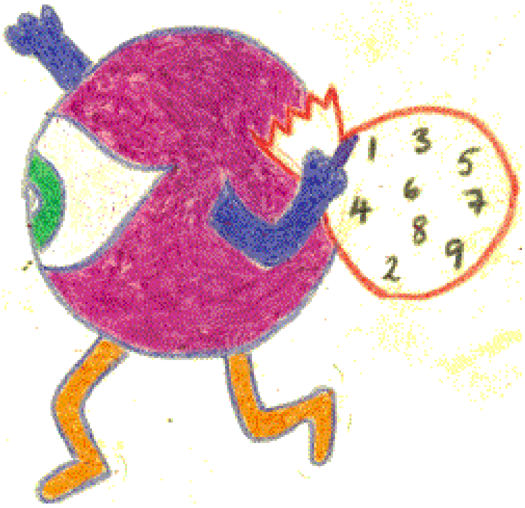
Relevant & Engaging Statistics & Data Handling

Using a Spreadsheet such as Excel to Analyse and Present Data



Centre for
Statistical Education

Chapter 7

Second Edition: published in the UK in 2009 by

The Royal Statistical Society Centre for Statistical Education
The University of Plymouth
Plymouth
PL4 8AA
UK

The advice and information in this booklet are believed to be true and accurate at the date of printing, but neither the authors, nor the publisher can accept any legal responsibility or liability for errors or omissions.

We would like to thank the many contributors to this book. These include: Doreen Connor, Neville Davies, John Marriott, Alan Catley (Chapter 4) and Mark Crowley (Chapter 6) and those who reviewed and helped in it's production including Alison Davies, Claire Webster, Peter Holmes & Elizabeth Gibson.

This booklet is aimed at all secondary level teachers: there are hints and tips that we hope will be useful to support the teaching and learning of statistics and data handling. We hope that you find the material useful. Please email or write to us with suggestions for improvements. We will try to respond to all communications.

Doreen Connor
The Royal Statistical Society Centre for Statistical Education
The University of Plymouth
2009

email info@censusatschool.org.uk

www.censusatschool.org.uk

www.rsscse.org.uk

Chapter 7

Using a Spreadsheet such as Excel to Analyse and Present Data

It can appear very easy to analyse data and produce charts and diagrams in a spreadsheet such as Excel, but very often the default ones produced are not the best ones, and some are even wrong! In this chapter we offer some *dos* and *don'ts* when using spreadsheet software, and go through some worked examples that can help produce a correct statistic or graph.
Throughout this chapter we are using the Excel spreadsheet software. The sample of data we show as an example is available to download from http://www.censusatschool.org.uk/get-data/results/phase-7-0607 towards the bottom of this webpage.

Firstly it is important that you understand that unfortunately Excel was not

- devised with mathematics or statistics teachers in mind as the main user but as a business package.
- designed with error messages to stop you doing things that are possible but mathematically and statistically wrong.

Dos

Do:

- get to know your software, its strengths and limitations;
- try out any new technique with simple numbers and small data sets first to see what happens before attempting things with real data;
- match graph types with data types (see below);
- match calculations with data types (see below);
- be critical of the output and be prepared to change it;
- make sure that representations and calculations are interpreted.

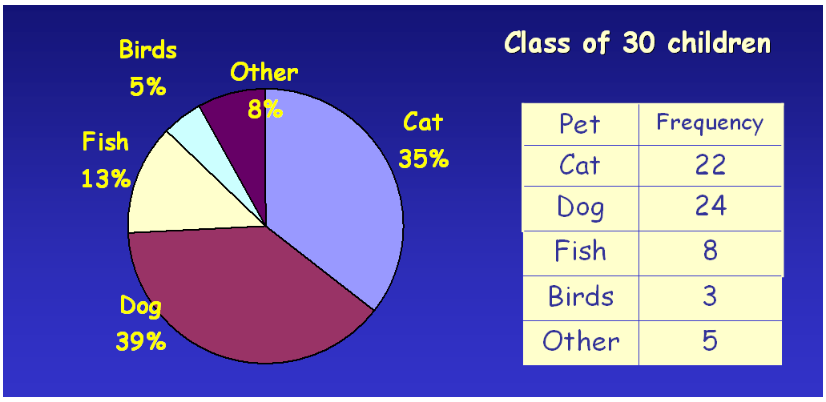
Matching graphs and calculations with data types

