Select*

For example: summarising sex for group 1 and group 2 would produce (where female is coded as '1' and male as '0'):

**Frequencies**

Frequency analysis for sex_group_1:

Total = 11

| Value | Frequency | Relative % | Cumulative | Cumulative Relative % |
|---|---|---|---|---|
| 0 | 5 | 45.454545 | 5 | 45.454545 |
| 1 | 6 | 54.545455 | 11 | 100 |

Frequency analysis for sex_group_2:

Total = 9

| Value | Frequency | Relative % | Cumulative | Cumulative Relative % |
|---|---|---|---|---|
| 0 | 2 | 22.222222 | 2 | 22.222222 |
| 1 | 7 | 77.777778 | 9 | 100 |

# 7 Displaying data

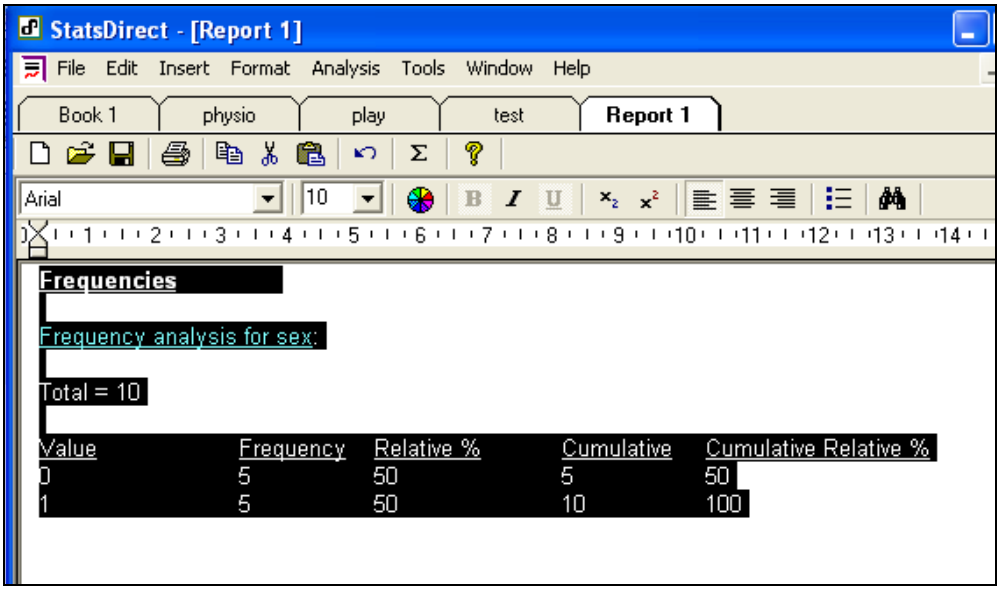
## 7.1 Bar and pie charts

Firstly, you will need to create a frequency table for the data as in chapter 6. If summarising the entire sample go to section **6.1,** if comparing groups go to section **6.2.**

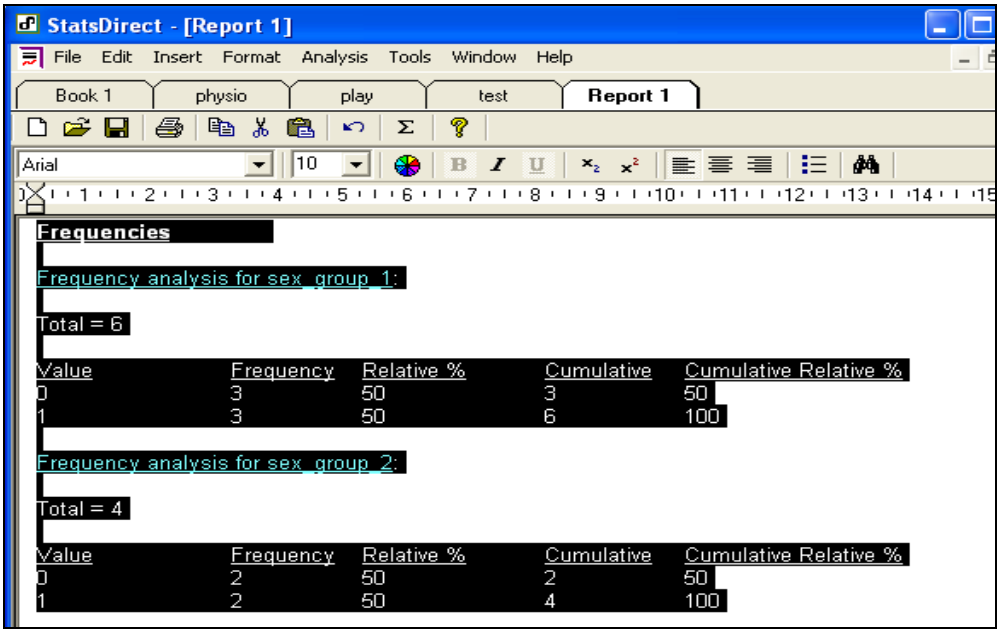
Having got the frequencies displayed in the report, you need to copy them into a data sheet.

- Go to the report with the frequencies in.
- Highlight the table you have created:

If it is a summary for the entire sample it should look like:




And if it is a summary for comparing groups like:

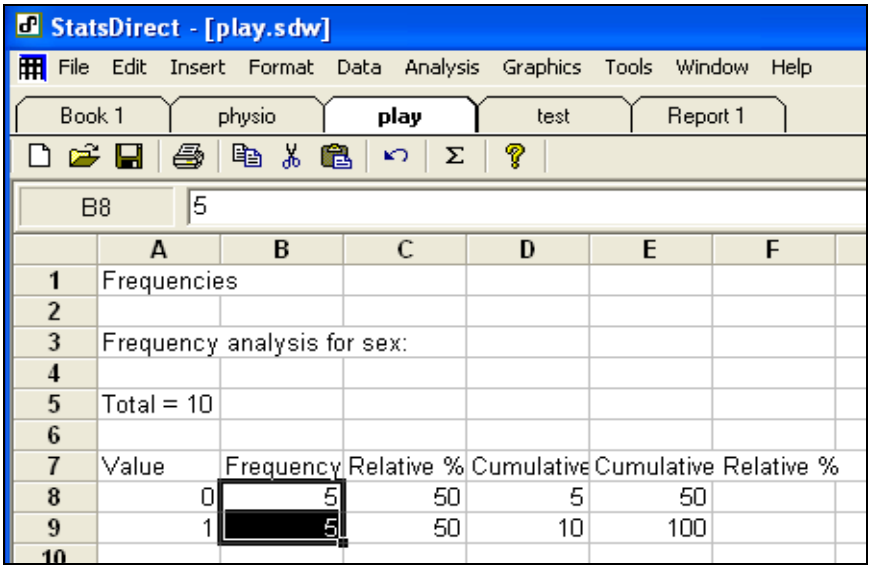



- *Edit > Copy*
- Go back to the dataset you were working from by clicking the appropriate tab at the top of the screen.
- *Insert > Sheet*
- *Edit > Paste*

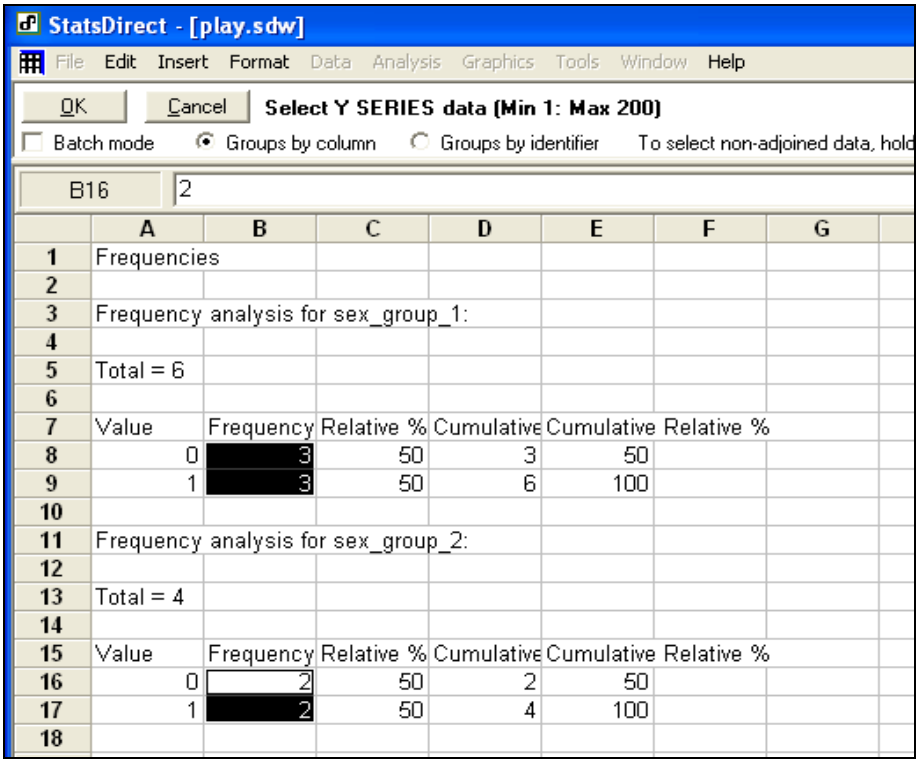
Now create the graph:

- *Graphics > Chart Maker*
- Check *Groups by column*
- Highlight the frequencies

For graph for <u>all sample</u>, as below:



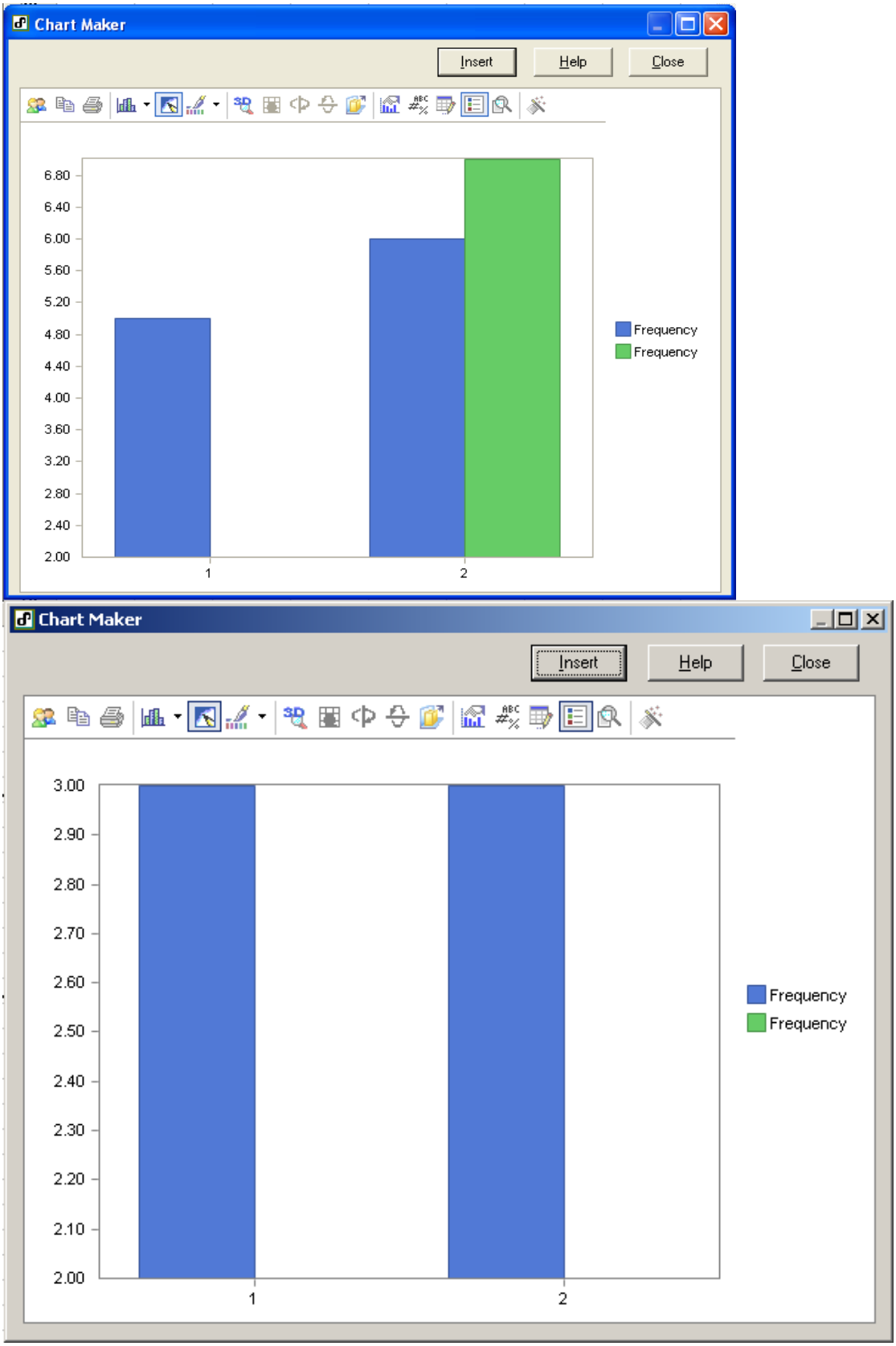
Or if for a graph comparing groups, as below (hold down the CTRL button when trying to make more than one selection at the same time):




| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Frequencies | | | | | | |
| 2 | | | | | | | |
| 3 | Frequency analysis for sex_group_1: | | | | | | |
| 4 | | | | | | | |
| 5 | Total = 6 | | | | | | |
| 6 | | | | | | | |
| 7 | Value | Frequency | Relative % | Cumulative | Cumulative Relative % | | |
| 8 | 0 | 3 | 50 | 3 | 50 | | |
| 9 | 1 | 3 | 50 | 6 | 100 | | |
| 10 | | | | | | | |
| 11 | Frequency analysis for sex_group_2: | | | | | | |
| 12 | | | | | | | |
| 13 | Total = 4 | | | | | | |
| 14 | | | | | | | |
| 15 | Value | Frequency | Relative % | Cumulative | Cumulative Relative % | | |
| 16 | 0 | 2 | 50 | 2 | 50 | | |
| 17 | 1 | 2 | 50 | 4 | 100 | | |
| 18 | | | | | | | |

