MAS187/AEF258

University of Newcastle upon Tyne

2005-6

Contents

Chapter 3

More graphical methods for presenting data

3.1 Introduction

We have seen some basic ways in which we might present data graphically. These methods will often provide the mainstay of business presentations. There are, however, other techniques which are useful and offer advantages in some applications over histograms and bar charts.

3.2 Percentage Relative Frequency Histograms

When we produced frequency tables in Chapter 2, we included a column for percentage relative frequency. This contained values for the frequency of each group, relative to the overall sample size, expressed as a percentage. Recall the data on service time (in seconds) for calls to a credit card service centre:

214.8412	220.6484	216.7294	195.1217	211.4795
195.8980	201.1724	185.8529	183.4600	178.8625
196.3321	199.7596	206.7053	203.8093	203.1321
200.8080	201.3215	205.6930	181.6718	201.7461
180.2062	193.3125	188.2127	199.9597	204.7813
198.3838	193.1742	204.0352	197.2206	193.5201
205.5048	217.5945	208.8684	197.7658	212.3491
209.9000	197.6215	204.9101	203.1654	192.9706
208.9901	202.0090	195.0241	192.7098	219.8277
208.8920	200.7965	191.9784	188.8587	206.8912

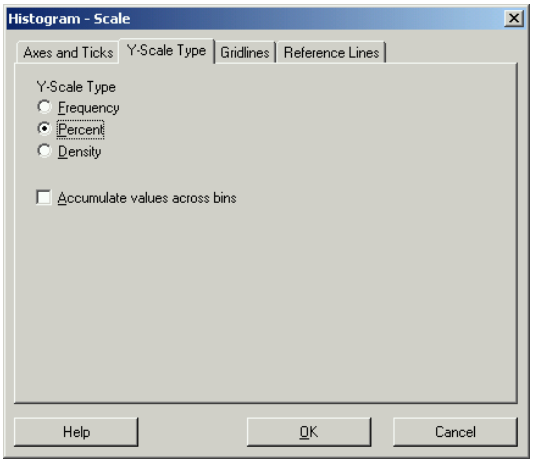
A percentage relative frequency table for these data is

Service time	Frequency	Relative Frequency (%)
$175 \leq time < 180$	1	2
$180 \leq time < 185$	3	6
$185 \leq time < 190$	3	6
$190 \leq time < 195$	6	12
$195 \leq time < 200$	10	20
$200 \leq time < 205$	12	24
$205 \leq time < 210$	8	16
$210 \leq time < 215$	3	6
$215 \leq time < 220$	3	6
$220 \leq time < 225$	1	2
Totals	50	100

You can easily plot these data like an ordinary histogram, except, instead of using frequency on the vertical axis (y-axis), you use the percentage relative frequency.

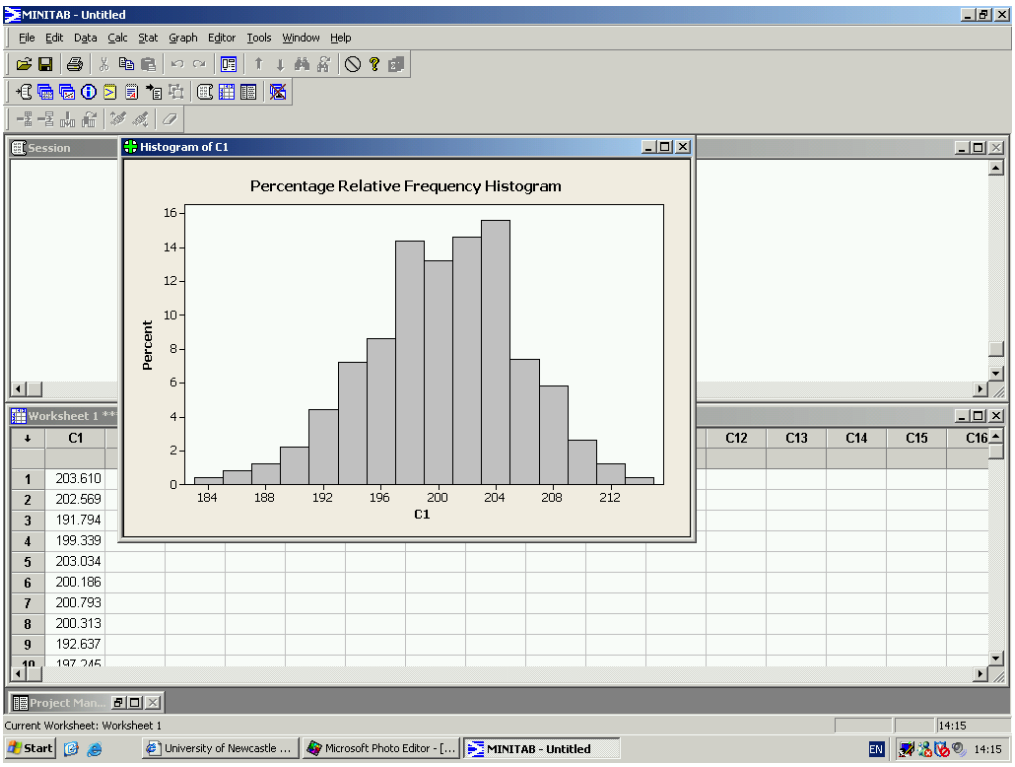
This can be done in MINITAB as follows.

1. Place the data to be graphed in a column of the worksheet. For illustrative purposes 500 observations have been generated in column C1.
2. `Graph > Histogram`
3. As with ordinary histograms, select the `Simple` graph format, click on **OK**, select column C1 under `Graph variables`.
4. Select `Scale...` then `Y-Scale Type` and check the `Percent` button



5. Click on **OK** and again on **OK**.

This produces the following histogram:

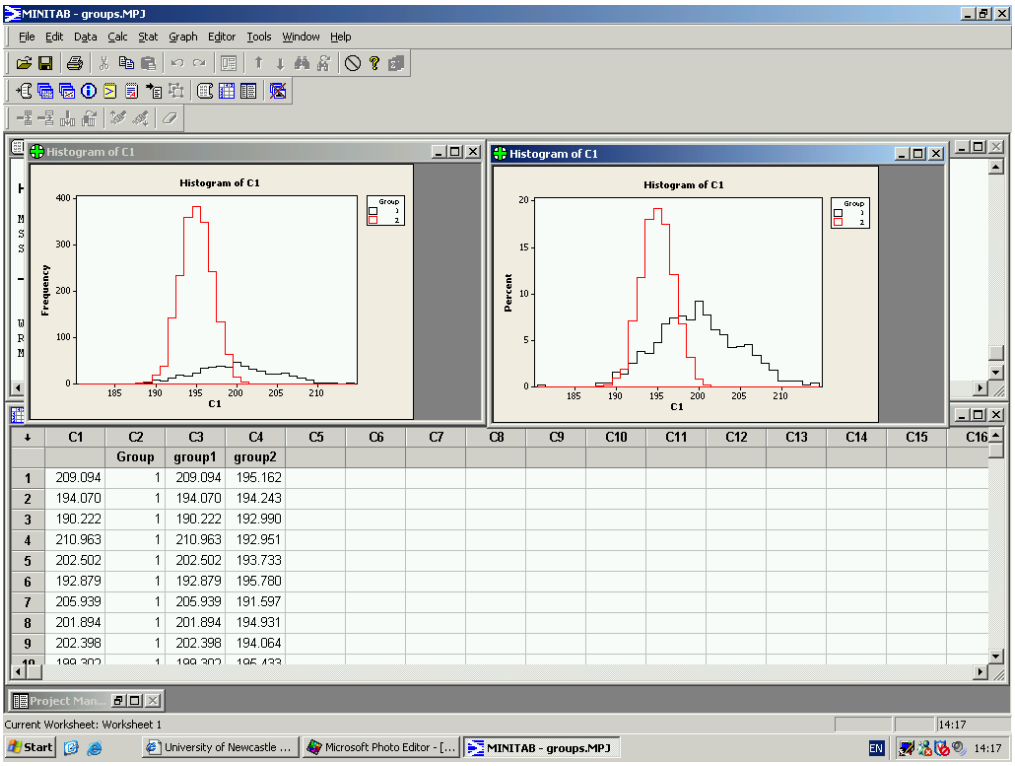


Note that the y-axis now contains the relative percentages rather than the frequencies.

You might well ask why we would want to do this? These percentage relative frequency histograms are useful when comparing two samples that have different numbers of observations. If one sample were larger than the other then a frequency histogram would show a difference simply because of the larger number of observations. Looking at percentages removes this difference and enables us to look at relative differences. It is really just a matter of making the vertical scales comparable.

In the following graph there are data from two groups and four times as many data points for one group as the other. The left-hand plot shows an ordinary histogram and it is clear that the

comparison between groups is masked by the quite different sample sizes. The right-hand plot shows a histogram based on (percentage) relative frequencies and this enables a much more direct comparison of the distributions in the two groups.



	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
		Group	group1	group2												
1	209.094	1	209.094	195.162												
2	194.070	1	194.070	194.243												
3	190.222	1	190.222	192.990												
4	210.963	1	210.963	192.951												
5	202.502	1	202.502	193.733												
6	192.879	1	192.879	195.780												
7	205.939	1	205.939	191.597												
8	201.894	1	201.894	194.931												
9	202.398	1	202.398	194.064												

Overlaying histograms on the same graph can sometimes not produce such a clear picture, particularly if the values in both groups are close or overlap one another significantly.

3.3 Relative Frequency Polygons

These are a natural extension of the relative frequency histogram. They differ in that, rather than drawing bars, each class is represented by one point and these are joined together by straight lines. The method is similar to that for producing a histogram.

1. Produce a percentage relative frequency table.
2. Draw the axes
   - The x-axis needs to contain the full range of the classes used.
   - The y-axis needs to range from 0 to the maximum percentage relative frequency.
3. Plot points, pick the mid point of the class interval on the x-axis and go up until you reach the appropriate percentage value on the y-axis and mark the point. Do this for each class.
4. Join the points together with straight lines.

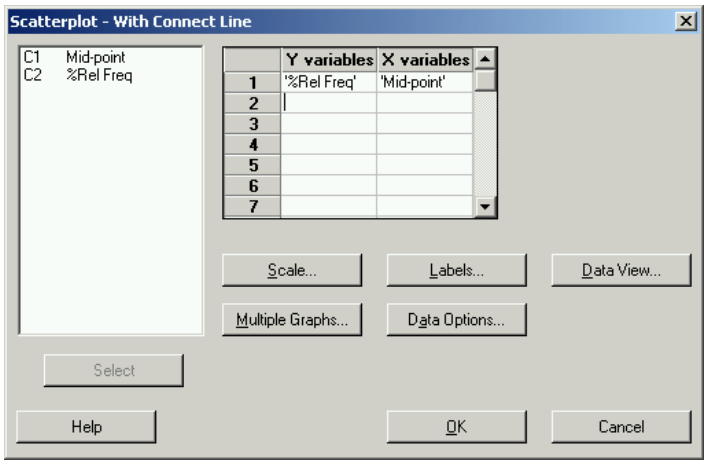
Consider the following simple example.

Class Interval	Mid Point	% Relative Frequency
$0 \leq x < 10$	5	10
$10 \leq x < 20$	15	20
$20 \leq x < 30$	25	35
$30 \leq x < 40$	35	25
$40 \leq x < 50$	45	10

We can draw this easily by hand.

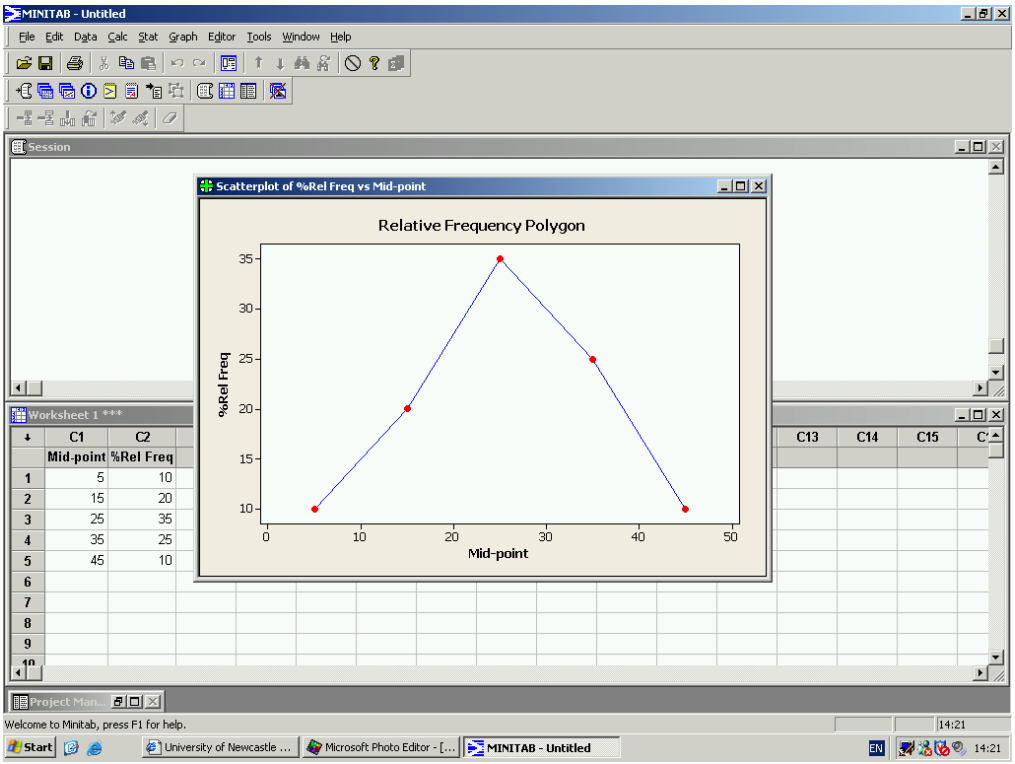
Alternatively you can use **MINITAB**.