Not all numbers, or data, carry the same amount of information. We need to know what type of data we are dealing with before we know what type of calculation or graph may be appropriate. Here are the main types of data:

Data – Nominal Scale: Here measurements are placed in categories such as Male or Female, Red, Yellow or Blue. They may sometimes be given codings e.g. 1 = Male and 2 = Female. **It makes no sense at all to do any arithmetic with these codings.**

Data measured on a **Nominal** or **Categorical** scale puts the items in categories where the order has no meaning and so does not matter. Appropriate graphs for this are pie charts (if there are not too many categories), bar charts and pictograms Pie charts emphasise the proportion each category is of the whole; bar charts emphasise the relative frequencies between the categories. Since the order does not matter, the categories may be put in decreasing order of frequency, but they often are not. The categories for a Pie chart must be mutually exclusive. This means that the categories must not overlap and the total number of data items must be the number of the complete set of values. For example, take a look at the following Incorrect Pie Chart.
Can you understand why this pie chart is wrong? – If so then you understand what mutually exclusive means!

If you are having problems, look at the total number of children in the class, this is 30 but if you add up how many pets they have the total comes to more than 30. in other words some of the children have more than one pet so converting the data into a pie chart which shows proportions of a whole does not make any sense!



Class of 30 children

Pet	Frequency
Cat	22
Dog	24
Fish	8
Birds	3
Other	5

Pie Chart pertaining to show the types of pets that the children in the class have.

Data – Ordinal scale: here measurements are meaningful in terms of their order. Ranking in terms of preference 1 (best) to 5 (worst) of food flavours, for example does give an order but does NOT mean that a food with flavour 5 is actually 5 times worse than one ranked 1.
Others could be made ordinal by giving definitions say to small, medium and large household sizes, but this can mean losing important information. Although pie charts are used for these data they lose the impact of the order of category. Bar charts, including stacked or composite bar charts are better. With ordinal data it is possible to calculate the median and quartiles (maybe as categories) as well as the items for *nominal* or *categorical* scales.

Data – Ratio or Interval scale: measurements made on a scale of equal units, such as height in metres or time in seconds. Data measured on an *interval* or *ratio* scale have measures associated with them on which we can do meaningful arithmetic. Examples are ***school year***, ***height***, ***household size***, ***time taken to get to school***. There is a distinction between measures that are discrete, i.e can only take certain values such as number of eggs or house numbers and those that are continuous and can take any value. Pie charts are not appropriate. Bar charts should not be used, rather, for *interval* or *ratio* scales, histograms should be used. Frequency polygons can be used. The whole range of statistical calculations can be used.