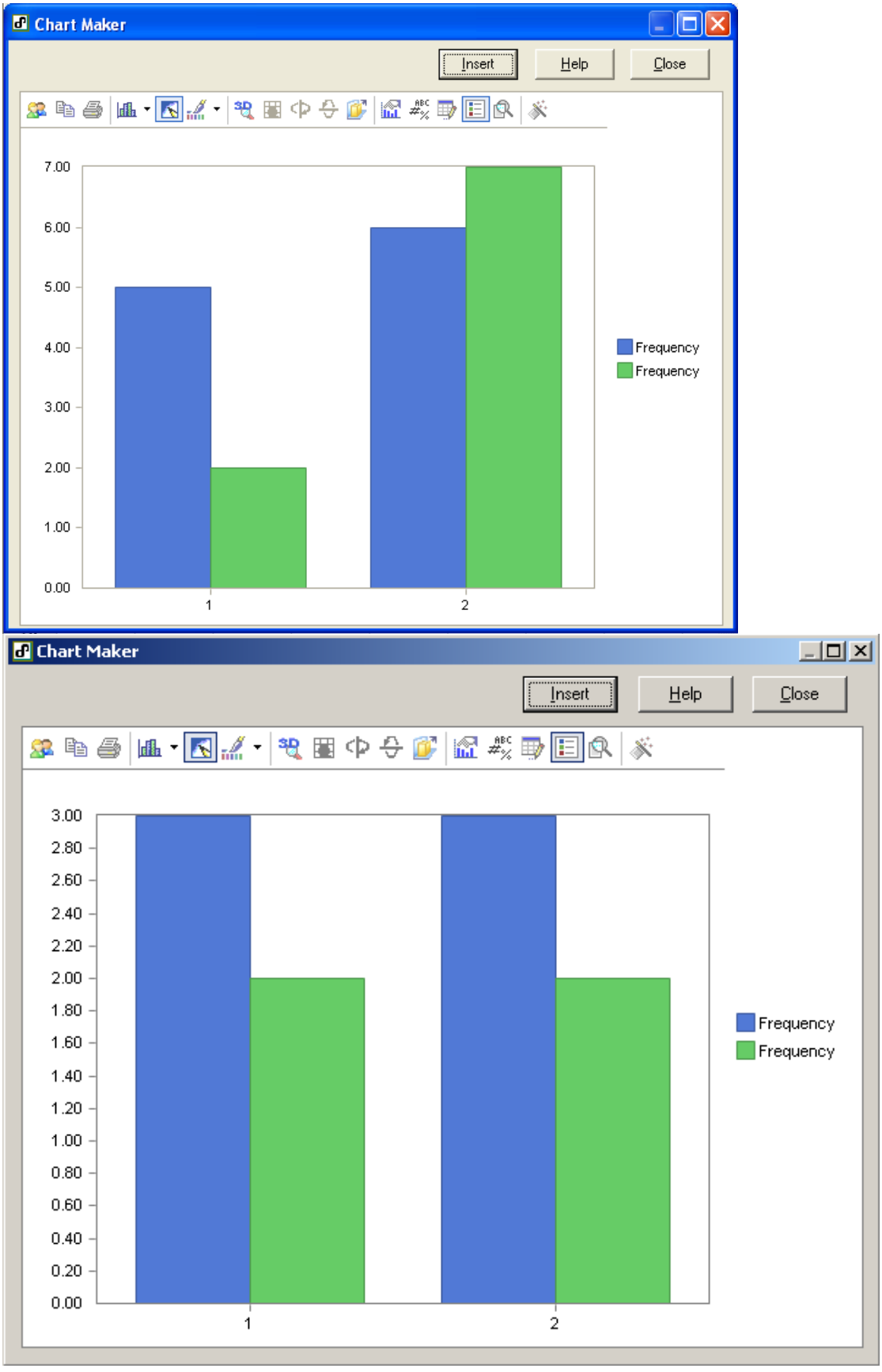
*OK > Cancel > Cancel*

A window similar to the following should appear (though not necessarily with this graph):



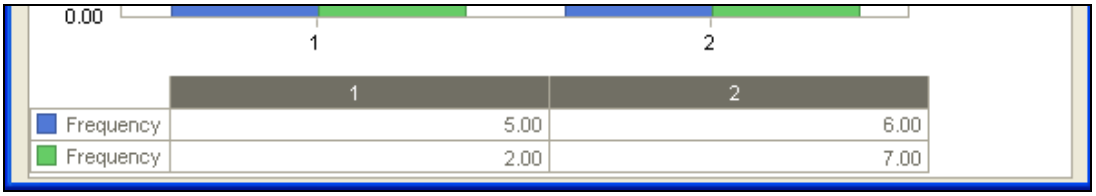

### 7.1.1. Editing the Y-axis

As above, the Y-axis may not begin from 0. This can be amended by clicking on the *Axes Settings* button  then *Options…* setting *Minimum:* to 0 and clicking *OK*. For the example above, after these commands the graph looked as below:




### 7.1.2. Editing the legend

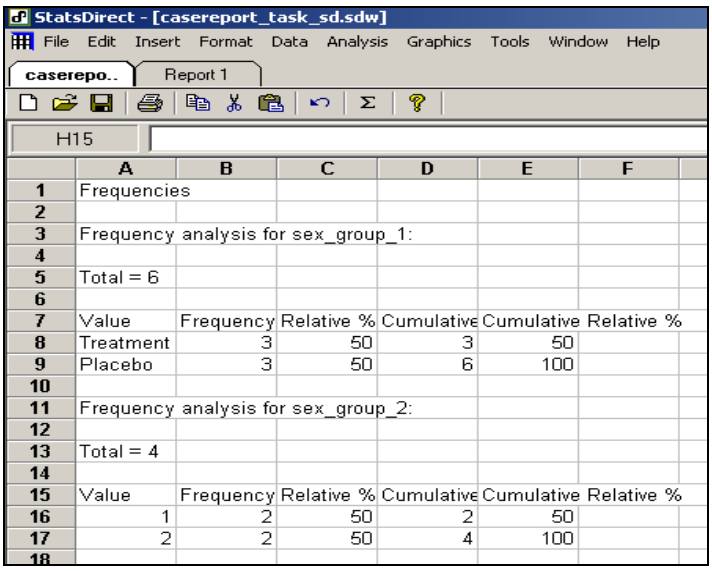
To edit the 'legend' (the portion of the chart which tells what each colour represents) or values of the graph, click the *Data Editor* button  . The bottom part of the chart should look as below:




In this table, you can double click and alter any of the words or values. By pressing ENTER after any change, the change is automatically made on the graph. Once you are done with this table, you can click the *Data Editor* button again to get rid of it.

### 7.1.3. Editing the X-axis

One thing you can't change in *Data Editor* are the labels for the X-axis. This can be done by creating one corresponding column of labels next to a set of your values. For the example above, if you wanted the X-axis to represent, e.g., placebo and treatment groups, you could enter these labels into the spreadsheet as follows:



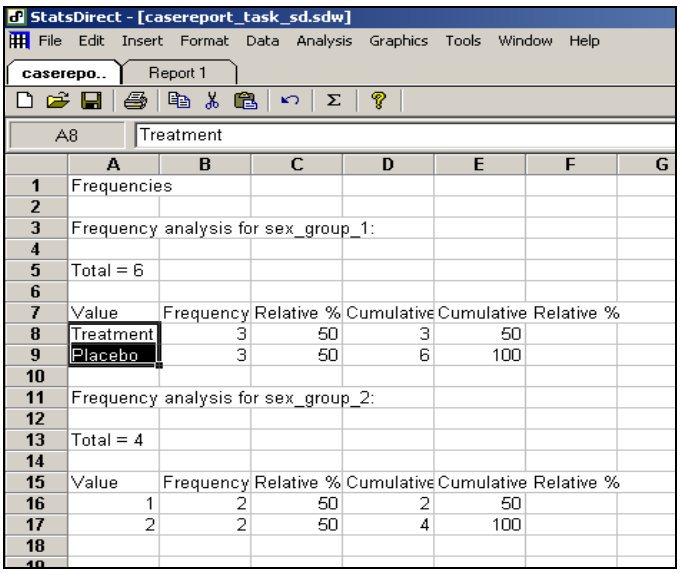

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Frequencies | | | | | |
| 2 | | | | | | |
| 3 | Frequency analysis for sex_group_1: | | | | | |
| 4 | | | | | | |
| 5 | Total = 6 | | | | | |
| 6 | | | | | | |
| 7 | Value | Frequency | Relative % | Cumulative | Cumulative | Relative % |
| 8 | Treatment | 3 | 50 | 3 | 50 | |
| 9 | Placebo | 3 | 50 | 6 | 100 | |
| 10 | | | | | | |
| 11 | Frequency analysis for sex_group_2: | | | | | |
| 12 | | | | | | |
| 13 | Total = 4 | | | | | |
| 14 | | | | | | |
| 15 | Value | Frequency | Relative % | Cumulative | Cumulative | Relative % |
| 16 | 1 | 2 | 50 | 2 | 50 | |
| 17 | 2 | 2 | 50 | 4 | 100 | |
| 18 | | | | | | |

This time do the following:

- *Graphics > Chart Maker*
- Check *Groups by column*
- Highlight the frequencies

*OK*

Highlight the labels 'Treatment' and 'Placebo' as below:

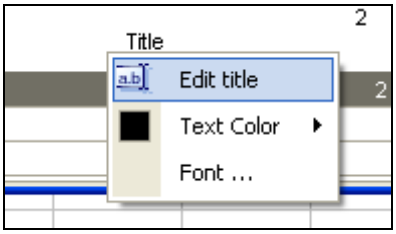



| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Frequencies | | | | | | |
| 2 | | | | | | | |
| 3 | Frequency analysis for sex_group_1: | | | | | | |
| 4 | | | | | | | |
| 5 | Total = 6 | | | | | | |
| 6 | | | | | | | |
| 7 | Value | Frequency | Relative % | Cumulative | Cumulative Relative % | | |
| 8 | Treatment | 3 | 50 | 3 | 50 | | |
| 9 | Placebo | 3 | 50 | 6 | 100 | | |
| 10 | | | | | | | |
| 11 | Frequency analysis for sex_group_2: | | | | | | |
| 12 | | | | | | | |
| 13 | Total = 4 | | | | | | |
| 14 | | | | | | | |
| 15 | Value | Frequency | Relative % | Cumulative | Cumulative Relative % | | |
| 16 | 1 | 2 | 50 | 2 | 50 | | |
| 17 | 2 | 2 | 50 | 4 | 100 | | |
| 18 | | | | | | | |

*OK > Cancel*

### 7.1.3. Adding a title

A title may be given to the chart by right-clicking on the chart and selecting *Edit title* as below:

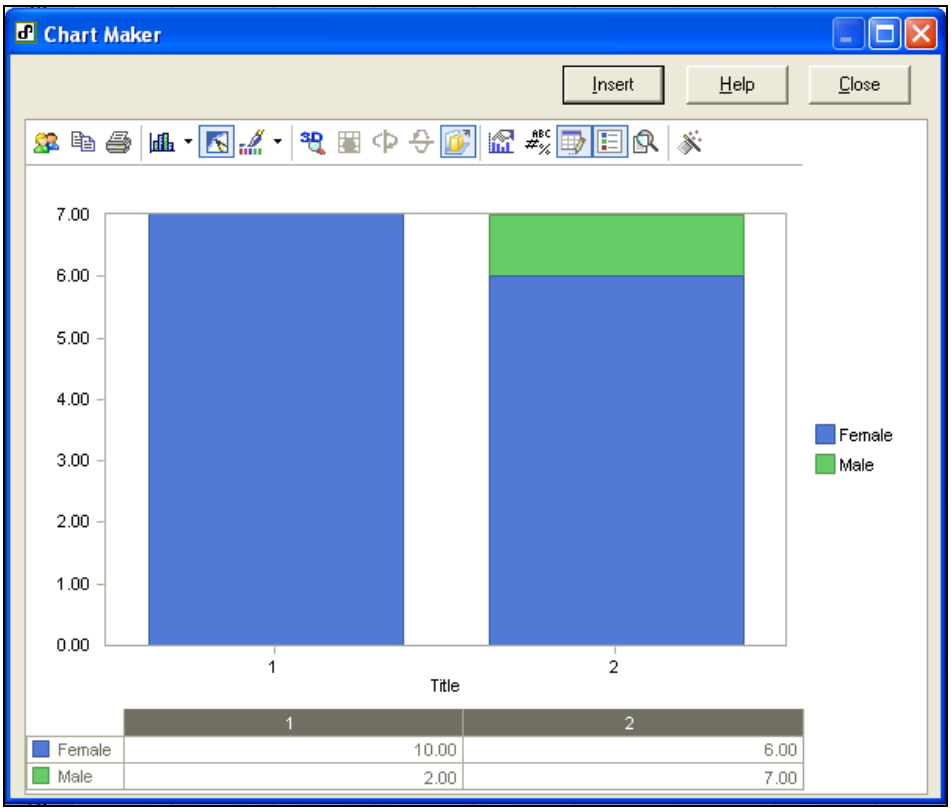



### 7.1.4. Switching between bar chart, pie chart, etc.

The option *Gallery*  allows you to choose between various ways of plotting the same data, such as a pie chart instead.