1. Place the data in the worksheet using column C1 for the mid-points and column C2 for the percentage relative frequencies.
2. `Graph > Scatterplot...`
3. Select the `With Connect Line` option and click on **OK**.
4. Enter the column with the percentage frequencies (C2) under `Y variables` and the column with the midpoints (C1) under `X variables`



5. Add a title by clicking on **Labels...** etc.
6. Click on **OK**.

These instructions produce the graph

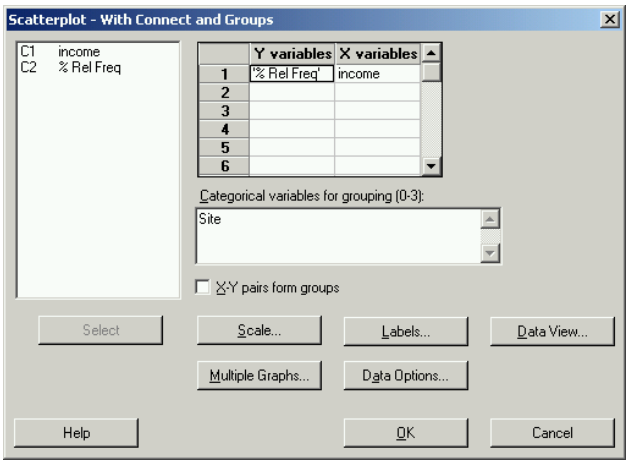


These percentage relative frequency polygons are of most use however for comparison between two samples. Consider the following data on gross weekly income collected from two sites in Newcastle. Let us suppose that many more responses were collected in Jesmond so that a direct comparison of the frequencies using a standard histogram is not appropriate. Instead we use relative frequencies.

Weekly Income (£)	**West Road** (%)	**Jesmond Road** (%)
$0 \leq income < 100$	9.3	0.0
$100 \leq income < 200$	26.2	0.0
$200 \leq income < 300$	21.3	4.5
$300 \leq income < 400$	17.3	16.0
$400 \leq income < 500$	11.3	29.7
$500 \leq income < 600$	6.0	22.9
$600 \leq income < 700$	4.0	17.7
$700 \leq income < 800$	3.3	4.6
$800 \leq income < 900$	1.3	2.3
$900 \leq income < 1000$	0.0	2.3

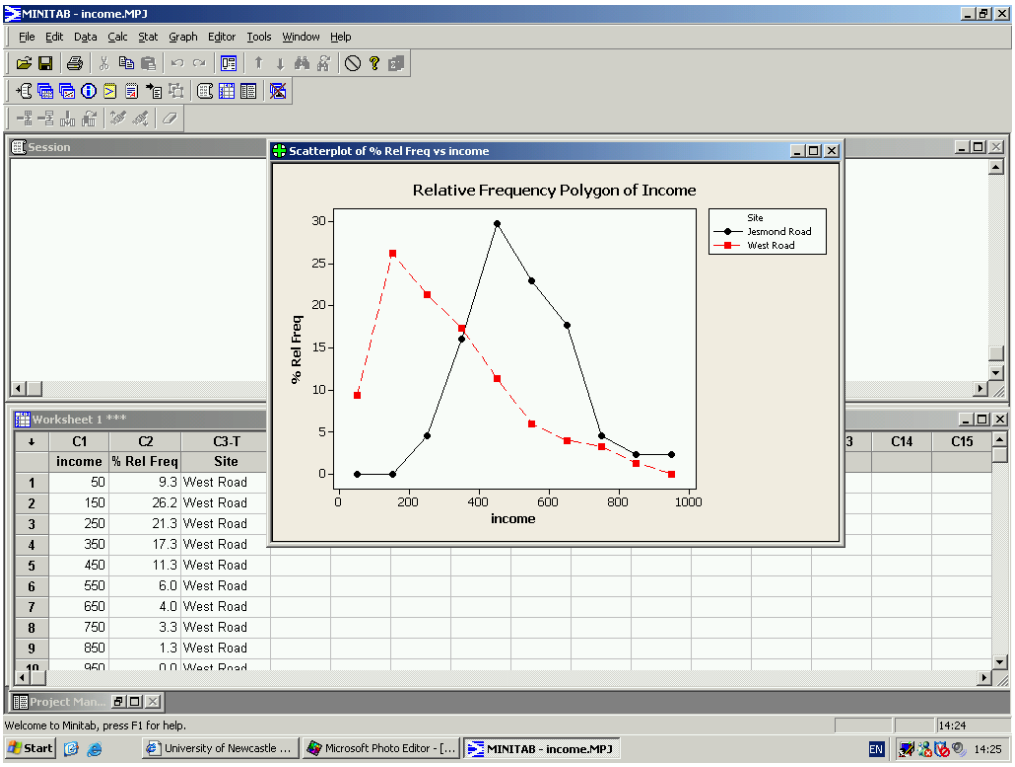
We can produce a graph containing polygons for both locations using MINITAB instructions very similar to those above:

1. Place the data in the worksheet using column C1 for the mid-points, column C2 for the percentage relative frequencies and column C3 for the site where the data were taken.
2. `Graph > Scatterplot...`
3. Select the `With Connect and Groups` option and click on **OK**
4. Enter the column with the percentage frequencies (C2) under `Y variables` and the column with the midpoints (C1) under `X variables`. Also enter the `Site` column (C3) in the box for `Categorical variables for grouping`



5. Add a title by clicking on **Labels...** etc.
6. Click on **OK**.

The polygon produced looks like



We can clearly see the differences between the two samples. The line connecting the circles represents the data from the West Road and the line connecting the boxes represents those for Jesmond Road. The distribution of incomes on the West Road is skewed towards lower values, whilst those on Jesmond Road are more symmetric. The graph clearly shows that income in the Jesmond Road area is higher than that on the West Road.