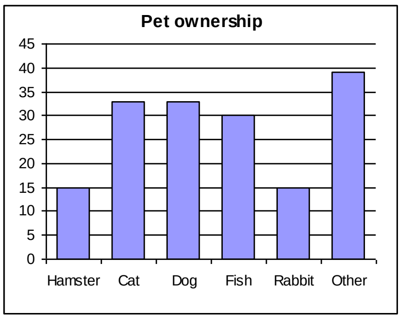
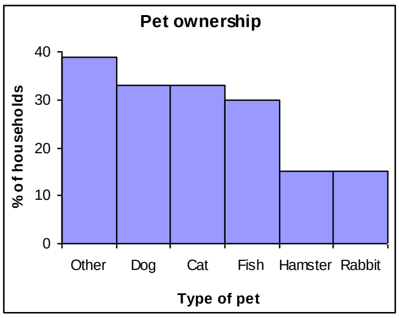
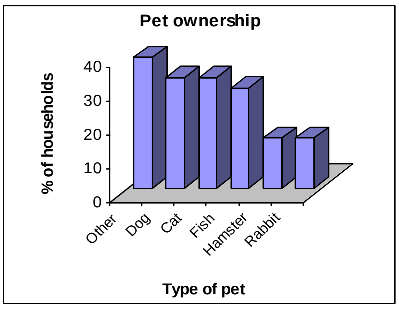
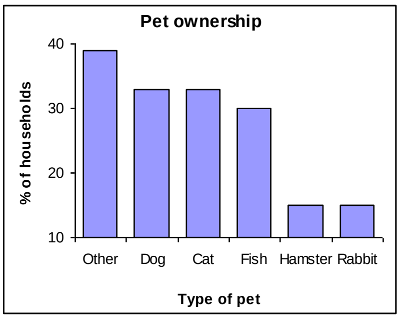
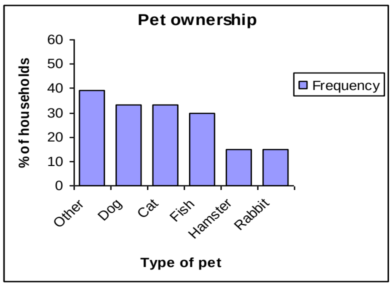
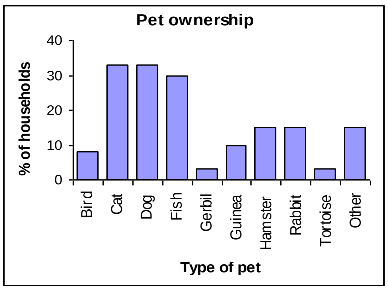
We have been careful to talk about categorical scales, interval scales and so forth. Within these scales there are measurements that are discrete or continuous. In many textbooks this distinction is confused by talking about *categorical data*, *discrete data* etc. These phrases should be interpreted as 'data measured on a categorical scale' and 'data measured on a discrete scale'. From a practical point of view all data are discrete, they are simply measured on different scales.

Don’ts

Don’t:

- expect Excel to do what you think it ought to do;
- use a technique/formula that you have not checked out beforehand;
- use graphs or calculations which are not appropriate to the type of data;
- draw three-dimensional pie charts, bar charts or histograms in Excel as, although they might look ‘nice’, they are misleading as the third dimension they show has no meaning and often misleads.

Even the simplest of graphs such as a bar chart can present a host of problems to the uninitiated. In the next table we present some bar charts that can be drawn with Excel using some of the responses that were received to the **Pet** question in phase 1 of the *CensusAtSchool* project. Underneath each chart we list errors, or *Don’ts*, that could be discussed with learners. Are these things simple aesthetics or is there a right and wrong way to draw a bar chart?

Problem Bar Charts using some of the Pet data from phase 1 of *CensusAtSchool*	
 The columns are not in order of size. Arguably the horizontal axis needs no title. The vertical axis has no title and too many marked points.	 There should be gaps between the columns.
 The axis labels are not horizontal. The 3-D visual effect distorts the picture.	 The non-zero vertical origin exaggerates any differences
 The vertical axis extends too far. The axis labels are not horizontal. The legend is unnecessary, cramping the plot area.	 The axis labels are not horizontal. The columns are not in order of size. Use horizontal bars rather than vertical labels.