### 7.1.4. Creating a stacked bar chart

To stack the data, e.g., make a stacked bar-chart, simply click the *Clustered (Z-Axis)* button  . Doing this for the example above gave the new chart shown below:

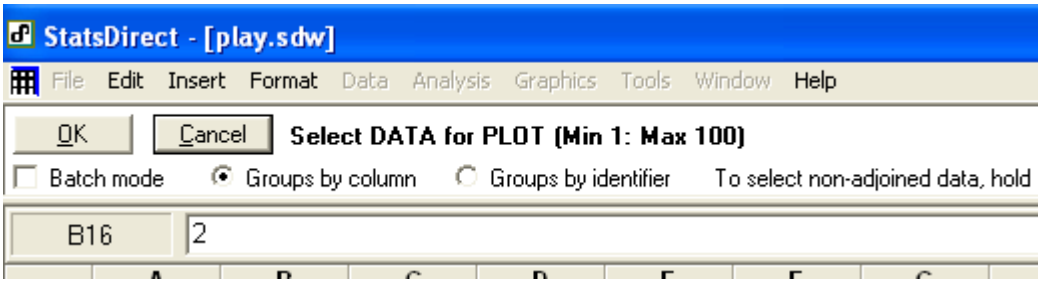



## 7.2 Dot plots, box and whisker plots and histograms

### 7.2.1 Dot plots for all of the sample

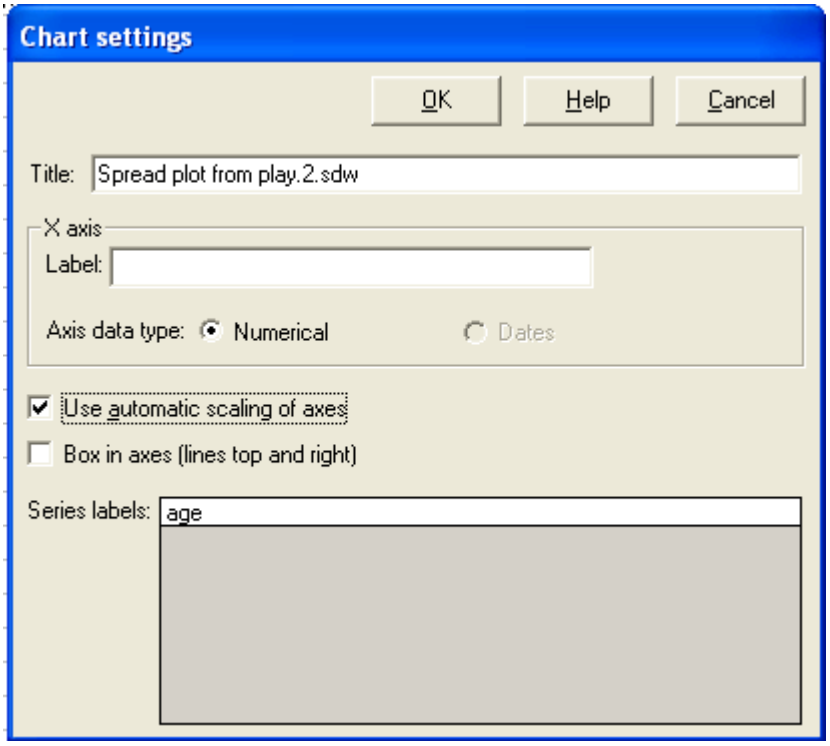
- *Graphics > Spread*

The top of the screen will change to:




- Check *Groups by column*
- Grey header of variable to be summarised, for example 'age'
- *OK*

The following window will appear



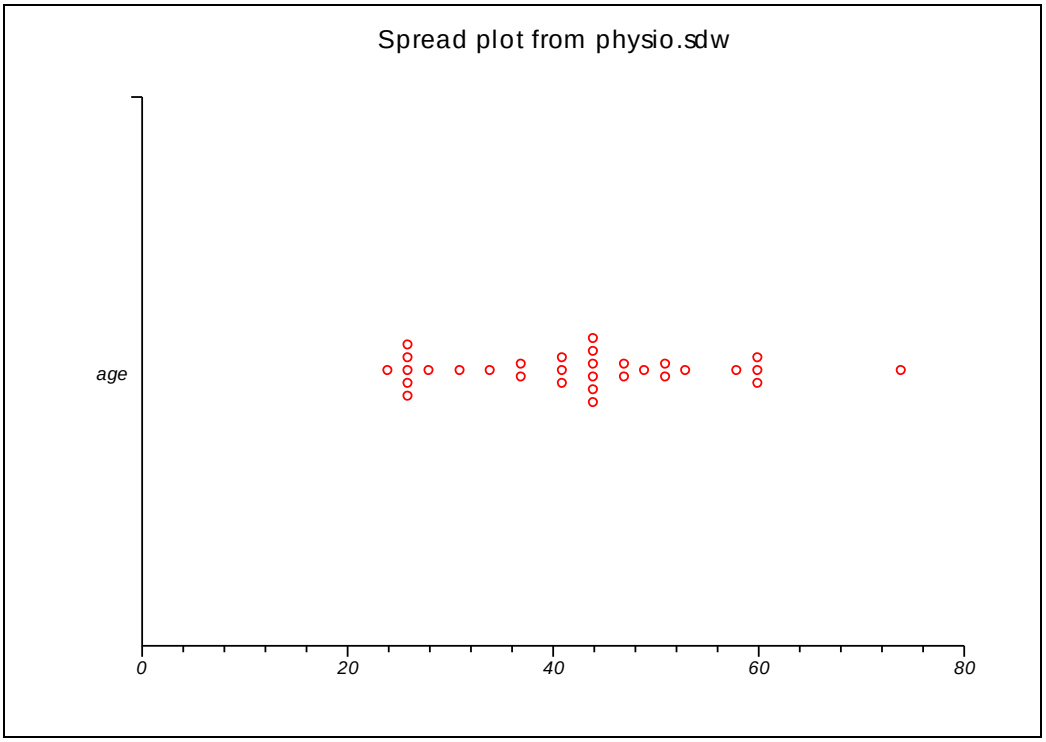

If necessary uncheck *Use automatic scaling of axes*, and make changes to *Title:* or *X axis Label:*

*OK*

In the next window:

Appropriate report > *Select*

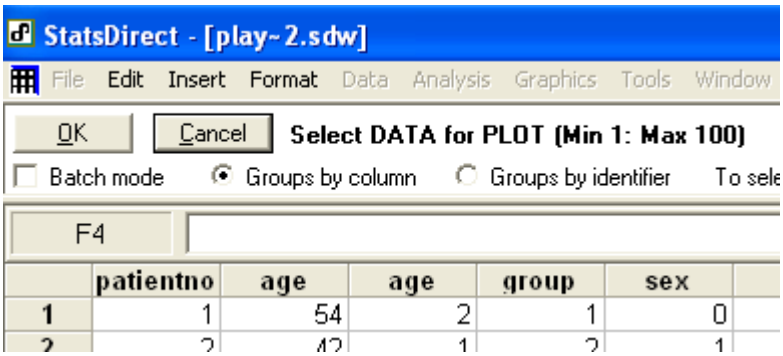
The graph will appear in the *Report*, e.g.:




### 7.2.2 Box and whisker plots for all of the sample

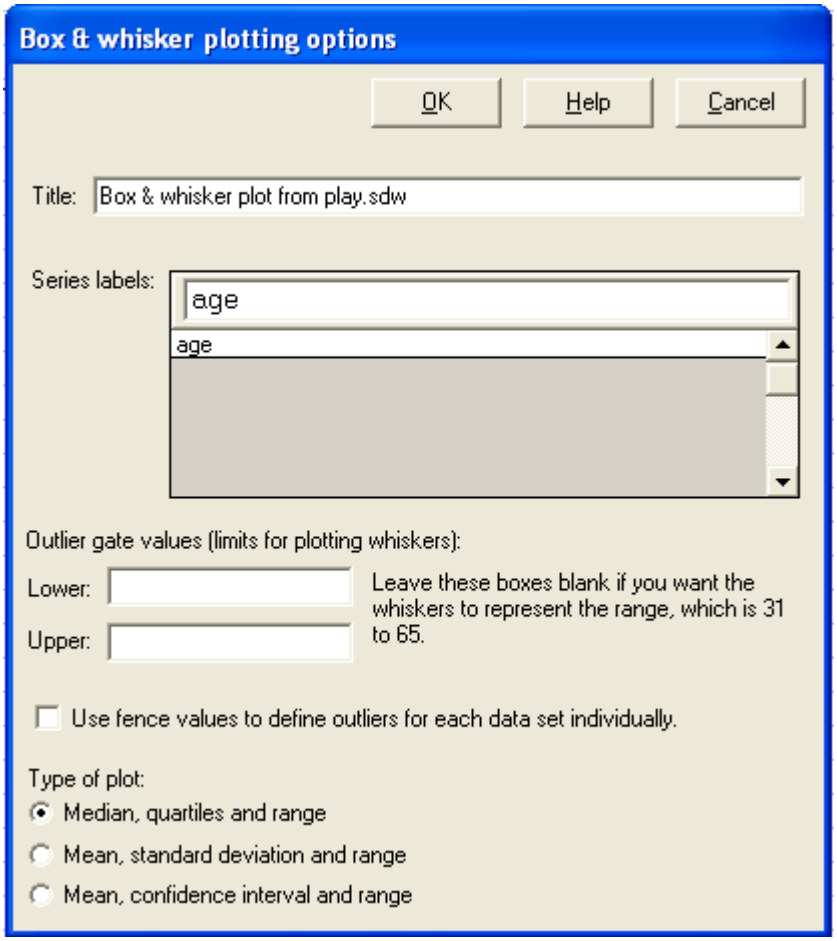
- *Graphics > Box & Whisker*

The top of the screen will change to:




- Check *Groups by column*
- Grey header of variable to be summarised
- *OK*

The following window will appear:

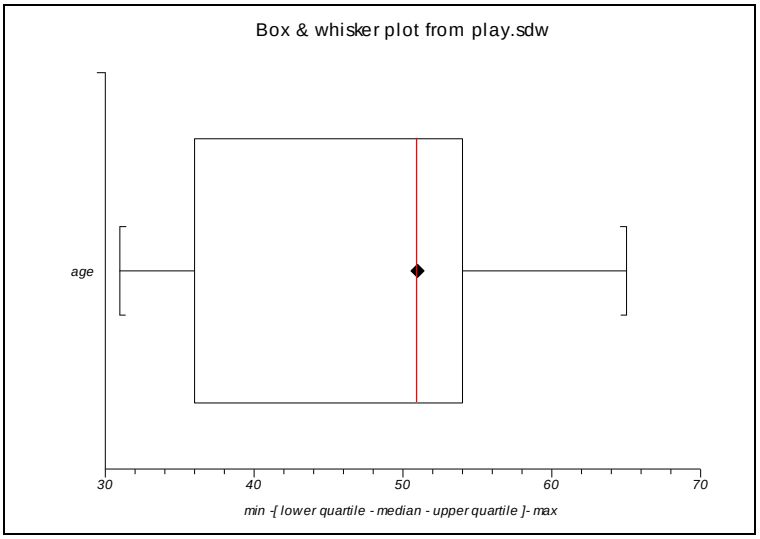



- If necessary, make changes to *Title:* or *Series labels:*
- Check *Use fence values to define outliers for each data set individually.*
- Check *Median, quartiles and range*
- *OK*

In the next window:

- Appropriate report > *Select*

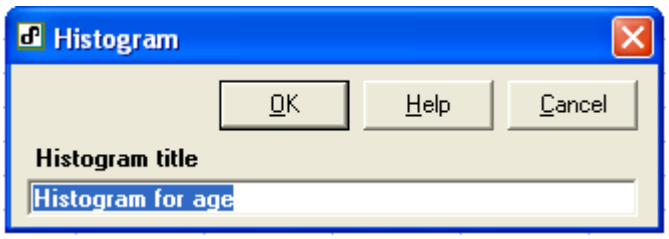
Graph will appear in *Report* e.g.:




### 7.2.3 Histograms for all of the sample

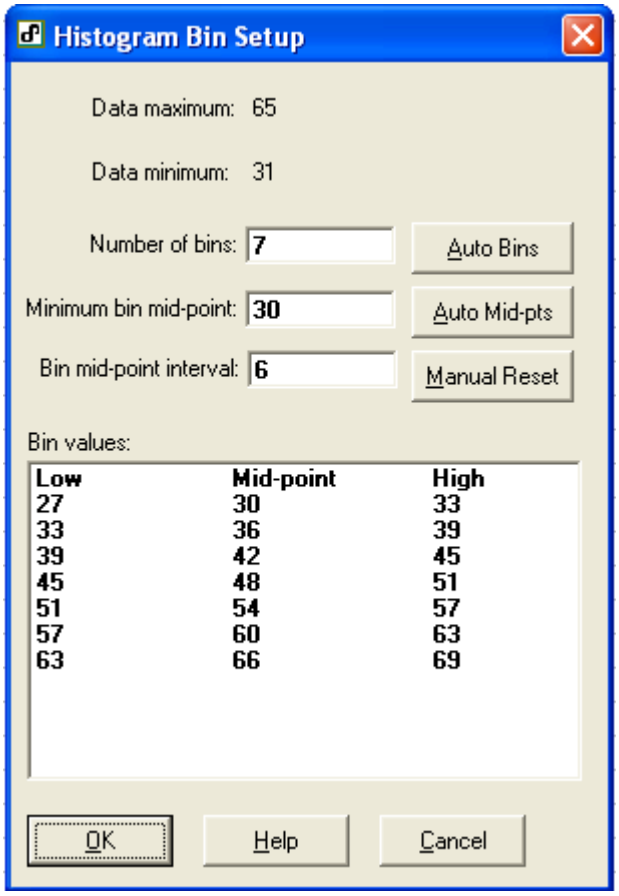
- *Graphics > Histogram*
- Grey header of variable to summarise

The next window allows you to change the title:




- If necessary change *Histogram title*
- *OK*

The next window asks about the grouping of the bars:

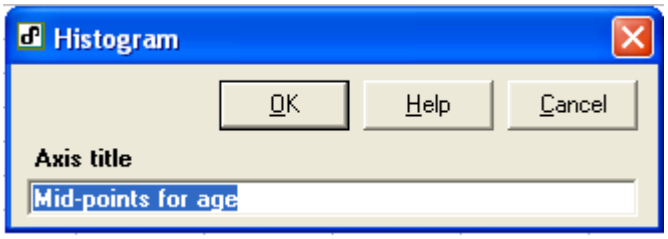



*OK* to start off with

A box will appear with the message “Overlay a normal curve?”:

*No*

The next window allows you to change the axis label:

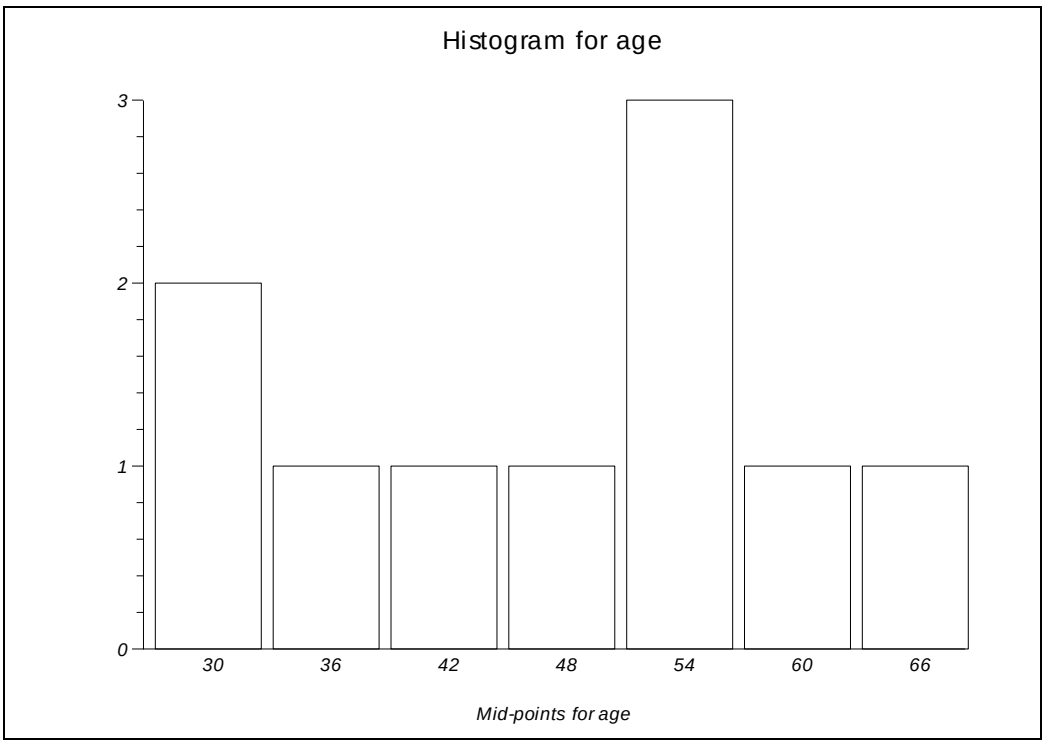



*OK*

In the next window:

Appropriate report > *Select*

Graph will appear in the *Report* e.g.:

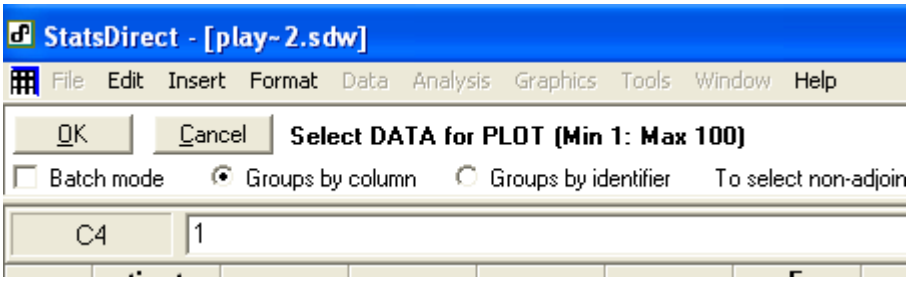



If you are not happy with the grouping, you can re-do the graph and make appropriate changes in the *Histogram Bin Setup* window.