3.4 Cumulative Frequency Polygons (Ogive)

Cumulative percentage relative frequency is also a useful tool. The cumulative percentage relative frequency is simply the sum of the percentage relative frequencies at the end of each class interval. Consider the example from the previous section.

Class Interval	**% Relative Frequency**	**Cumulative % Relative Frequency**
$0 \leq x < 10$	10	10
$10 \leq x < 20$	20	30
$20 \leq x < 30$	35	65
$30 \leq x < 40$	25	90
$40 \leq x < 50$	10	100

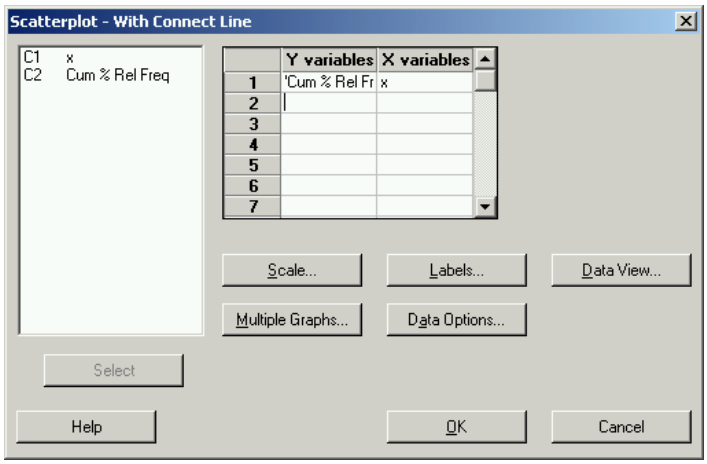
At the upper limit of the first class the cumulative % relative frequency is simply the % relative frequency in the first class 10. However at the end of the second class, at 20, the cumulative % relative frequency is $10 + 20 = 30$. The cumulative % relative frequency at the end of the last class must be 100.

The corresponding graph, or *ogive*, is simple to produce by hand.

1. Draw the axis.
2. Label the x-axis with the full range of the data and the y-axis from 0 to 100%.
3. Plot the cumulative % realtive frequency at the end point of each class.
4. Join the points, starting at 0% at the lowest class boundary.

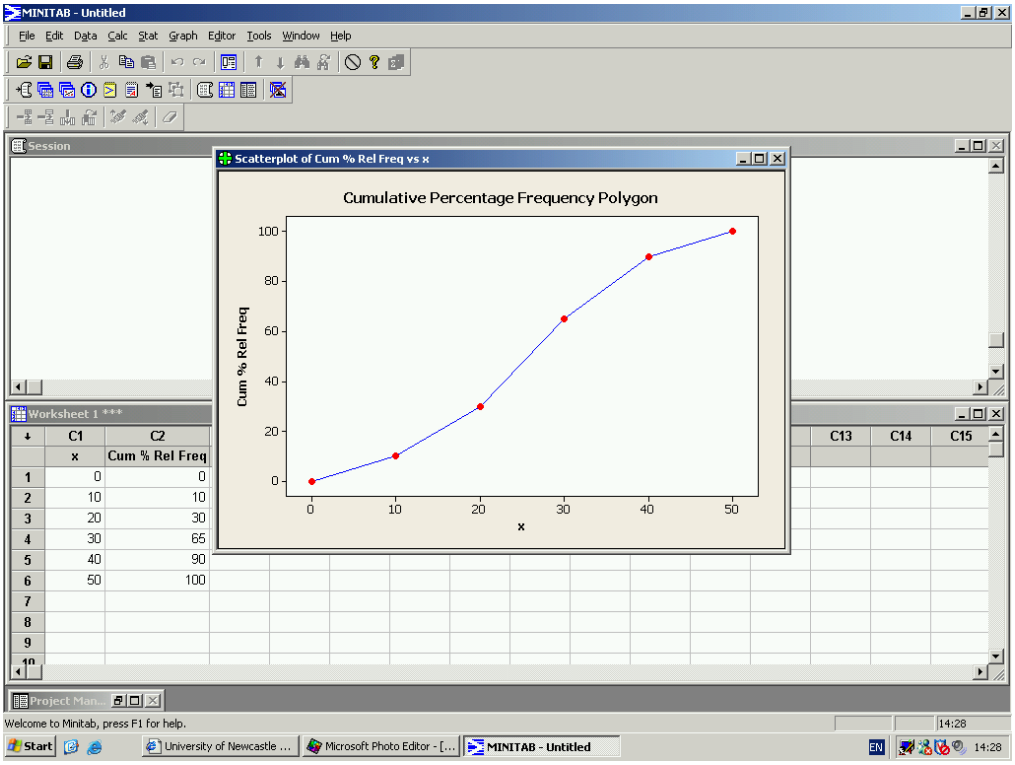
This graph can be produced using the following **MINITAB** instructions:

1. In column **C1**, enter the end points of the class intervals, as well the starting point of the smallest class.
2. In column **C2**, enter 0 against the starting point and the cumulative percentage relative frequencies against the relevant end point.
3. `Graph > Scatterplot...`
4. Select the `With Connect Line` option and click on **OK**
5. Enter the column with the percentage frequencies (C2) under `Y variables` and the column with the midpoints (C1) under `X variables`