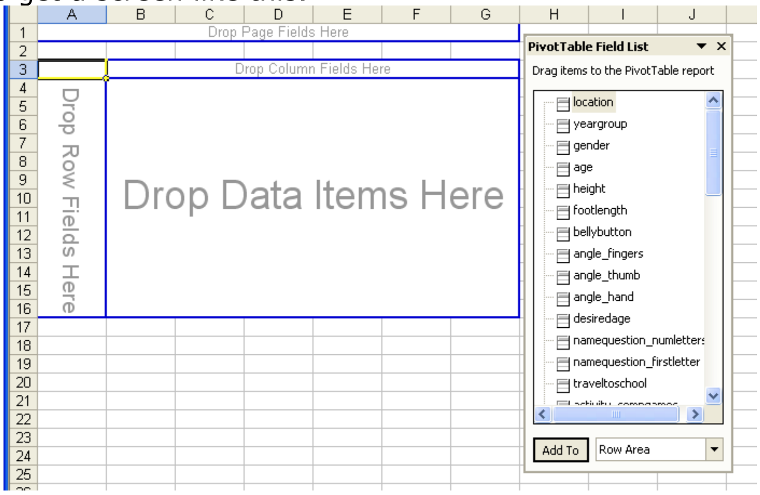
In the following sections we will attempt to guide you how to use Excel to show the right ways to generate simple charts and diagrams. For this purpose we are going to use a datafile of 100 random responses from the CensusAtSchool project. The data is available for download from the site at http://www.censusatschool.org.uk/get-data/results/phase-7-0607 If you wish to work through the examples with the booklet you will need to choose the Sample A random dataset from the bottom of the page, and open the file in Excel.

Two way or Pivot tables

As it stands it is difficult to see all the information in the spreadsheet. One way to get some insight into overall patterns is to construct some two-way tables. In Excel these are called *pivot tables*. Initially these can be quite confusing to construct because of all the possible different options. Because Excel was developed for business use, the default is to add up values in cells (generally this was money) whereas we usually want to count how many pupils have different properties. We therefore add a new column to the spreadsheet called *Count.*

Basic Two way or Pivot Tables

1. In cell **AS1,** type *Count.* In cell **AS2** type in the number 1. Move the cursor to the bottom right of cell **AS2** until it becomes a black cross. Now drag down to the bottom of **AS101**. The number 1 should appear in each of the cells in column AS.
2. Suppose we wanted to know how many boys and how many girls there were in each of the school years in our data set. Click anywhere in the spreadsheet. Go to **Data > PivotTable and PivotChart report** and click **Next > Next > Finish** to get a screen like this.



3. Drag ***Yeargroup*** to *Drop Column Fields Here* in the table, ***Gender*** to *Drop Row Fields Here*. Use the arrow to find ***Count*** and drag this to *DATA*. You should get the following table.

Sum of Count	Year					
Gender	7	8	9	10	11	Grand Total
female	10	5	22	7	1	45
male	10	9	25	8	3	55
Grand Total	20	14	47	15	4	100

4. Can you interpret this table in simple terms? How many females and males are there in the data? Have we got equal numbers in each year group?

If you just want to count how many pupils are in each year you can drag ***Gender*** back to the list and then ***Yeargroup*** to *ROW* and ***Count*** to *DATA* to get an even simpler table. This can be useful before drawing bar charts, pie charts etc.

Sum of Count		
Yeargroup		Total
7	7	20
8	8	14
9	9	47
10	10	15
11	11	4
	Grand Total	100

More Complex Pivot Tables

You will have noticed that the pivot table you have just produced went on to a new sheet in Excel. It is useful to keep track of these sheets by naming them.