### 7.2.4 Dot plot for comparing between groups

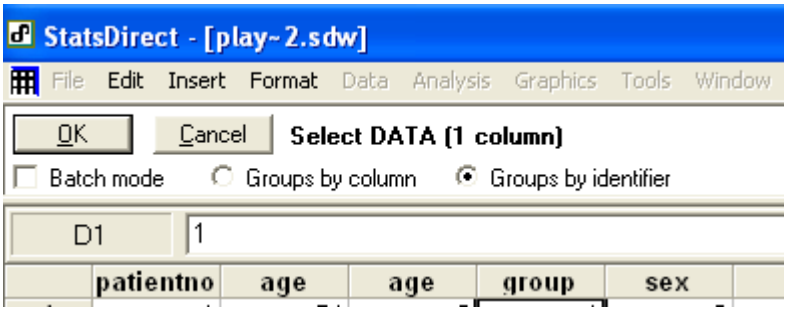
- *Graphics > Spread*

The top of the screen which change to:



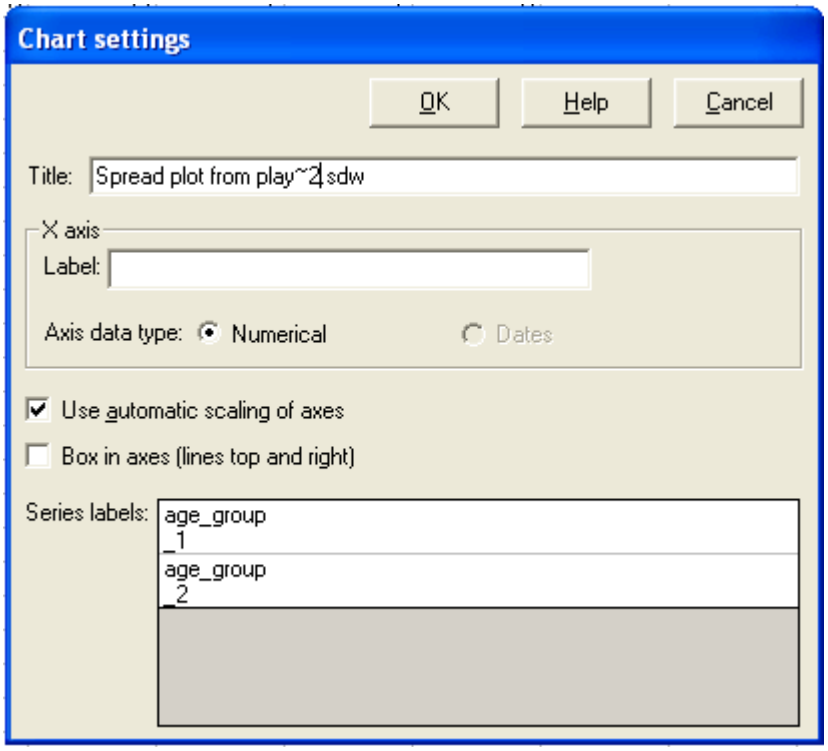

- *Groups by identifier*
- Grey header of grouping variable, e.g. 'group'
- *OK*

The screen will change to:




- Grey header of variable to be summarised e.g. 'age'
- *OK*

The following window will appear:

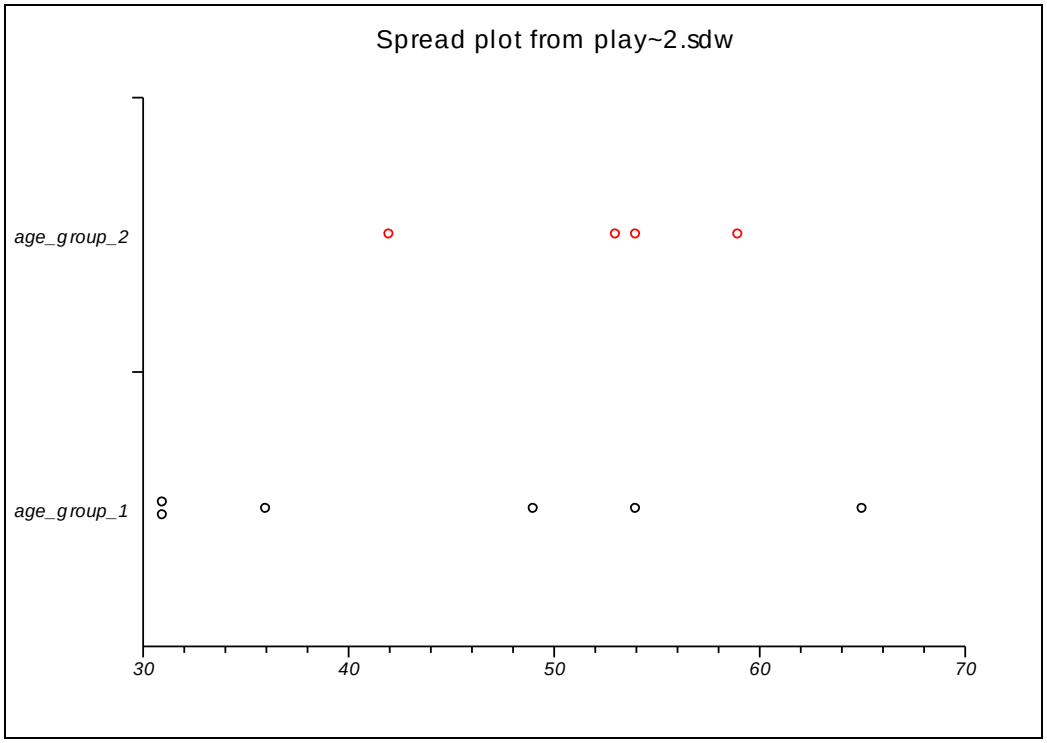



- If necessary, make changes to *Title*: or *X axis Label*
- *OK*

In the next window:

- Appropriate report > *Select*

A dot plot similar to that shown below should appear:



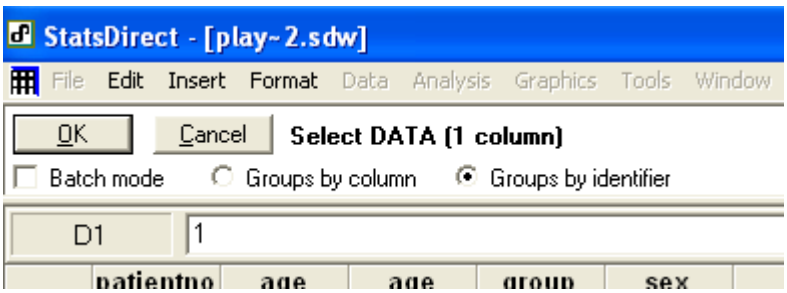

### 7.2.5 Box and whisker plots when comparing between groups

- *Graphics > Box and Whisker*

The top of the screen changes to:

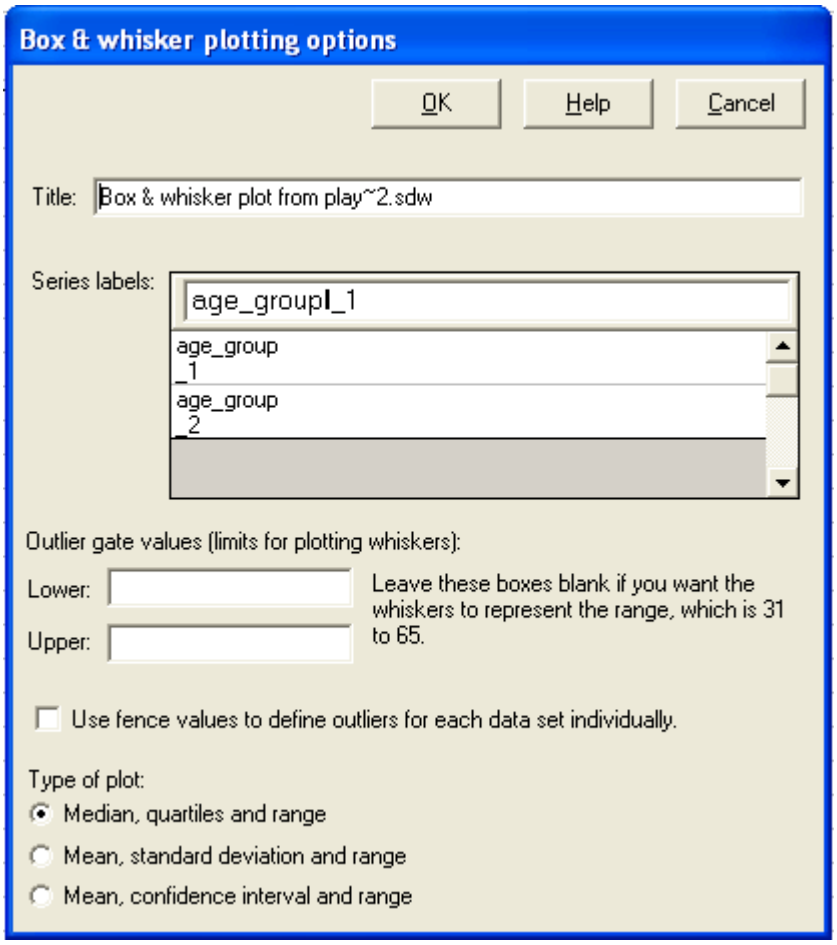

- *Groups by identifier*
- Grey header of grouping variable e.g. 'group'
- *OK*

The screen will change to:




- Grey header of variable to be summarised e.g. 'age'
- *OK*

The following window will appear:



- If necessary make changes to *Title:* or *Series labels:*
- Check *Use fence values to define outliers for each data set individually*
- *Median, quartiles and range*
- *OK*

In the next window:

- Appropriate report *> Select*

A graph will appear in the *Report* e.g.:


### 7.3 Limits of agreement

- *Analysis > Agreement > Continuous (Intra-Class...)*
- *Groups by column*
- Grey header of measures to be compared
- *OK*

In the next window:

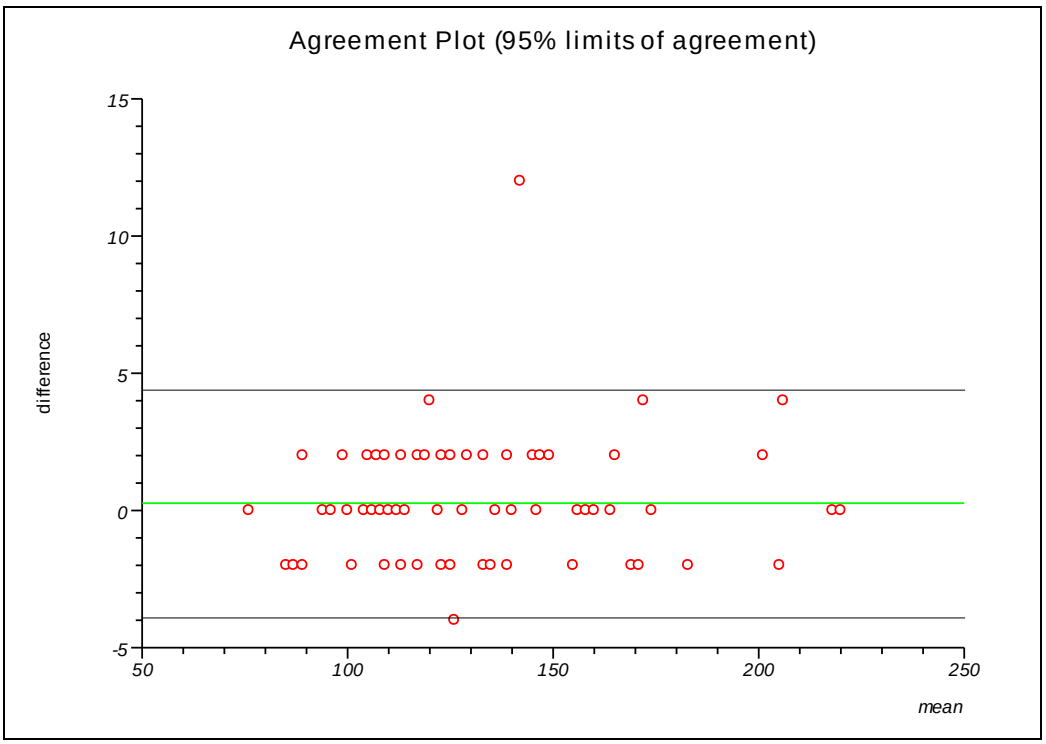
- Appropriate report *> Select*

A graph and summary will appear in the *Report,* e.g.:

**Agreement**

Variables: j1, r1

95% Limits of agreement = -3.871086 to 4.435792



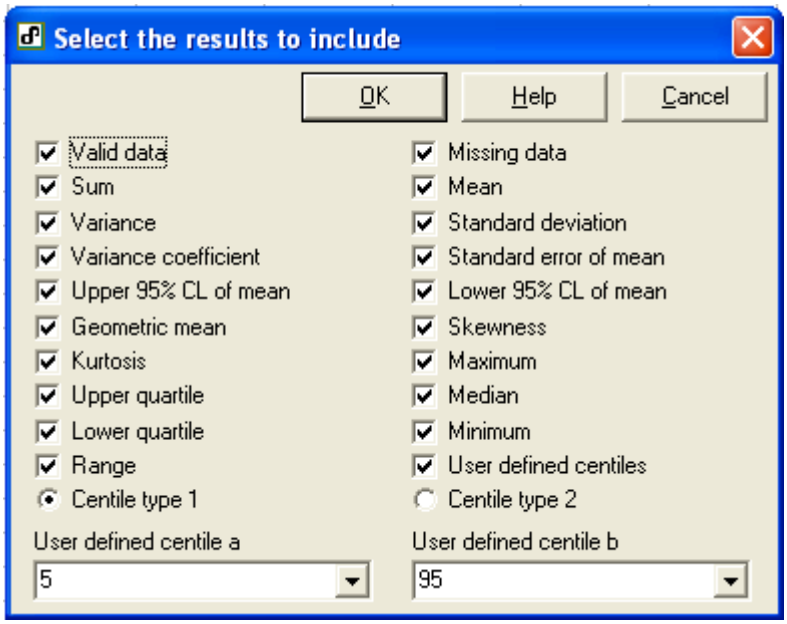
## 8 Descriptive statistics

Note: If you would prefer the categories to be presented in a different order, sort by the variable you would like to be considered first (see section 4.2).