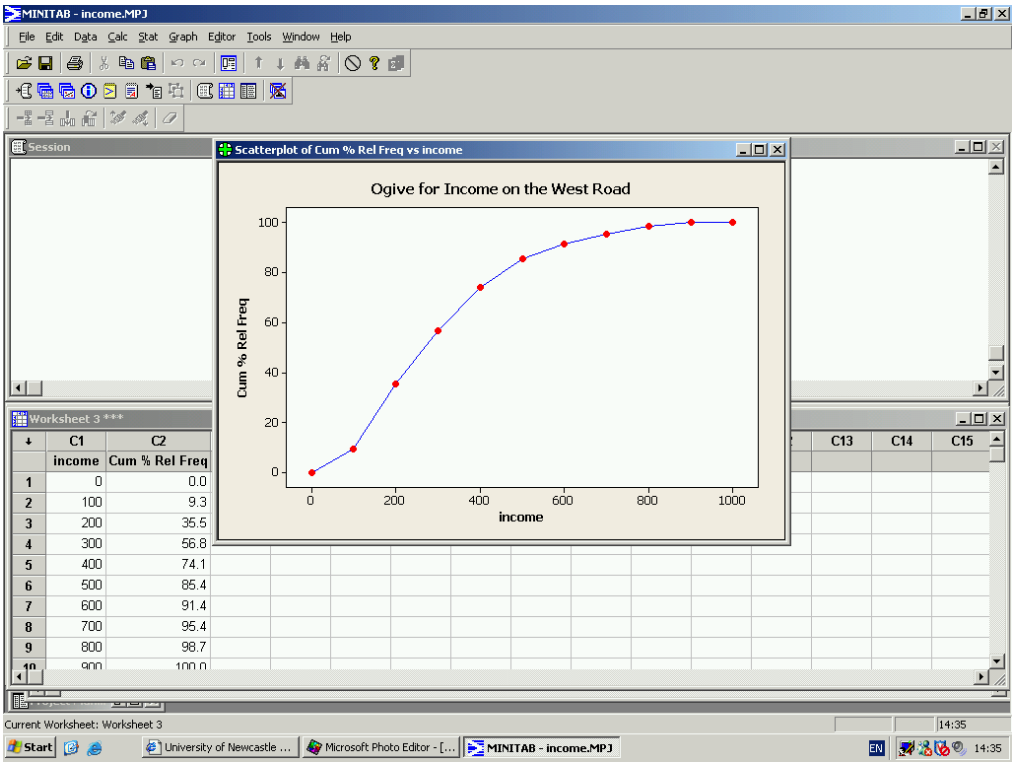


6. Add a title by clicking on **Labels...** etc.
7. Click on **OK**.

This produces the following graph:



Applying this procedure to the income data from the West Road survey gives the ogive:



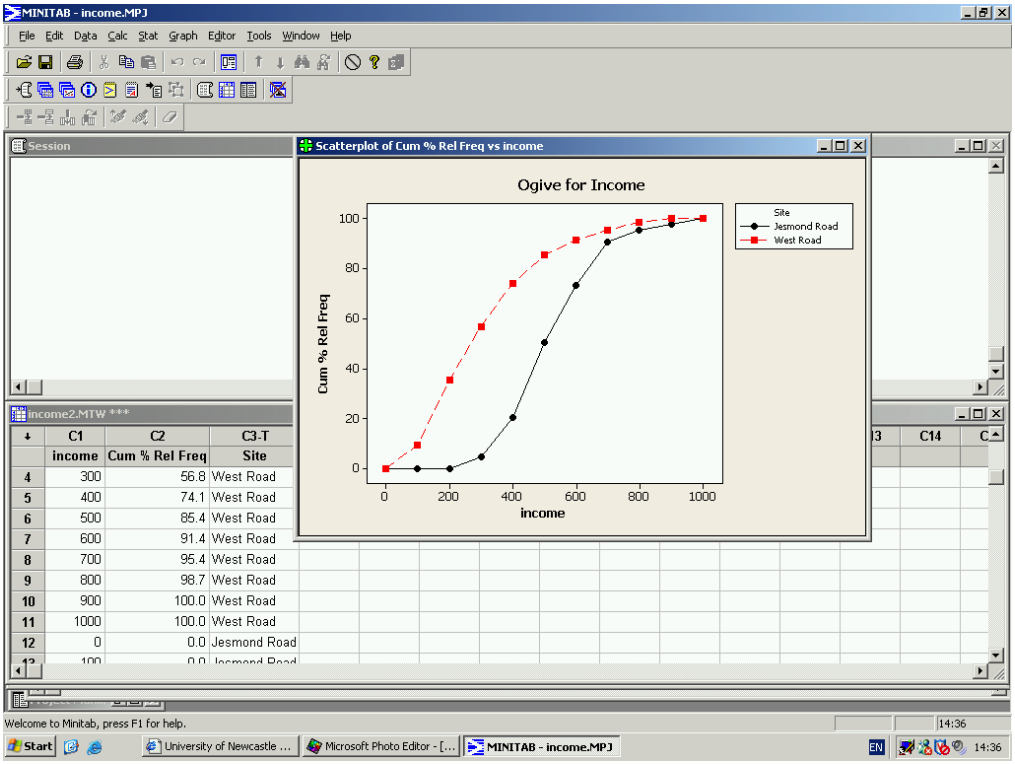
This graph instantly tells you many things. To see what percentage of respondents earn less than £x per week.

1. Find x on the x-axis and draw a line up from this value until you reach the ogive.

2. From this point trace across to the y-axis.
3. Read the percentage from the y-axis.

If we wanted to know what percentage of respondents in the survey on the West Road earn less than £250 per week, we simply find £250 on the x-axis, trace up to the ogive and then trace across to the y-axis and we can read a figure of about 47%. The process obviously works in reverse. If we wanted to know what level of income 50% of respondents earned, we would trace across from 50% to the ogive and then down to the x-axis and read a value of about £300.

Ogives can also be used for comparison purposes. The following plot contains the ogives for the income data at both the West Road and Jesmond Road sites.



	C1	C2	C3-T
	income	Cum % Rel Freq	Site
4	300	56.8	West Road
5	400	74.1	West Road
6	500	85.4	West Road
7	600	91.4	West Road
8	700	95.4	West Road
9	800	98.7	West Road
10	900	100.0	West Road
11	1000	100.0	West Road
12	0	0.0	Jesmond Road

It clearly shows the ogive for Jesmond shifted to the left of that for the West Road. This tells us that the surveyed incomes are higher on Jesmond Road. We can compare the percentages of people earning different income levels between the two sites quickly and easily.

This technique can also be used to great effect for examining the changes between before and after the introduction of a marketing strategy. For example, daily sales figures of a product for a period before and after an advertising campaign might be plotted. Here a comparison of the two ogives can be used to help assess whether or not the campaign has been successful.

3.5 Pie Charts

Pie charts are simple diagrams for displaying categorical or grouped data. These charts are commonly used within industry to communicate simple ideas, for example market share. They are

used to show hte proportions of a whole. They are best used where there are only a handful of categories to display.