1. Double click on the tab at the bottom of the spreadsheet with the pivot table on it. (It will say something like *Sheet 1).* It is now highlighted and you can type in your name for it, say *Year & gender table.* You can re-order the sheets by dragging these tabs.
2. The first pivot table we constructed was very simple and not very informative. Suppose we wanted to know how old these children would like to live until in different years and what were the differences between males and females. Find your way back to the *original data* sheet and click somewhere in the table.
3. Go to **Data > PivotTable and PivotChart report** and click **Next > Next** and choose **No > Finish.** Now drag ***Yeargroup*** and ***Gender*** to COLUMN, ***desiredage*** to ROW and ***Count*** to DATA. This gives the numbers of pupils in each category. Interpret these data. Name this new sheet *Desired Age data.*
4. You will notice that there are 33 different ages. It might make the picture clearer if we grouped the amounts of money.
5. Right click on any item in the ***desiredage*** column and **Group and Outline > Grouping** brings up a small table. The table shows that the figures go from 65 to 150 and suggests a grouping of size 10 (i.e. 10 years). Type in 20 against *By,* to give class intervals of 20 years. Click on OK. You will find that the amounts of years have been grouped as 65 - 84, 85 - 104 etc.
6. Interpret the table you get and investigate other groupings.

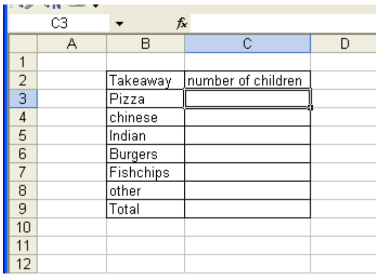
To ungroup the table, right click on a cell in the ***desiredage*** column then **Group and outline > Ungroup**.

Producing charts and diagrams

Bar charts: getting the right picture

Excel will draw bar charts, but they do not always come out as you might expect.
Using the same data as in the previous section. Open the datasheet. To draw a bar chart from the data follow the instructions below.

1. Excel will only draw bar charts once the data has been put into a frequency table. It will NOT do this for you so you can either use the pivot table with a chart attached or you need to first produce a frequency table of your data.
2. We want to produce a bar chart to show the children's favourite takeaway food. The data is in column AA and they could have chosen pizza, Indian, Chinese, burgers fishchips or other . To produce our frequency table choose a fresh sheet, rename it *Barcharts* and produce the outline of our table as below:



	A	B	C	D
1				
2		Takeaway	number of children	
3		Pizza		
4		chinese		
5		Indian		
6		Burgers		
7		Fishchips		
8		other		
9		Total		
10				
11				
12				

3. In C3 type the following formula which will tell the spreadsheet to go to sheet 1 where our data is and count the responses giving this answer from the column AA

   =COUNTIF(Sheet1!AA:AA, "pizza")

   You should find that 37 children gave this response.
   Copy the formula into the cells C4 to 8 changing the type of fast food each time.
   To check we have got all 100 children type the following formula into the total cell in C9

   =SUM(C3:C8)
4. Click on B2 and drag to C8 so that the cells in the table are highlighted.
5. Click on the chart wizard icon in the toolbar (it looks like a 3d bar chart) and choose *column graph.*
6. Now do **Next > Next** until you get a screen like this.

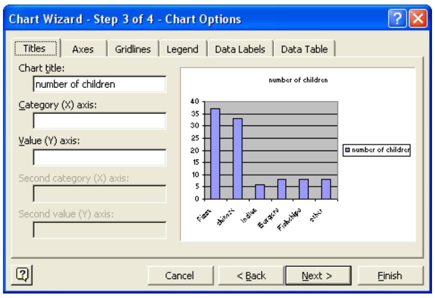


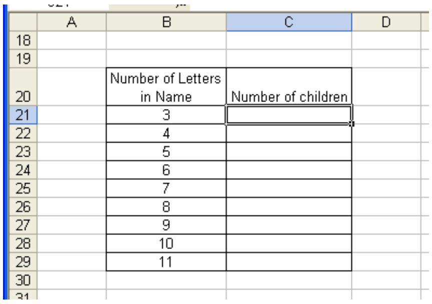
Chart title: number of children
Category (X) axis:
Value (Y) axis:
Second category (X) axis:
Second value (Y) axis:

7. Type in your chosen chart title, name for the x-axis and *frequency* for the y-axis. You can remove the little box on the right of the graph by going to the *Legend* tab and removing the tick against the box *show legend.*
8. Click on **Next > Finish** to complete your bar chart.