8.1 For all of the sample:

- *Analysis > Descriptive > Univariate summary*
- *Groups by column*
- Grey header of variable to summarise
- *OK*

The following screen will appear:




- Customise the statistics to be displayed by unchecking any unwanted statistics
- *OK*
- Appropriate report > *Select*

8.2 Comparing groups

- *Analyse > Descriptive > Univariate summary*
- *Groups by identifier*
- Grey header of grouping variable > *OK*
- Grey header of variable to summarise > *OK*

The same screen as above will appear:

- Customise the statistics to be displayed by unchecking any unwanted statistics
- *OK*
- Appropriate report > *Select*

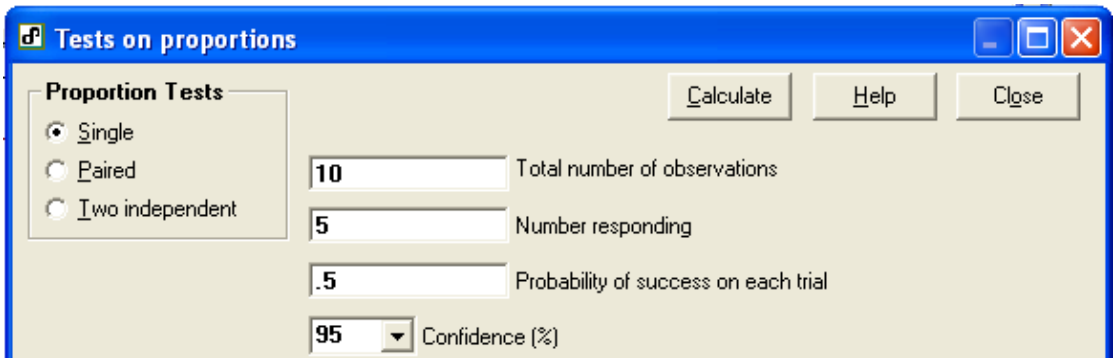
# 9 Confidence intervals

## 9.1 Confidence intervals for a proportion

Firstly produce a summary of the frequencies (through *Analysis>Frequencies* section 7.1.1)

- Go to *Report* where output is displayed
- *Analysis > Proportions > Single*

Use the output from the *Frequencies* analysis to enter the total and proportion of successes. For the example in section 7.1.1 the top of the window will look as below:




- *Calculate*

For the above example the following output would be given:

**Single proportion**

Total = 10, response = 5
Proportion = 0.5

Exact (Clopper-Pearson) 95% confidence interval = 0.187086 to 0.812914

Using null hypothesis that the population proportion equals 0.5
Binomial one sided P = 0.623
Binomial two sided P > 0.9999

Approximate (Wilson) 95% mid-P confidence interval = 0.236593 to 0.763407

Binomial one sided mid-P = 0.5
Binomial two sided mid-P > 0.9999

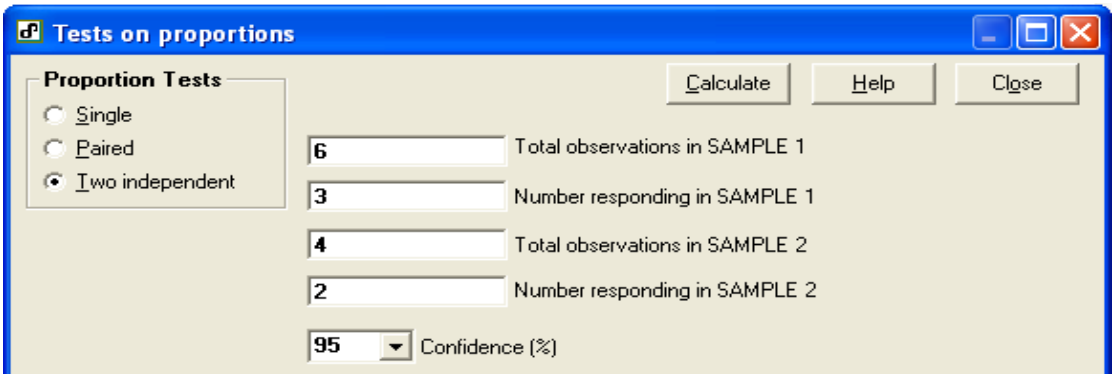
- *Close > Yes >* Appropriate *Report > Select*

### 9.2 For a difference in proportions

Firstly produce a summary of the frequencies (through *Analysis>Frequencies* section 7.1.2)

- *Report* where output is displayed
- *Analysis > Proportions > Two Independent*

Use the output from the *Frequencies* analysis to enter the total and proportion of successes for the two groups. For the example in section 7.1.2 the top of the window will look as below:




- *Calculate*

For the above example the following output would be given:

**Two independent proportions**
Total 1 = 6, response 1 = 3
Proportion 1 = 0.5
Total 2 = 4, response 2 = 2
Proportion 2 = 0.5

Proportion difference = 0

Approximate (Miettinen) 95% confidence interval = -0.550153 to 0.550153

Exact two sided (mid) P > 0.9999

Standard error of proportion difference = 0.322749
Standard normal deviate (z) = 0
Approximate two sided P > 0.9999
Approximate one sided P = 0.5

- *Close > Yes >* Appropriate *Report > Select*

### 9.3 For a mean

See section 8.1

### 9.4 For a difference in means

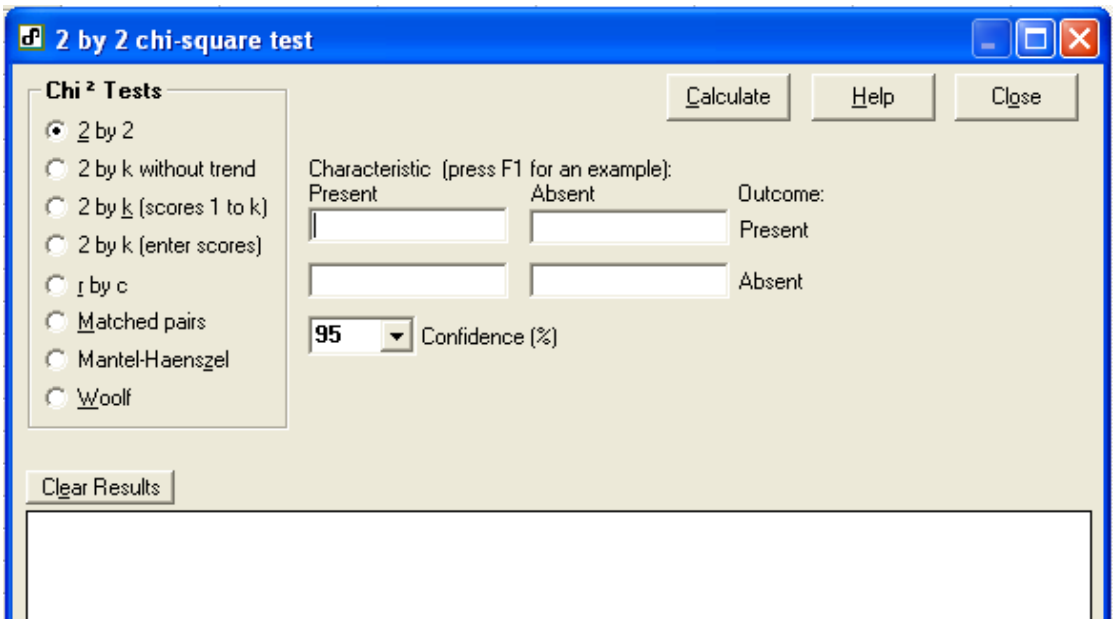
See section 10.2

## 10 Statistical tests

### 10.1 Chi-squared test, Chi-squared test for trend and Fishers' Exact test (comparing a categorical outcome across two or more groups)

*Analysis > Chi-Square Tests > 2 by 2*

A similar screen to below will appear:

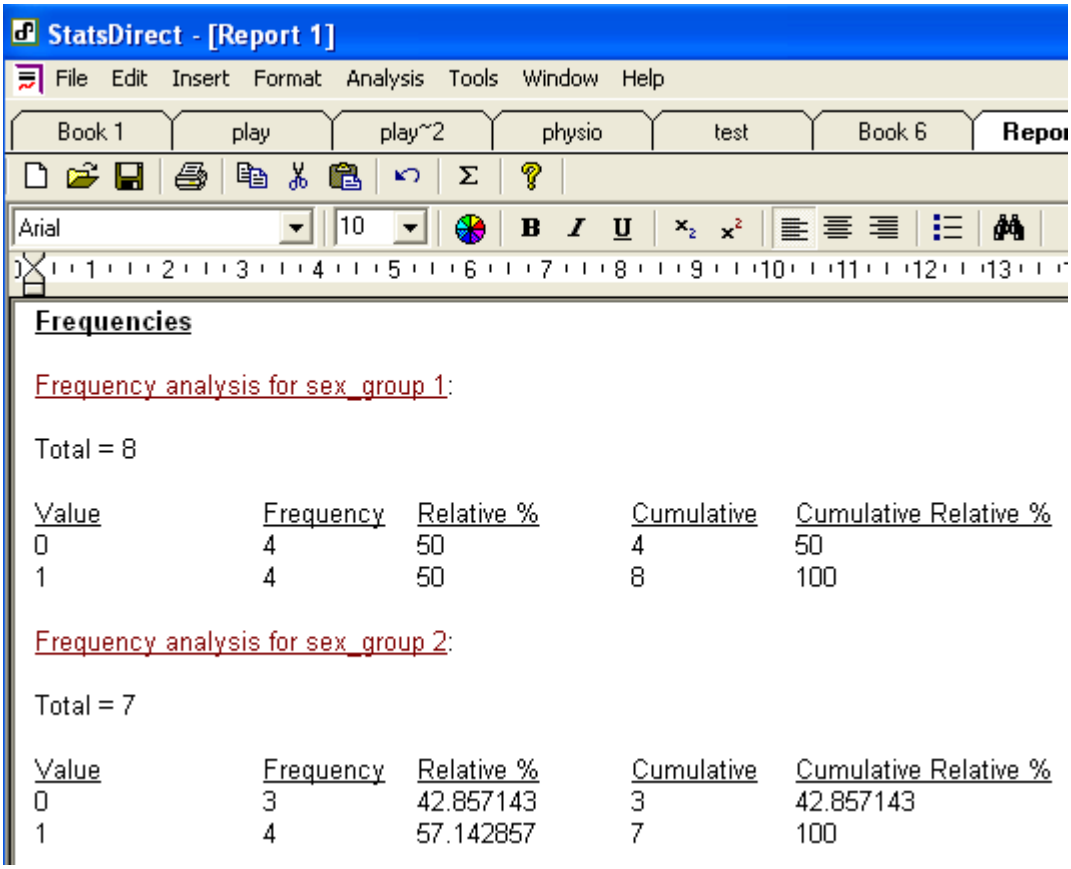



#### 10.1.1 Chi-squared test and Fishers' Exact test

Check either *2 by 2* (test categorical outcome of two groups) *or 2 by k without trend* (testing categorical outcome of k groups) or *r by c* (testing r categorical outcomes across c groups) as appropriate.

Enter the frequencies (which can be obtained via a frequency analysis – see Section 6) in the 2x2 set of cells under *Characteristic (press F1 for an example):* considering the columns to represent the explantory factor and the rows to represent the outcome measure.

For example, for the following *Frequencies* output where 'sex' is a dummy variable ('0' for female, '1' for male) and fall into either of groups 1 and 2:




**Frequencies**

Frequency analysis for sex_group 1:

Total = 8

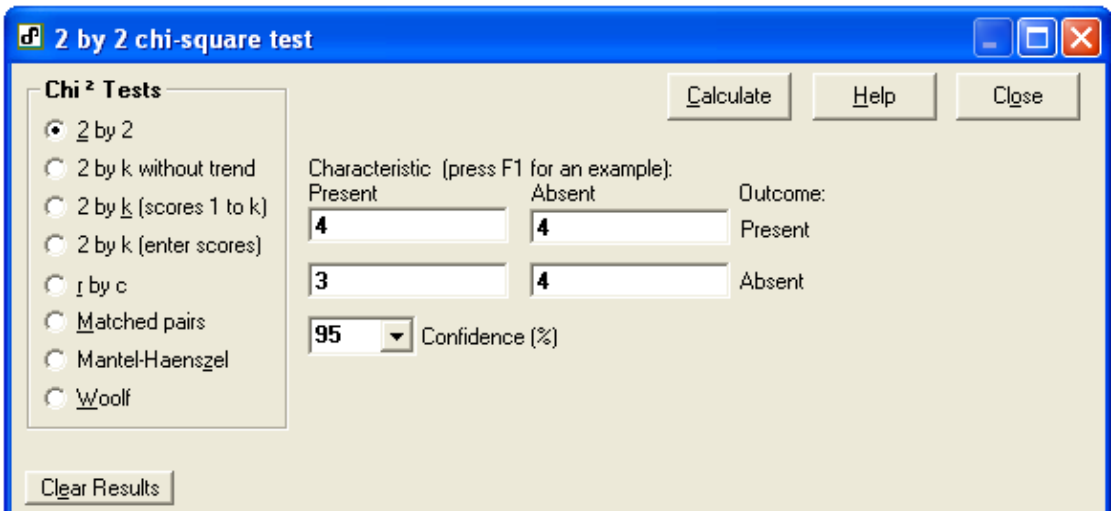
| Value | Frequency | Relative % | Cumulative | Cumulative Relative % |
|---|---|---|---|---|
| 0 | 4 | 50 | 4 | 50 |
| 1 | 4 | 50 | 8 | 100 |

Frequency analysis for sex_group 2:

Total = 7

| Value | Frequency | Relative % | Cumulative | Cumulative Relative % |
|---|---|---|---|---|
| 0 | 3 | 42.857143 | 3 | 42.857143 |
| 1 | 4 | 57.142857 | 7 | 100 |

Group 1 could be represented as the *Absent* column, and group 2 could be considered as the *Present* column, and female could be represented as the *Absent* row, and male the *Present* row (think of *Absent* and *Present* as '0's and '1's). This data could be entered as follows:

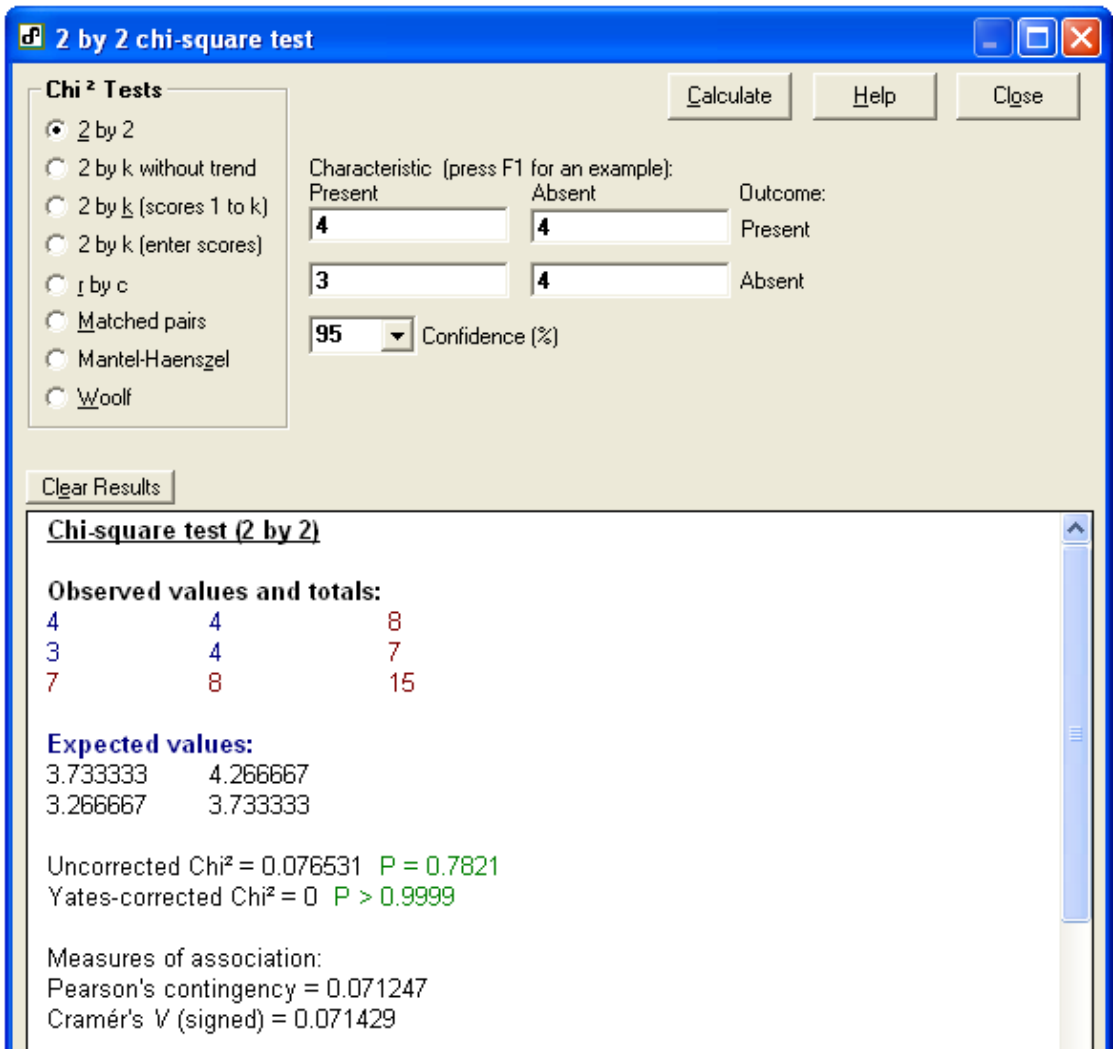



*Calculate*

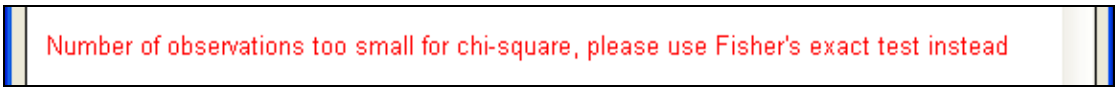
If 2 by 2:

Check *Neither > OK*

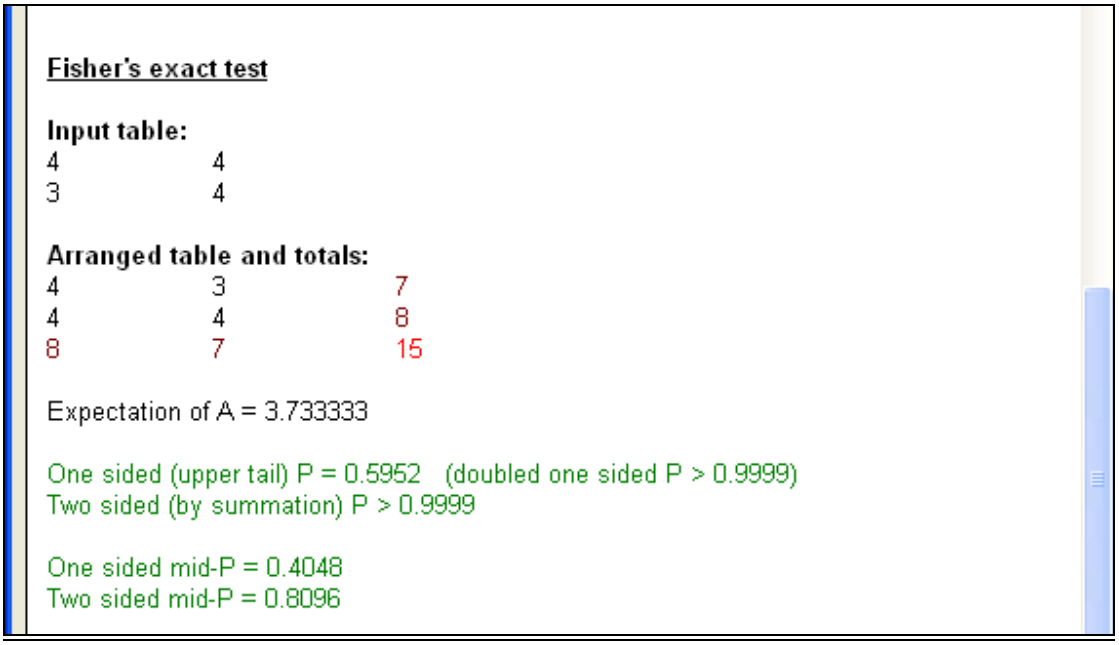
Decide whether a Chi-squared test is appropriate using the expected values that have been produced in the results section of the window. It is not appropriate if there is an expected value less than 1, or if more than 20% of the expected values are less than 5. Here is some of the output for the sex/group example:




None of the *Expected values:* are less than 1, but all of the values are less than 5 – therefore a Chi-squared test is not approriate in this case. Fortunately, newer versions of StatsDirect tell you this on the next line:




The results of a Fisher's Exact test are given just below this:



**Fisher's exact test**

**Input table:**

| | |
|---|---|
| 4 | 4 |
| 3 | 4 |

**Arranged table and totals:**

| | | |
|---|---|---|
| 4 | 3 | 7 |
| 4 | 4 | 8 |
| 8 | 7 | 15 |

Expectation of A = 3.733333

One sided (upper tail) P = 0.5952 (doubled one sided P > 0.9999)
Two sided (by summation) P > 0.9999

One sided mid-P = 0.4048
Two sided mid-P = 0.8096

### 10.1.2 Chi-squared test for trend

- Do as in 10.1.1, but in the *2 by 2 chi-squared test* box, check *2 by k (scores 1 to k)* instead of *2 by 2*. (this may also be reached by *Analysis > Chi-Square Tests > 2 by k > Trend Scores 1 to k*)
- Under *Enter frequencies for each row:* enter the number of successes and failures per subject
- *Calculate*

An example of an output is given below:

| | **Successes** | **Failures** | **Total** | **Per cent** |
|---|---|---|---|---|
| Observed | 16 | 6 | 22 | 72.73 |
| Expected | 15.6 | 6.4 | | |
| Observed | 53 | 13 | 66 | 80.3 |
| Expected | 46.8 | 19.2 | | |
| Observed | 42 | 16 | 58 | 72.41 |
| Expected | 41.127273 | 16.872727 | | |
| Observed | 27 | 15 | 42 | 64.29 |
| Expected | 29.781818 | 12.218182 | | |
| Observed | 18 | 14 | 32 | 56.25 |
| Expected | 22.690909 | 9.309091 | | |
| **Total** | **156** | **64** | **220** | **70.91** |

Total $Chi^2$ = 7.149102 |Chi| = 2.67378 (4 DF) P = 0.1282

$Chi^2$ for linear trend = 5.49603 |Chi| = -2.344361 (1 DF) P = 0.0191

Remaining $Chi^2$ (non-linearity) = 1.653072 (3 DF) P = 0.6474

## 10.2 T-test (parametric test comparing a numerical outcome across two groups)

- *Analysis > Parametric > Unpaired t*
- Check *Group by identifiers*
- Grey header of explanatory factor > *OK*
- Grey header of outcome measure > *OK*
- Select appropriate report > *Select*

Output of a similar format to below will be displayed in the report:

**Unpaired t test**

Mean of ttagi_group_1 = 159.19 (n=16)
Mean of ttagi_group_2 = 137.47 (n=15)

Assuming equal variances
Combined standard error = 22.47
df = 29
t = 0.97
One sided P = 0.1708
Two sided P = 0.3416

95% confidence interval for difference between means = -24.23 to 67.67

Assuming unequal variances
Combined standard error = 22.54
df = 28.23
t(d) = 0.96
One sided P = 0.1717
Two sided P = 0.3434

95% confidence interval for difference between means = -24.38 to 67.82

Comparison of variances
Two sided F test is not significant
No need to assume unequal variances

## 10.3 Mann-Whitney U test (non-parametric test comparing a numerical outcome across two groups)

- *Analysis > Non-Parametric > Mann-Whitney*
- Check *Group by identifiers*
- Grey header of explanatory factor > *OK*
- Grey header of outcome measure > *OK*
- Select the appropriate report > *Select*

Output of a similar format to below will be displayed in the report:

**Mann-Whitney U test**

Observations (x) in ttagi_group_1 = 16  median = 159  rank sum = 289.5
Observations (y) in ttagi_group_2 = 15  median = 105
U = 153.5  U' = 86.5

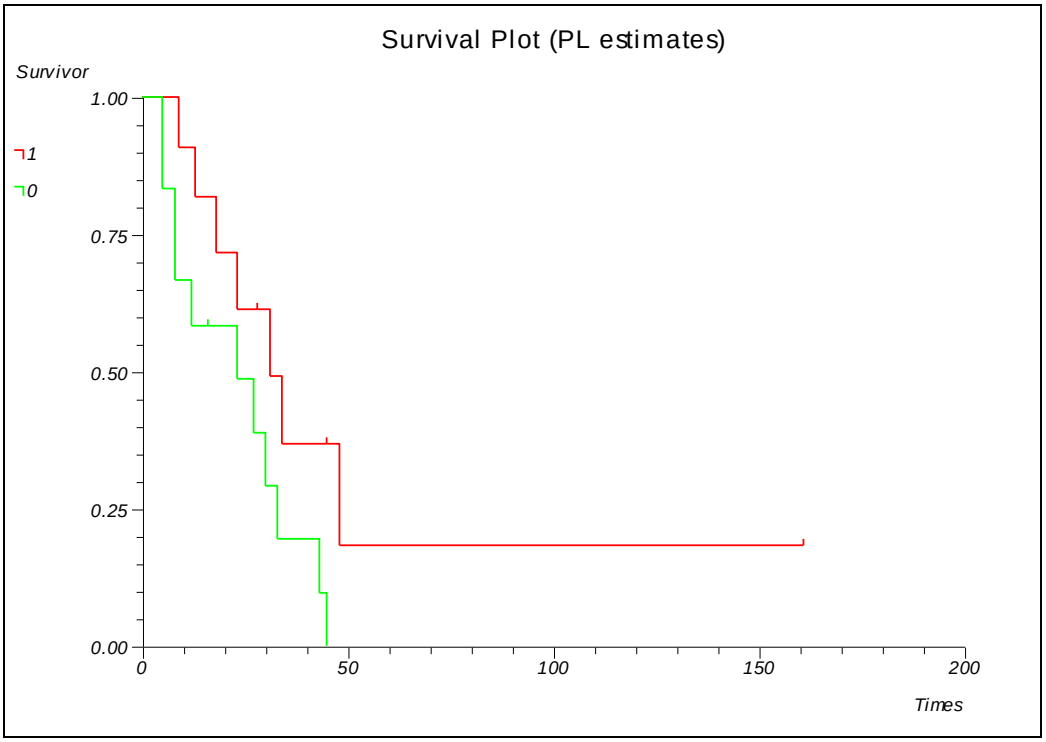
Exact probability (adjusted for ties):
Lower side P = 0.0956 ($H_1$: x tends to be less than y)
Upper side P = 0.9044 ($H_1$: x tends to be greater than y)
Two sided P = 0.1911 ($H_1$: x tends to be distributed differently to y)

95.1% confidence interval for difference between medians or means:
K = 71  median difference = 21
CI = -19 to 77

## 10.4 Kaplan-Meier graph and Log-Rank test (comparing a time-to-event outcome across two or more groups)

- *Analysis > Survival Analysis > Kaplan-Meier*
- Grey column header of time to event measure > *OK*
- Grey column header of measure which indicates whether event occurred > *OK*
- Grey column header of explanatory factor > *OK*
- *No*
- Select appropriate report > *Select*
- *OK*

An example of a survival plot which will be included in the output is given below:




- Appropriate datasheet
- *Analysis > Survival Analysis > Log rank & Wilcoxon*
- Grey column header of explanatory factor > OK
- Grey column header of time to event measure > OK
- Grey column header of measure which indicates whether event occurred > OK
- *Cancel*
- *Yes*
- *OK*
- Select appropriate report > *Select*

Output of a similar format to below will be displayed in the report:

**Logrank and Wilcoxon tests**

**Log-rank (Peto):**

For group 1 (treat = 1)
Observed deaths = 7
Extent of exposure to risk of death = 10.689336
Relative rate = 0.654858

For group 2 (treat = 0)
Observed deaths = 11
Extent of exposure to risk of death = 7.310664
Relative rate = 1.504651

test statistics:
-3.689336 3.689336

variance-covariance matrix:
0.249529 -4.007551
-4.007551 4.007551

Chi-square for equivalence of death rates = 3.396389 P = 0.0653

Hazard Ratio (approximate 95% confidence interval)
Group 1 vs. Group 2 = 0.435223 (0.169899 to 1.114892)

Conditional maximum likelihood estimates:
Hazard Ratio = 0.397818
Exact Fisher 95% confidence interval = 0.118548 to 1.199547
Exact Fisher one sided P = 0.0561, two sided P = 0.0808
Exact mid-P 95% confidence interval = 0.133888 to 1.092582
Exact mid-P one sided P = 0.0372, two sided P = 0.0744

**Generalised Wilcoxon (Tarone-Ware):**

test statistics:
-12.837917 12.837917

variance-covariance matrix:
0.018091 -55.276335
-55.276335 55.276335

Chi-square for equivalence of death rates = 2.981604 P = 0.0842

## 10.5 ANOVA
**(parametric procedure comparing a numerical outcome across more than two groups)**

- *Analysis > Analysis of Variance > One Way…*
- Grey column header of explanatory factor > *OK*
- Grey column header of outcome > OK

Output of a similar format to below will be displayed in the report:

**One way analysis of variance**

Variables: outcome_group_1, outcome_group_2, outcome_group_3

| Source of Variation | Sum Squares | DF | Mean Square |
|---|---|---|---|
| Between Groups | 9646.333333 | 2 | 4823.166667 |
| Within Groups | 14779.666667 | 15 | 985.311111 |
| Corrected Total | 24426 | 17 | |

F (variance ratio) = 4.89507 P = 0.0231

To paste into report:

- *Close*
- *Yes*
- Select appropriate report > *Select*

### Kruskal-Wallis test (non-parametric alternative to ANOVA)

- *Analysis > Analysis of Variance > Kruskal-Wallis…*
- Grey column header of explanatory factor *> OK*
- Grey column header of outcome > *OK*

Output of a similar format to below will be displayed in a window:

**Kruskal-Wallis test**

Variables: outcome_group_1, outcome_group_2, outcome_group_3

Groups = 3
df = 2
Total observations = 18

T = 3.973684
P = 0.1371

Adjusted for ties:

T = 3.981903
P = 0.1366

To paste into a report:

- *Close*
- *Yes*
- Select appropriate report > *Select*

## 11 Creating dummy variables

Note

Older versions of StatsDirect create dummy variables in the order it comes across them in the datafile. If you require a particular level to be the reference category (the level not represented by a dummy), then sort your data by this variable before doing the following procedure. It is best to have a reference category that is clinically sensible or one with many cases in it.