A pie chart consists of a circle divided into segments, one segment for each category. The size of each segment is determined by the frequency of the category and measured by the angle of the segment. As the total number of degrees in a circle is 360, the angle given to a segment is $360°$ times the fraction of the data in the category, that is

$$\text{angle} = \frac{\text{Number in category}}{\text{Total number in sample}(n)} \times 360.$$

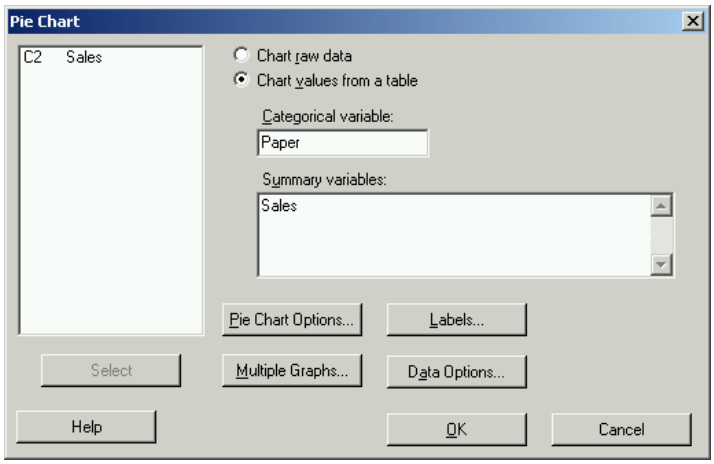
Consider again the data on newspaper sales to 650 students.

Paper	**Frequency**	**Degrees**
The Times	140	77.5
The Sun	200	110.8
The Sport	50	27.7
The Guardian	120	66.5
The Financial Times	20	11.1
The Mirror	80	44.3
The Daily Mail	10	5.5
The Independent	30	16.6
Totals	650	360.0

The pie chart is constructed by first drawing a circle and then dividing it up with segments with angles calculated using this formula.

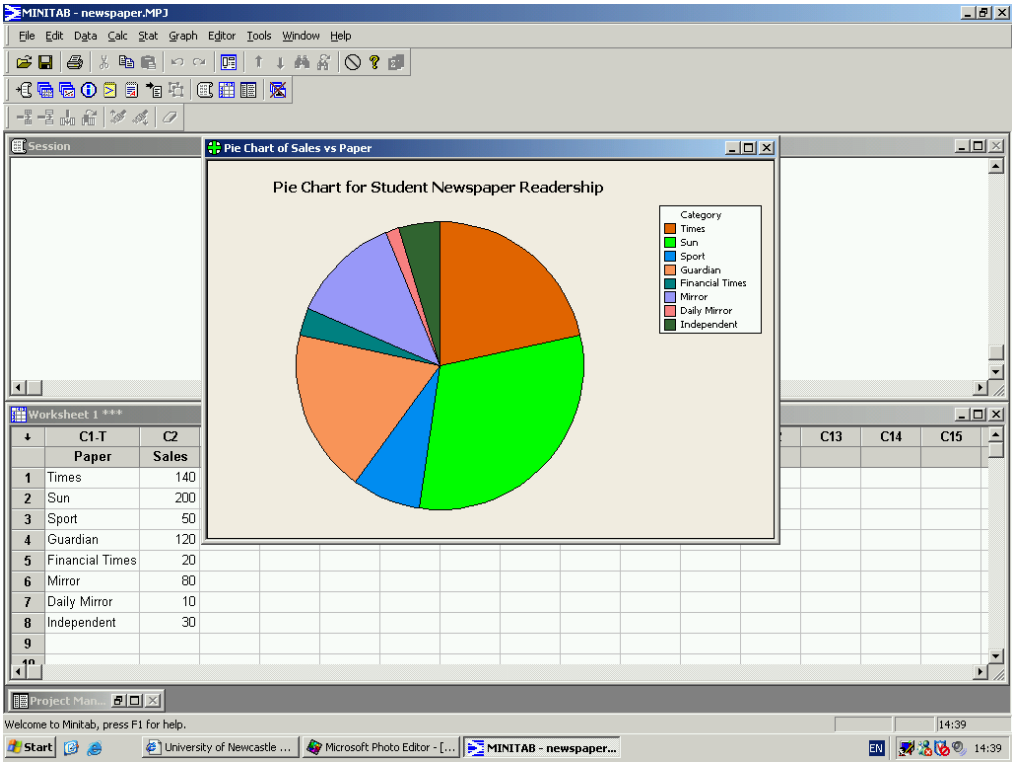
In **MINITAB**, a pie chart for these data would be obtained as follows:

1. Enter data into worksheet, with category name in column C1 and frequencies in column C2.
2. `Graph > Pie Chart...`
3. Check the button for `Chart values from a table`
4. Enter the `Category` column under `Categorical variable:` and the `Frequency` column under `Summary variables:`



5. Add a title and click **OK**

This produces the following pie chart



It shows that The Sun, The Times and The Guardian are the most popular papers.

Note that the pie chart is a simple circle. Some computer software will draw “perspective” pie charts, pie charts with slices detached etc. It is best to avoid such gimmicks which merely obscure the information contained in the chart.

3.6 Time Series Plots

So far we have only considered data where we can (at least for some purposes) ignore the order in which the data come. Not all data are like this. One exception is the case of time series data, that is, data collected over time. Examples include monthly sales of a product, the price of a share at the end of each day or the air temperature at midday each day. Such data can be plotted simply using time as the x-axis.

Consider the following data on the number of computers sold (in thousands) by quarter (January-March, April-June, July-September, October-December) at a large warehouse outlet.

Quarter	Units Sold
Q1 2000	86.7
Q2 2000	94.9
Q3 2000	94.2
Q4 2000	106.5
Q1 2001	105.9
Q2 2001	102.4
Q3 2001	103.1
Q4 2001	115.2
Q1 2002	113.7
Q2 2002	108.0
Q3 2002	113.5
Q4 2002	132.9
Q1 2003	126.3
Q2 2003	119.4
Q3 2003	128.9
Q4 2003	142.3
Q1 2004	136.4
Q2 2004	124.6
Q3 2004	127.9

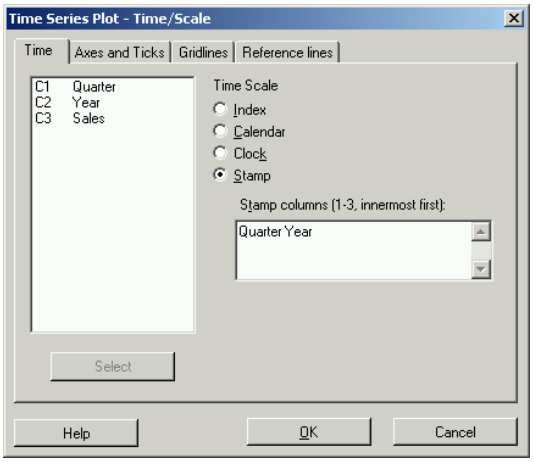
By hand, a time series plot is constructed as follows:

1. Draw the x-axis and label over the time scale.
2. Draw the y-axis and label with an appropriate scale.
3. Plot each point according to time and value.
4. Draw lines connecting all points.