9. By clicking on appropriate areas of the bar chart (e.g. the grey background) you can change it. Explore the menus to present your bar chart as well as you can.

What happens when the variable is numeric rather than categories as in the example you have just done? Let us use the data for how many letters are in your first name? the variable is in column L and called. *namequestion_numletters*

10. First produce a frequency table with the fast food data. Try to do this on your own first but the instructions below will lead you through this.
11. Go to your *Barcharts sheet, scroll down to row 20* and produce the outline of our table as below:



	A	B	C	D
18				
19				
20		Number of Letters in Name	Number of children	
21		3		
22		4		
23		5		
24		6		
25		7		
26		8		
27		9		
28		10		
29		11		
30				

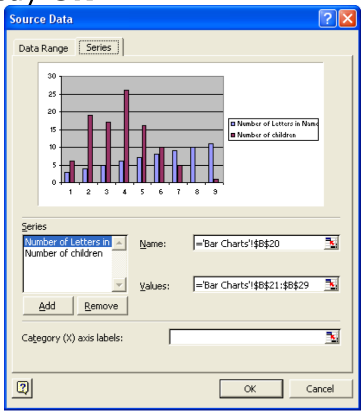
12. In C21 type the following formula which will tell the spreadsheet to go to sheet 1 where our data is and count the responses giving this answer from the column L

    =COUNTIF(Sheet1!L:L, 3)

    You should find that 6 children gave this response.
    Copy the formula into the cells C22 to 29 changing the number each time.
    To check we have got all 100 children type the following formula into the total cell in C30

    =SUM(C21:C29)

8. Now we can draw our Bar Chart; Highlight the cells where your data is which is cells B20 to C29 if you have followed the instructions above and use the chart wizard to draw a bar chart (column graph).
9. How does it differ from the one you drew for table 1? What has it done? You should be able to see now why Excel can be very hard to produce even simple bar charts!.
10. To get rid of the superfluous "Number of letters in Children's name" data you need to put your cursor somewhere over the bar chart, right click **Source Data > Click on the Series tab** highlight the Number of letters series and remove it, say **OK**



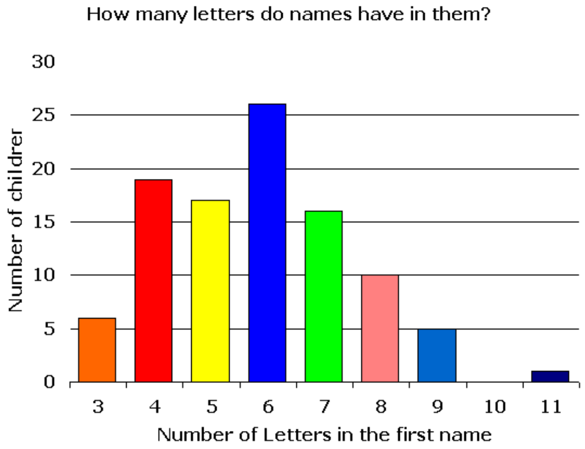
Series: Number of Letters in / Number of children

Name: ='Bar Charts'!$B$20

Values: ='Bar Charts'!$B$21:$B$29

Category (X) axis labels:

11. To get Excel to pick up the correct labels for each bar it needs text rather than a number so in B21 change the 3 into Three and in B29 change the 11 into eleven. Now redraw your bar chart which should now be rather nearer the bar chart you expected.
12. As a final twist change you can now change the 3 and 11 back to numbers in your table and the graph will change!
13. Experiment with the menus to try and produce a nice bar chart

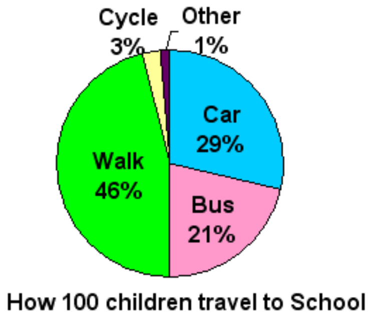


Experiment with some of the other data in the spreadsheet to perfect your bar chart drawing skills.