Example: The two dummy variables for the factor 'partner?' that would be created are 'yes' and 'no'. 'unknown' is the reference category, see below:

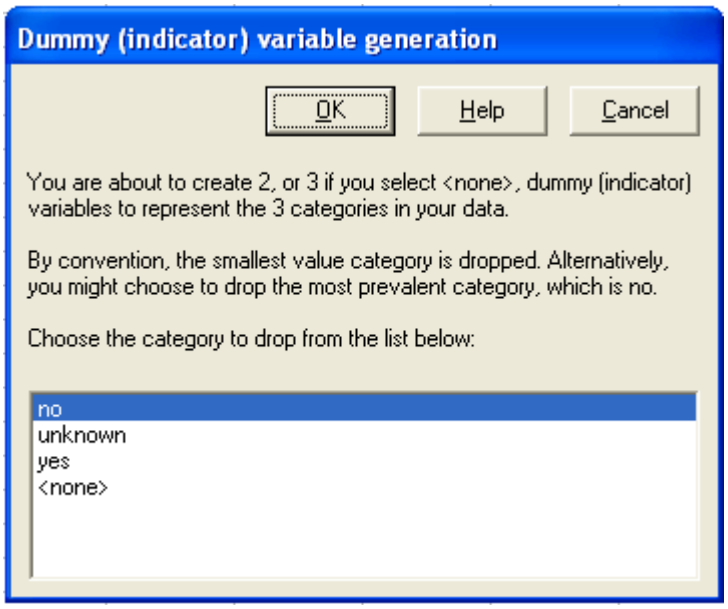
| ID | partner? | partner? (yes) | partner? (no) |
|---|---|---|---|
| 1 | yes | 1 | 0 |
| 2 | yes | 1 | 0 |
| 3 | no | 0 | 1 |
| 4 | yes | 1 | 0 |
| 5 | unknown | 0 | 0 |
| 6 | no | 0 | 1 |

If you wanted to compare to the 'yes' level, then sort by '*partner?*' first, as below:

| ID | partner? | partner? (no) | partner? (unknown) |
|---|---|---|---|
| 3 | no | 1 | 0 |
| 6 | no | 1 | 0 |
| 5 | unknown | 0 | 1 |
| 1 | yes | 0 | 0 |
| 2 | yes | 0 | 0 |
| 4 | yes | 0 | 0 |

- *Data > Dummy Variables*
- Grey header of explanatory factor > *OK*

A window similar to below will appear:



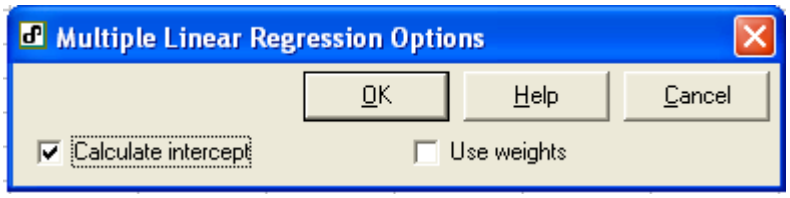

Check that it is creating the number of variables you were expecting.

- Under "Choose the category to drop from the list below:" select the variable you wish to be the reference category > *OK*
- Name of the sheet you are already in > *Select*

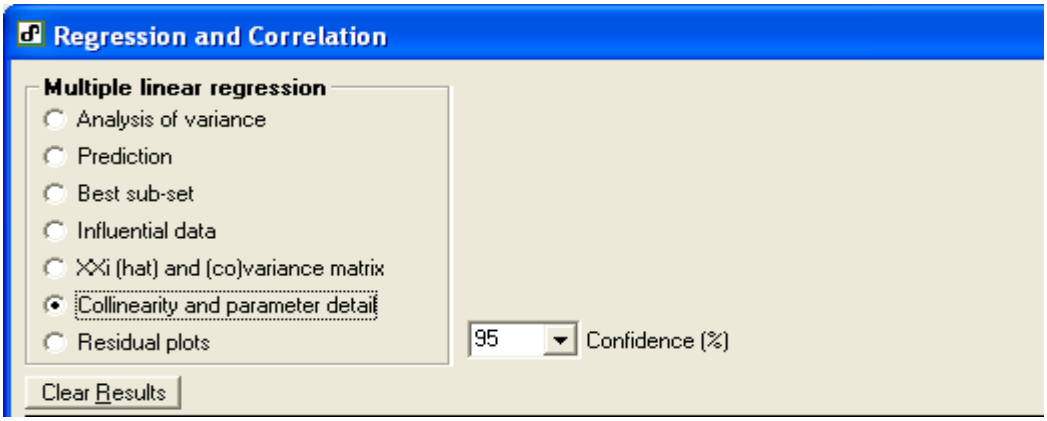
The new variables will be created. It is a good idea to rename them to something appropriate.

## 12 Linear regression (numerical, normally distributed outcome)

*Analysis > Regression & Correlation > Multiple Linear...*




*OK*
Grey header of outcome > *OK*
*Group by column*
Grey headers of explanatory factors > *OK*
Check *Colinearity and parameter detail* as below:




*Calculate*

An example of some of the output is given below:

*Difference between groups*

*p-value*

**Multiple linear regression**

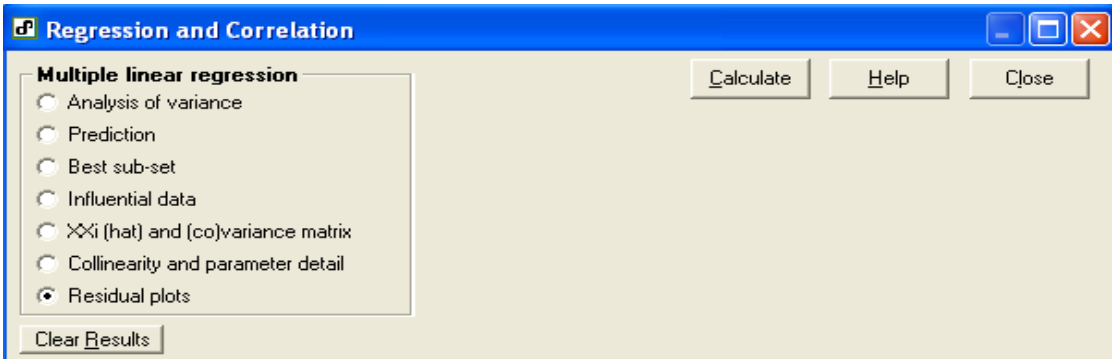
| | | | |
|---|---|---|---|
| Intercept | b0 = 89.650208 | t = 1.566659 | P = 0.1288 |
| group | b1 = -9.278291 | t = -0.428648 | P = 0.6716 |
| tt0gi | b2 = 0.488542 | t = 2.468731 | P = 0.0202 |
| age | b3 = -0.226022 | t = -0.254855 | P = 0.8008 |

ttagi = 89.650208 -9.278291 group +0.488542 tt0gi -0.226022 age

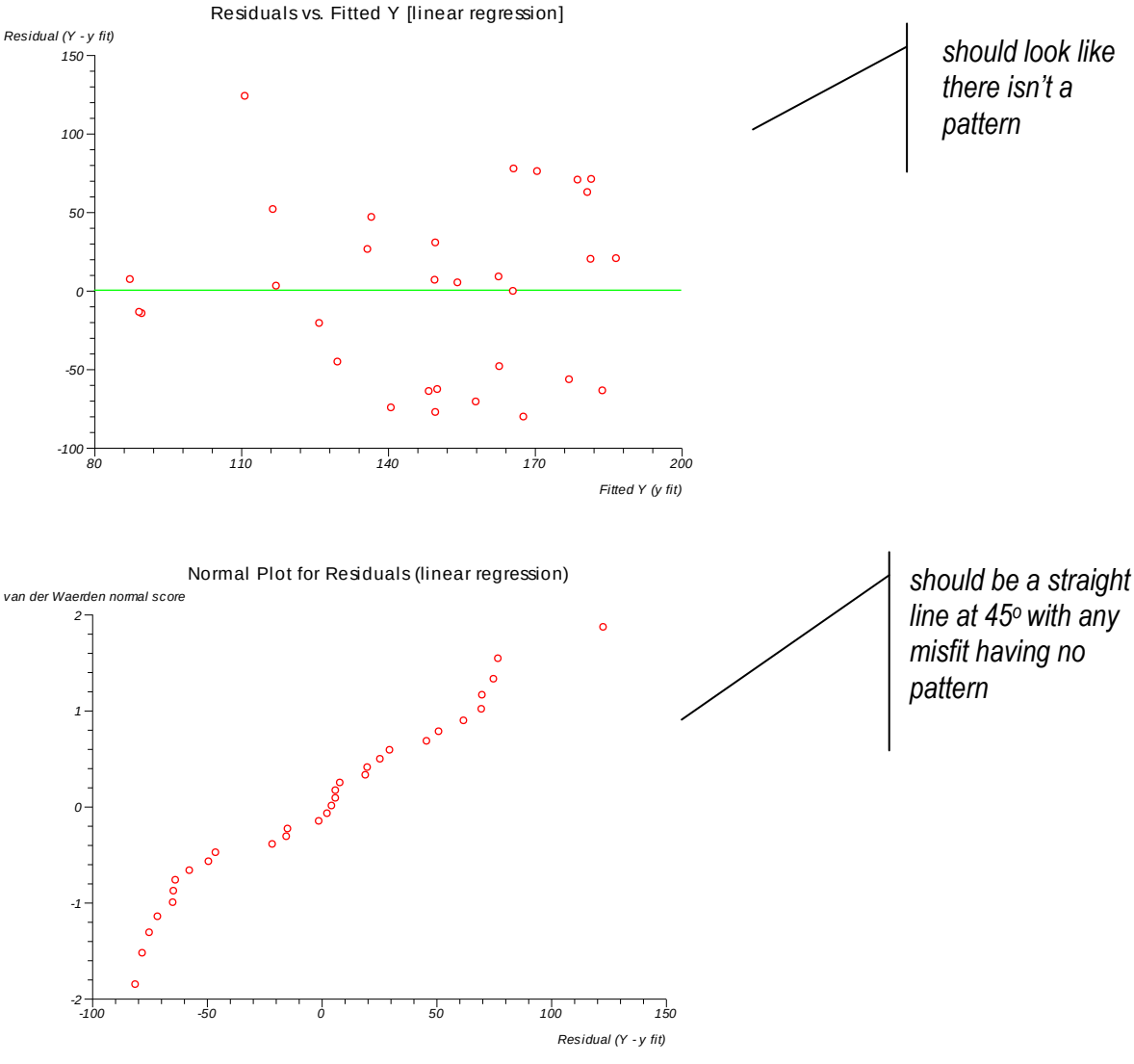
**Multiple linear regression - collinearity and parameter detail**

| | Coefficient | Standard Error | 95% Confidence Interval |
|---|---|---|---|
| constant | 89.650205 | 57.22381 | -24.797415 to 204.097825 |
| groups | -9.278291 | 21.645465 | -52.569221 to 34.012639 |
| tt0gi | 0.488542 | 0.197892 | 0.092758 to 0.884326 |
| age | -0.226022 | 0.886865 | -1.999752 to 1.547708 |

*Residual plots* as below > *Calculate*



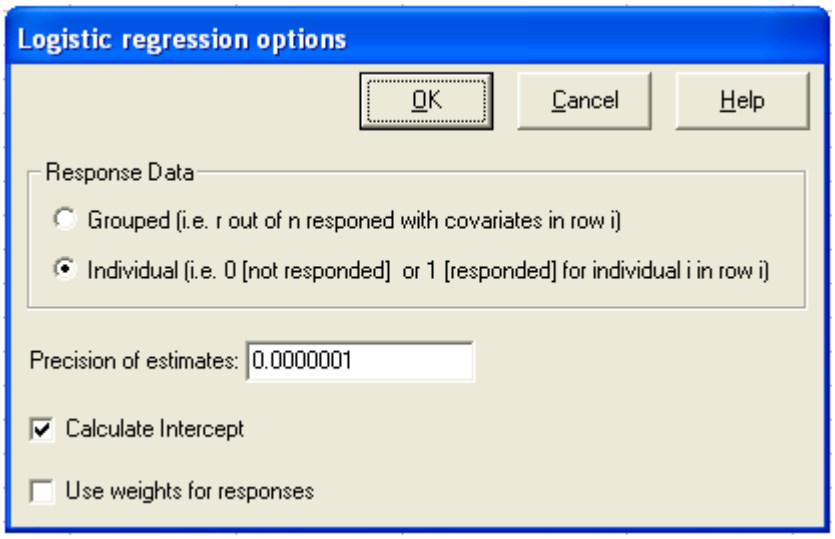

An example of some of the plots is given below:



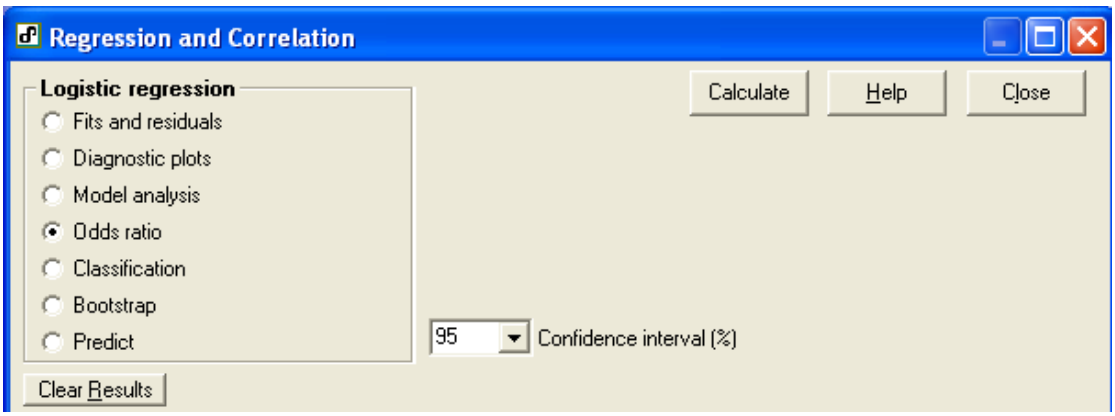
- *Close > Yes*
- Appropriate *Report > Select*

## 13 Logistic regression (binary outcome)

- *Analysis > Regression & Correlation > Logistic...*
- Check *Individual (i.e. 0 [not responded])…* and *Calculate Intercept* as below




- *OK*
- Grey header of *outcome* > *OK*
- Check *Groups by column*
- Grey headers of *explanatory factors* > *OK*
- Check *Odds Ratio* as below:




- *Calculate*

An example of the output is given below:

**Logistic regression**

Deviance (likelihood ratio) chi-square = 116.4203 df = 5 P < 0.0001

*p-value for group effect*

| | | | |
|---|---|---|---|
| Intercept | b0 = -3.328618 | z = -3.357776 | P = 0.0008 |
| psych ill? | b1 = 0.718923 | z = 3.514854 | P = 0.0004 |
| Age | b2 = 0.053512 | z = 4.325273 | P < 0.0001 |
| family | b3 = -0.719272 | z = -3.426963 | P = 0.0006 |
| home | b4 = -2.197231 | z = -9.467613 | P < 0.0001 |
| other | b5 = -0.964006 | z = -0.674511 | P = 0.5 |

logit Y = -3.328618 +0.718923 psych ill? +0.053512 Age -0.719272 family -2.197231 home -0.964006 other

*odds ratio for group effect*

*confidence interval of odds ratio for group effect*

**Logistic regression - odds ratios**

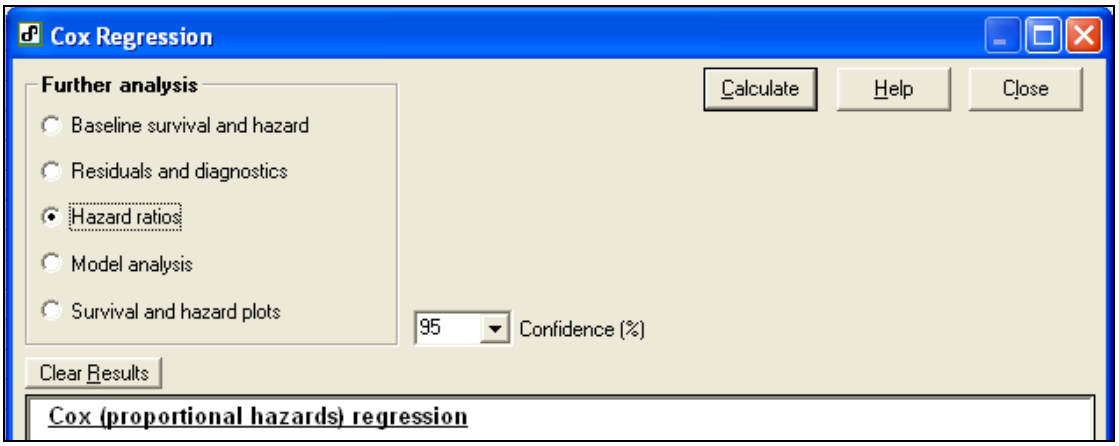
| Parameter | Estimate | Odds Ratio | 95% CI |
|---|---|---|---|
| Constant | -3.328618 | | |
| psych ill? | 0.718923 | 2.052221 | 1.374424 to 3.064274 |
| Age | 0.053512 | 1.05497 | 1.029696 to 1.080864 |
| family | -0.719272 | 0.487107 | 0.322826 to 0.734987 |
| home | -2.197231 | 0.11111 | 0.070503 to 0.175106 |
| other | -0.964006 | 0.381362 | 0.023164 to 6.278672 |

*Close > Yes*

Select appropriate report > *Select*

## 14 Survival regression (time-to-event outcome)

- *Analysis > Survival Analysis > Cox Regression*
- Grey column header of time-to-event measure > *OK*
- Grey column header of measure which indicates whether the event occurred > *OK*
- Grey column headers of explanatory factors > *OK*
- *Cancel > OK*
- Check *Hazard ratios* (as below) > *Calculate*




An example of the output is shown below:

**Cox (proportional hazards) regression**

*p-value for group effect*

Deviance (likelihood ratio) chi-square = 16.943139 df = 4 P = 0.002

| | | | |
|---|---|---|---|
| group | b1 = 0.692197 | z = 2.503218 | P = 0.0123 |
| age | b2 = -0.013497 | z = -1.265433 | P = 0.2057 |
| histgrad (1) | b3 = -1.147459 | z = -1.557426 | P = 0.1194 |
| histgrad (2) | b4 = -0.521152 | z = -1.874593 | P = 0.0608 |

*hazard ratio for group effect*

**Cox regression – hazard ratios**

| Parameter | Hazard Ratio | (95% CI) |
|---|---|---|
| group | 1.998101 | 1.162093 to 3.435533 |
| age | 0.986594 | 0.966184 to 1.007435 |
| histgrad (1) | 0.317442 | 0.074908 to 1.345245 |
| histgrad (2) | 0.593836 | 0.34437 to 1.024018 |

| Parameter | Coefficient | Standard Error |
|---|---|---|
| group | 0.692197184570695 | 0.276522895871924 |
| age | -1.34966340772734E-02 | 1.06656286099365E-02 |
| histgrad (1) | -1.14745858762533 | 0.736765829930851 |
| histgrad (2) | -0.521152032660354 | 0.278008145176212 |

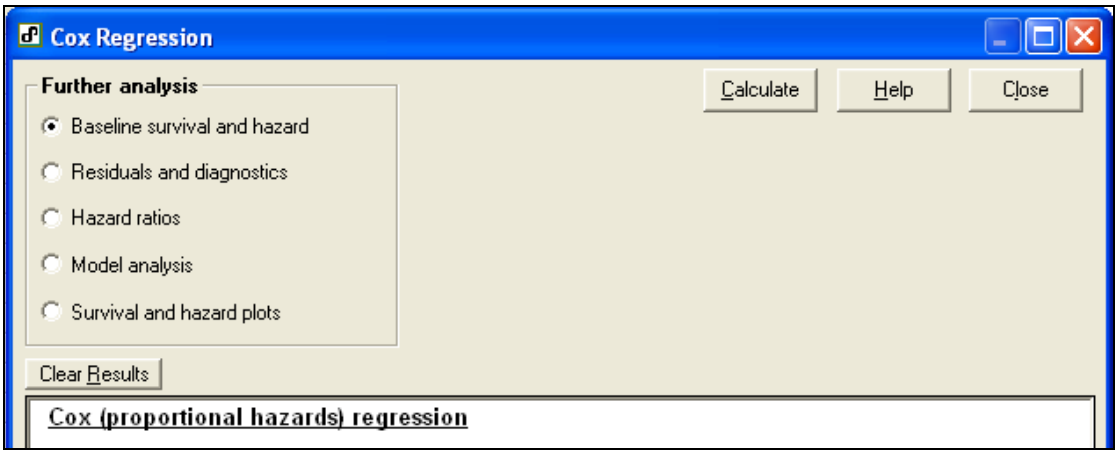
- *Close > Yes*
- Select appropriate report *> Select*

**To check constant hazard ratio assumption:**

Perform the analysis in the same way as before, but rather than clicking Cancel when asked to "select data for STRATA", click the grey column header of the appropriate categorical explanatory factors. Continue as before, but do not include that factor in list of predictors.

In the *Regression and Correlation* window:

- *Baseline survival and hazard* (as below) > *Calculate*




- *Yes*
- Select workbook with your data in > *Select*
- Go to workbook with your data in

Rename the headers in the new columns and delete the cells in those new columns which contained the original titles (note you will also need to move the entries for these new columns up by one row).

These next steps are for creating the variables which you need to plot for the complementary log-log plots:

- *Data > Apply Function*
- Select grey header of *Survival (baseline)* (one of the columns StatsDirect has just created) > *OK*
- "ln(-ln(V1))"
- *OK* > Select workbook with your data in > *Select*
- "complog strata factor" where factor is the name of the factor which you put as the strata
- *OK*
- *Data > Apply Function*
- Grey header of time to event measure > OK
- "ln(V1)"
- *OK* > Select workbook with your data in > *Select*

- "log time"
- *OK*

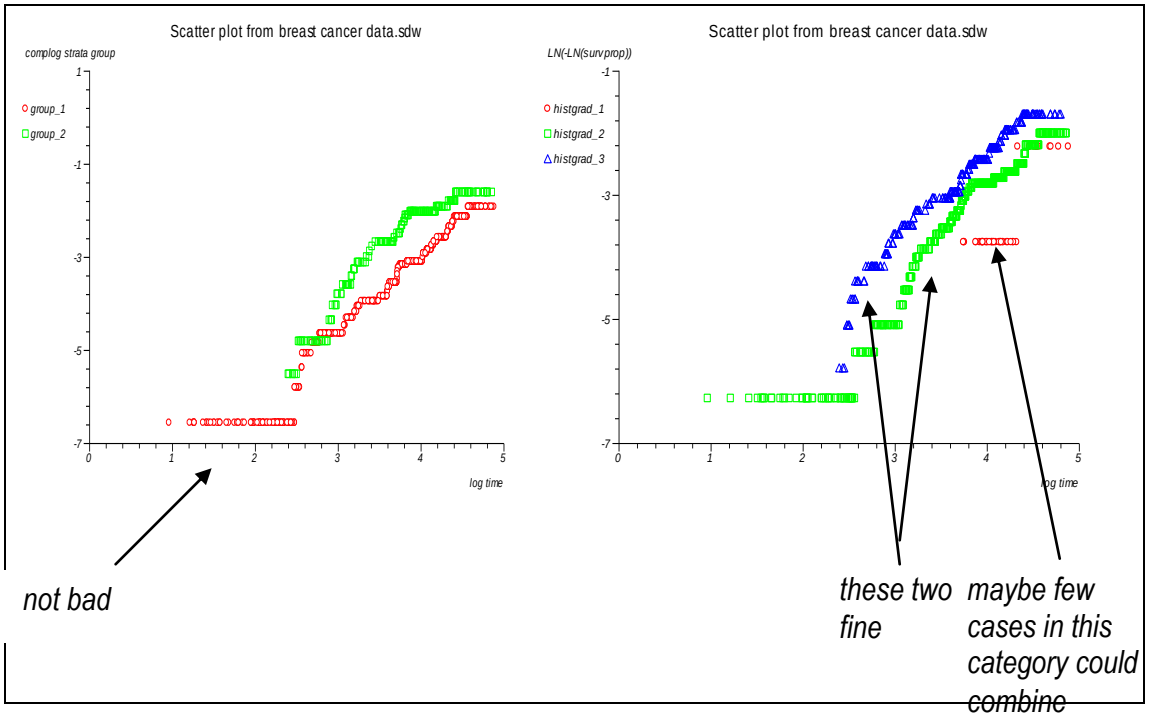
These next steps are for creating the complementary log log plot:

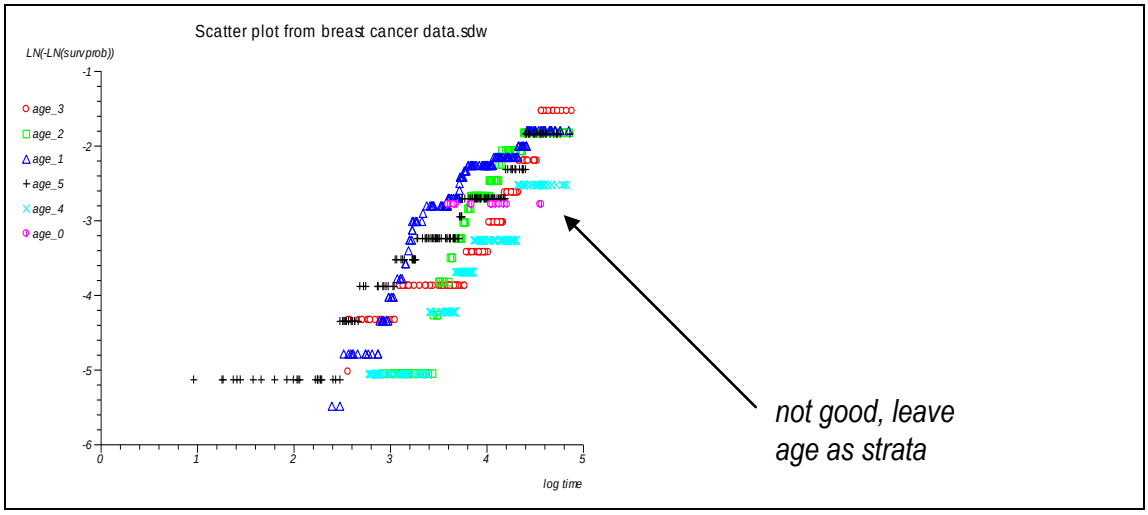
- *Graphics > Scatter*
- Check *Groups by identifier*
- Grey header of factor you entered as strata > *OK*
- Grey header of *complog strata factor* > *OK*
- Grey header of *log time* > *OK*
- *OK* > Select appropriate report > *Select*

Repeat for all explanatory factors. For any factors that aren't categorical, create a new variable which categorises the factor and use this to define the strata.

Examples of the graphs produced for the same example as above are given below:

Note: for each explanatory factor, for a constant hazard ratio, the category lines should be parallel and not cross.




Scatter plot from breast cancer data.sdw
LN(-LN(survprob))
age_3
age_2
age_1
age_5
age_4
age_0
-1
-2
-3
-4
-5
-6
0
1
2
3
4
5
log time
not good, leave
age as strata