In **MINITAB** the plot can be obtained using

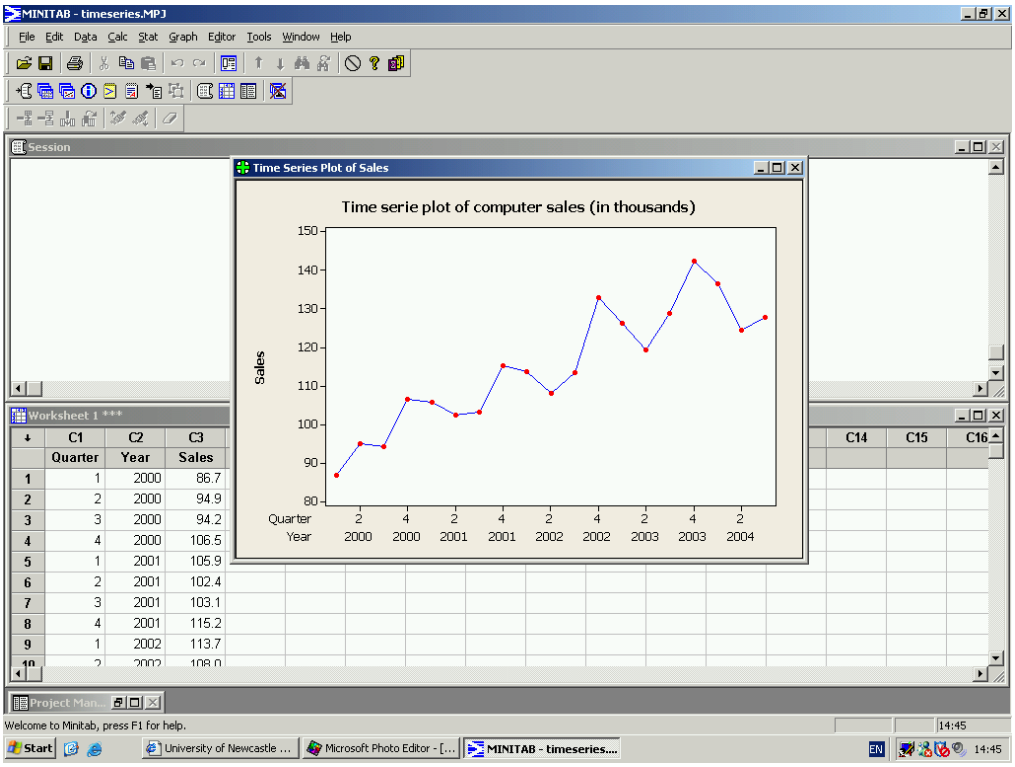
1. Enter the data into a worksheet, with the `Quarter, Year` and `Sales` in columns C1, C2 and C3.

2. Click on `Graph` and select `Time Series Plot...`
3. Select the `Simple` graph format and click on `OK`.
4. Enter the `Sales` column in the `Series:` box.
5. Now click on `Time/Scale...`, check the `Stamp` button and enter the `Quarter` and `Year` columns under `Stamp columns`



6. Click **OK**.
7. Add a title etc.
8. Click **OK**.

The time series plot is



The plot clearly shows us two things: firstly that there is an upwards trend to the data and secondly that there is some regular variation around this trend. We will come back to more sophisticated techniques for analysing time series data later in the course.

3.7 Scatter Plots

The final type of graph we are going to look at is *scatter plots*. These are used to plot two variables which you believe might be related, for example, height and weight, advertising expenditure and sales or age of machinery and maintenance costs.

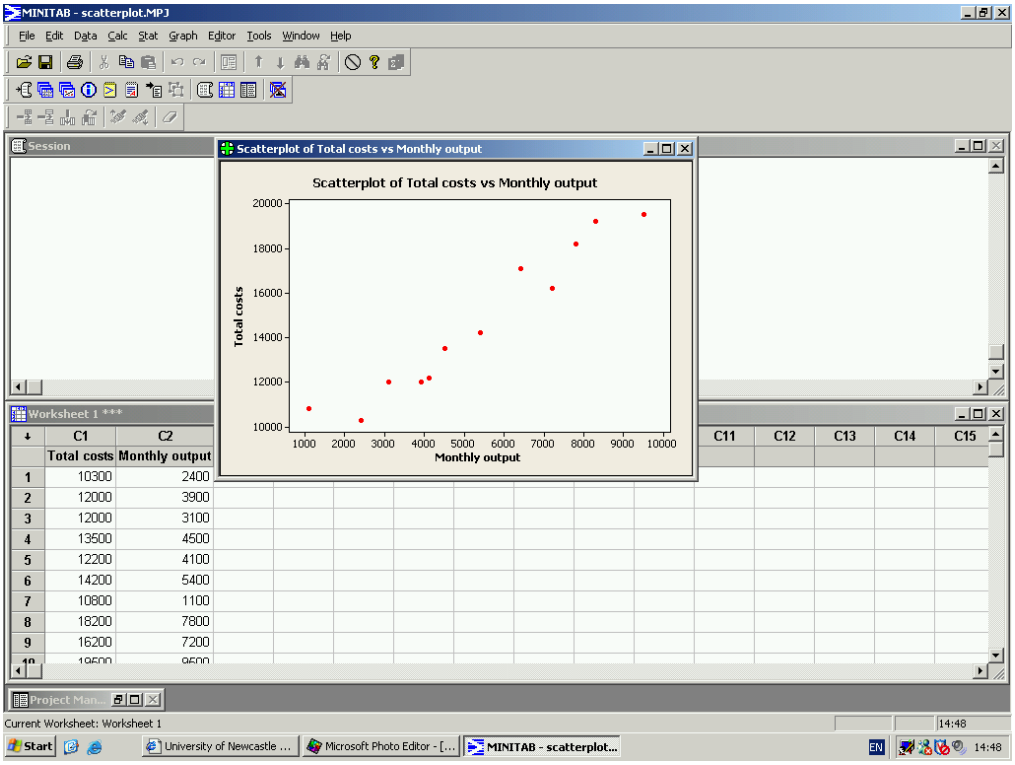
Consider the following data for monthly output and total costs at a factory.