Pie charts: getting the right picture

This exercise is similar to the one in the last section about Bar Charts and we use the same file of data.
Remember that if a Pie Chart is going to be drawn you must have mutually exclusive data as described in the first part of this chapter.
Go to a new sheet in your spreadsheet and name it *pie charts*.

1. Create a frequency table for the travel to school data (column N) in cells B2 to C7. Put the titles in row 2. Use the COUNTIF function just as you did in the bar charts section. Check you have all 100 items of data by putting a sum function in cell C8.
2. Click on B2 and drag to C9 so that the cells are highlighted.
3. Click on the chart wizard and choose pie chart.
4. Follow the instructions through to complete the drawing of the pie chart, insert any title etc at the appropriate page. You should finish up with a sensible pie chart.



5. But suppose we try to do the same thing with the data on the number of letters in names we used in the Bar chart section? Copy and paste the frequency table you produced for your second bar chart and put it into your Pie Charts sheet.. Highlight the table. Go to chart wizard and choose

pie charts. Click **Next > … > Finish.** What has happened? – look carefully You should find 9 sections although you only have 8 different numbers of letters and the proportions in the pie chart are not right (for example the third one down which is 6 letters but represented as 3 on your pie chart should be over ¼ of the pie). Can you work out what Excel has done?

6. Now try replacing 3 with three and 11 with eleven and drawing the pie chart. This should be better. Why? Once again you can now replace the words with numbers again and the pie chart will be fine.

Calculating Summary Statistics

Mean, Median and Quartiles

Label a new sheet *Summary.* We will once more use the original data set.

Mean
Finding the average height of our random sample

1. In cell **E102** type: =AVERAGE(E2:E101). This will give you the mean of the height data.
2. Move the cursor to the bottom right of cell **E102** until it becomes a black cross. Now drag this formula across cells **F102** and **G102** to get the means of foot length and belly button to floor. Note that AVERAGE function gives the arithmetic mean of the range chosen.

Median and Quartiles

1. Median has its own function, MEDIAN, but you can also use the quartile function QUARTILE and find the second quartile.
2. In **E103** type: =MEDIAN(E2:E101). Copy this across to the next two cells in the same row. You should get the medians of the data sets above
3. What rule is Excel using for the median?

Excel is not designed to produce Box and Whisker plots but with a little fudging we can make it do so. Follow the instructions to get a box and whisker plot of the age males and females in our sample want to live until:

4. Copy the 2 columns of C "*gender*" and K "*desiredage*" data into columns A and B on the *Summary* sheet

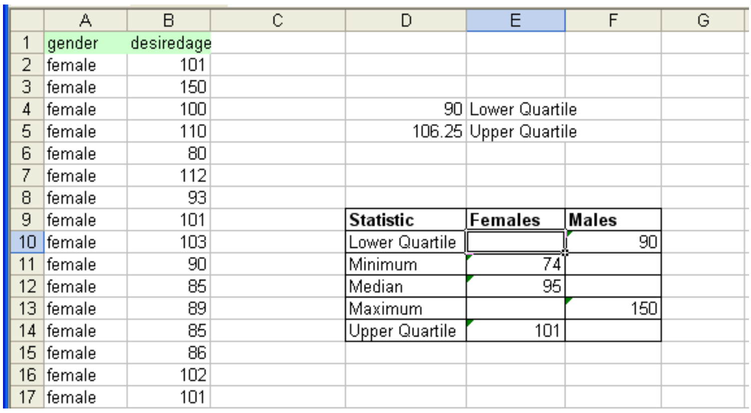
5. Work out the using the following formulae. In D4 type: =QUARTILE(A24:A101,1) to get the lower quartile (shown by the '1').
6. In D5 type =QUARTILE(A2:A101,3) the upper quartile, shown by the '3',
7. What rule are they using for quartiles? [They are not calculated using both (n+1)/4 and 3(n+1)/4].