Total costs (£)	**Monthly Output**
10300	2400
12000	3900
12000	3100
13500	4500
12200	4100
14200	5400
10800	1100
18200	7800
16200	7200
19500	9500
17100	6400
19200	8300

If you were interested in the relationship between the cost of production and the number of units produced you could easily plot this by hand.

1. The "response" variable is placed on the y-axis. Here we are trying to understand how total costs relate to monthly output and so the response variable is "total costs".
2. The variable that is used to try to explain the response variable (here, monthly output) is placed on the x-axis.
3. Plot the pairs of points on the graph.

This graph can be produced using **MINITAB**. (Select *Graph* then *Scatter Plot* then *Simple* and insert the required variables).



The plot highlights a clear relationship between the two variables: the more units made, the more it costs in total. This relationship is shown on the graph by the upwards trend within the data as monthly output increases so does total cost. A downwards sloping trend would imply that as output increased total costs declined, an unlikely scenario. This type of plot is the first stage of more sophisticated techniques which we will develop later in the course.

3.8 Exercises 3

1. Consider the following data for daily sales at a small record shop, before and after a local radio advertising campaign.

Daily Sales	Before	After
$1000 \leq sales < 2000$	10	7
$2000 \leq sales < 3000$	30	10
$3000 \leq sales < 4000$	40	25
$4000 \leq sales < 5000$	20	35
$5000 \leq sales < 6000$	15	37
$6000 \leq sales < 7000$	12	40
$7000 \leq sales < 8000$	10	20
$8000 \leq sales < 9000$	8	10
$9000 \leq sales < 10000$	0	5
Totals	145	189

   (a) Calculate the percentage relative frequency for before and after.

   (b) Calculate the cumulative percentage relative frequency for before and after.

   (c) Plot the relative frequency polygons for both on one graph and **comment**.

   (d) Plot the ogives for both on one graph.

   (e) Find the level of sales before and after that are reached on 25%, 50% and 75% of days.

2. The following table shows data for the monthly sales of a small department store and their monthly advertising expenditure.

Advertising Expenditure	Monthly Sales
52000	1200000
20500	650000
35000	870000
76000	1600000
65000	1200000
27000	850000
55000	1100000
39000	1000000
45000	1110000
27000	700000
38000	900000
52000	1150000

   Plot these data on an appropriate graph and comment on the relationship between advertising expenditure and monthly sales.

3. The data in table 3.1 give the amounts, in £, spent by 200 customers at a "Farmer's Market" stall who bought at least one item. Use a histogram to display the data. The data are also available from the module Web page.

19.32	5.74	12.52	8.57	9.97	5.43	5.76	12.63	0.67	10.92
9.59	0.39	19.92	11.25	7.71	10.91	4.77	23.88	18.13	8.25
3.84	5.17	21.78	11.06	8.29	12.43	16.68	12.03	4.29	11.06
5.73	6.95	10.92	5.67	19.66	12.69	19.84	5.78	7.33	3.42
9.13	2.80	5.11	4.35	12.58	15.71	24.78	13.88	5.38	14.59
11.98	14.48	15.18	13.37	7.64	5.10	1.54	6.46	4.85	7.54
14.45	11.26	6.48	3.50	6.59	3.30	8.35	2.53	8.19	6.39
13.41	4.96	13.18	46.59	26.42	14.81	4.21	12.89	14.92	18.02
7.82	6.45	3.92	2.28	3.97	14.35	6.72	8.84	4.88	1.88
7.34	2.75	9.71	4.29	11.37	10.00	5.04	5.76	8.74	2.14
15.11	1.37	3.68	10.99	2.75	20.77	7.39	5.92	12.57	6.57
11.56	10.86	7.37	4.44	9.24	18.48	3.71	9.19	10.61	7.85
12.57	6.65	10.54	14.54	14.00	9.73	14.37	2.56	2.01	0.85
8.39	11.66	4.65	17.29	6.12	18.36	4.89	5.89	10.44	5.35
12.10	8.43	26.18	8.92	9.79	10.93	5.92	18.00	6.01	2.68
10.40	0.91	11.46	16.73	19.16	12.06	15.22	10.53	6.78	6.33
7.67	4.76	7.38	21.10	10.86	14.88	6.35	8.02	5.29	1.16
19.93	3.38	4.08	5.88	5.32	9.41	29.92	17.19	11.72	10.10
8.01	3.98	4.95	2.13	1.57	10.08	17.81	5.78	4.77	17.80
4.31	20.42	2.28	2.40	26.99	9.17	2.86	14.58	36.25	20.96

Table 3.1: Amounts spent at a farmer's market stall

Industry	UK	Ireland
Agriculture	2.7	23.2
Mining	1.4	1.0
Manufacturing	30.2	20.7
Power supplies	1.4	1.3
Construction	6.9	7.5
Service Industries	16.9	16.8
Finance	5.7	2.8
Social and personal services	28.3	20.8
Transport and communications	6.4	6.1

Table 3.2: Percentages employed in different industries

Month	1980	1981	1982	1983	1984	1985
January	1998	1924	1969	2149	2319	2137
February	1968	1959	2044	2200	2352	2130
March	1937	1889	2100	2294	2476	2154
April	1827	1819	2103	2146	2296	1831
May	2027	1824	2110	2241	2400	1899
June	2286	1979	2375	2369	3126	2117
July	2484	1919	2030	2251	2304	2266
August	2266	1845	1744	2126	2190	2176
September	2107	1801	1699	2000	2121	2089
October	1690	1799	1591	1759	2032	1817
November	1808	1952	1770	1947	2161	2162
December	1927	1956	1950	2135	2289	2267

Table 3.3: Jeans sales in the UK

4. Table 3.2 shows the percentages employed in different industries in the UK and Ireland in the late 1970s. Use pie charts to compare the two countries' proportions.

5. Table 3.3 shows the estimated monthly sales of pairs of jeans, in 1000s, in the UK over six years. Use these data to make a time series plot. The data are also available from the module Web page.