(Note: Excel uses (n+1)/4 for Q1 (the first quartile), (n+1)/2 for Q2 (the median) and (3n+1)/4 for Q3 (the third quartile).

8. We want to compare male and female data so move the cursor to cell A1 and sort the data by clicking **Data > Sort** choose *gender* as the variable to sort and click **OK**. Data for all the females should now appear at the top of the list.

9. Now generate the boxplot statistics from the data sets. The results have to be displayed in the order as shown below for the 'trick' to work. Remember female data are in cells B2 to B46 and the male data are in B47 to B101.

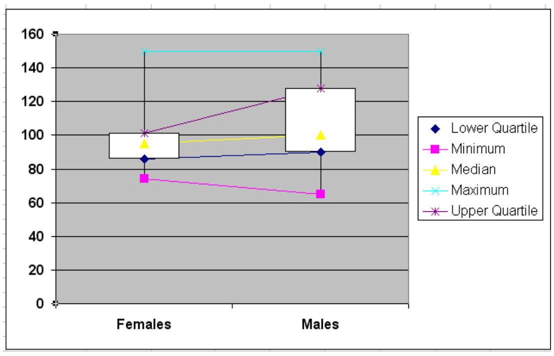
10. Use the functions MIN, MAX, QUARTILE and MEDIAN. Below is what your table should look like with some entries still to do. Complete your table.



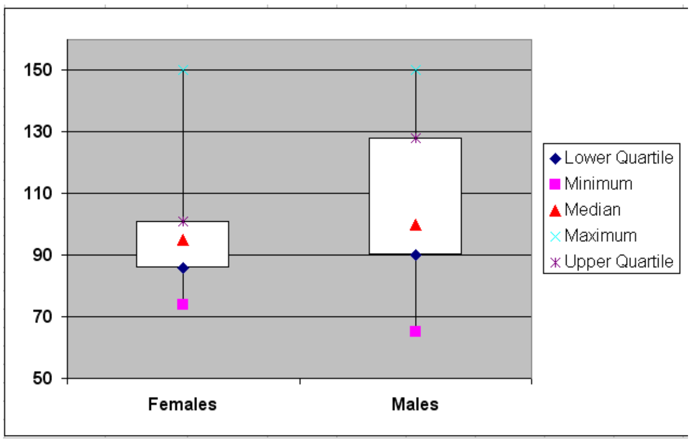
	A	B	C	D	E	F	G
1	gender	desiredage					
2	female	101					
3	female	150					
4	female	100		90	Lower Quartile		
5	female	110		106.25	Upper Quartile		
6	female	80					
7	female	112					
8	female	93					
9	female	101		**Statistic**	**Females**	**Males**	
10	female	103		Lower Quartile		90	
11	female	90		Minimum	74		
12	female	85		Median	95		
13	female	89		Maximum		150	
14	female	85		Upper Quartile	101		
15	female	86					
16	female	102					
17	female	101					

11. Select all the information including the labels, then click ***Insert*** > ***Chart*** and choose ***Line***. Select the sub-type '**Line with markers displayed at each data value**'. Click ***Next***, select '**Series in Rows'** and then click ***Finish***.

12. Now you need to do some tweaking. Right click on one of the lines or Data Series. Select ***Format Data Series***, go to ***Options*** and check **High-low lines** and **Up-down lines**. The **Gap Width** may be adjusted to change the width of the boxes. You should get something that looks like this:



13. The data lines may be removed by selecting each data series in turn and removing the lines. Right click a data series > ***Format Data Series*** > ***Patterns***> ***Line*** and choose **None**.
Use the menus to tweak further and you have your boxplot!


Females
Males
Lower Quartile
Minimum
Median
Maximum
Upper Quartile