

## Chapter 2

# Graphical methods for presenting data

### 2.1 Introduction

We have looked at ways of collecting data and then collating them into tables. Frequency tables are useful methods of presenting data; they do, however, have their limitations. With large amounts of data graphical presentation methods are often clearer to understand. Here, we look at methods for presenting graphical representations of data of the types we have seen previously.

### 2.2 Stem and Leaf plots

**Stem and leaf plots** are a quick and easy way of representing data graphically. They can be used with both discrete and continuous data. The method for creating a stem and leaf plot is similar to that for creating a grouped frequency table. The first stage, as with grouped frequency tables, is to decide on a reasonable number of intervals which span the range of data. The interval widths for a stem and leaf plot must be equal. Because of the way the plot works it is best to use “sensible” values for the interval width – i.e. 5, 10, 100, 1000; if a dataset consists of many small values, this interval width could also be 1, or even 0.1 or 0.01. Once we have decided on our intervals we can construct the stem and leaf plot. This is perhaps best described by demonstration.

Consider the following data: 11, 12, 9, 15, 21, 25, 19, 8. The first step is to decide on interval widths – one obvious choice would be to go up in 10s. This would give a **stem unit** of 10 and a **leaf unit** of 1. The stem and leaf plot is constructed as below.

0		8	9		
1		1	2	5	9
2		1	5		
<b>Stem   Leaf</b>					
$n = 8,$ stem unit = 10,   leaf unit = 1.					

You can clearly see where the data have been put. The stem units are to the left of the vertical line, while the leaves are to the right. So, for example, our first observation – 11 – is made up of a stem unit of one 10 and a leaf unit one 1. It is important to give an equal amount of space to each leaf value – by doing so, we can get a clear picture of any patterns in the data (it’s almost like a bar-chart on its side – but it still shows the “raw” observations!). Before producing a stem and leaf plot, it will probably help to first write down the data in ascending numerical order.

### Example 1: Percentage returns on a share

As you might imagine, the interval width does not have to be 10. The following numbers show the percentage returns on an ordinary share for 23 consecutive months:

-0.2   -2.1   1.0   0.1   -0.5   2.4   -2.3   1.5   1.2   -0.6   2.4   -1.2  
1.7   -1.3   -1.2   0.9   0.5   0.1   -0.1   0.3   -0.4   0.5   0.9

Here, the largest value is 2.4 and the smallest -2.3, and we have lots of decimal values in between. Thus, it seems sensible here to have a stem unit of 1 and a leaf unit of 0.1. A stem and leaf diagram for this set of returns then might look like:

-2		1	3						
-1		2	3	2					
-0		5	6	1	4				
0		2	1	9	5	1	3	5	9
1		0	5	2	7				
2		4	4						

**Stem   Leaf**

$n = 23$ ,   stem unit = 1,   leaf unit = 0.1.

### Example 2: Unemployment rates in the U.S.

Hopefully, that should all seem fine so far. So what can go wrong? Consider the following data, which are the percentage unemployment rates for 10 U.S. states:

17   18   15   14   12   19   20   21   24   15

If you were to choose 10 as the interval width (i.e. go up in 10s), the stem and leaf plot would look like

1		2	4	5	5	7	8	9
2		0	1	4				

**Stem   Leaf**

$n = 10$ ,   stem unit = 10,   leaf unit = 1.

Here, the interval width is too large, resulting in only two intervals for our data. With such few intervals it is difficult to identify any patterns in the data. We can get a better idea about what is going on if we choose a smaller interval width – say 5. Doing so gives the following stem and leaf plot:

1		2	4						
1		5	5	7	8	9			
2		0	1	4					

**Stem   Leaf**

$n = 10$ ,   stem unit = 10,   leaf unit = 1.

Notice now that there are two 1s in the stem – one for observations between 10 and 14 (inclusive) and another for observations between 15 and 19 (inclusive). Thus, the stem unit is still 10, but the interval width is now only 5. Changing the interval width like this produces a plot which starts to show some sort of pattern in the data – indeed, this is the intention of such graphical presentations. We could, however, go to the other extreme and have *too many* intervals. If this were the case, any pattern would again be lost because lots of intervals would contain no observations at all. So choose your interval width carefully!

### Example 3: Call centre data

Let us work through the following example. The observations in the table below are the recorded time it takes to get through to an operator at a telephone call centre (in seconds).

54	56	50	67	55	38	49	45	39	50
45	51	47	53	29	42	44	61	51	50
30	39	65	54	44	54	72	65	58	62

**Stem   Leaf**

$n =$

stem unit =

leaf unit =

### Example 4: Production line data

If there is more than one significant figure in the data, the extra digits are *cut*, not *rounded*, to the nearest value; that is to say, 2.97 would become 2.9, not 3.0. To illustrate this, consider the following data on lengths of items on a production line (in cm):

2.97 3.81 2.54 2.01 3.49 3.09 1.99 2.64 2.31 2.22

The stem and leaf plot for this is as follows:

```

1 | 9
2 | 0 2 3
2 | 5 6 9
3 | 0 4
3 | 8

```

$n = 10$ , stem unit = 1 cm, leaf unit = 0.1 cm.

Here the interval width is 0.5. This allows for greater clarity in the plot. Why do you think we *cut* the extra digits?

### 2.2.1 Using Minitab

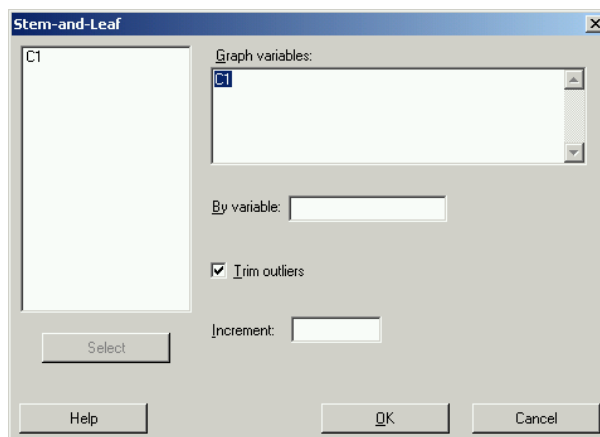
With the small data sets we have seen so far, it is obviously relatively easy to create stem and leaf plots by hand. With larger data sets this would be more problematic and certainly more time consuming. Fortunately, there are computer packages that will create these plots for us – Minitab is one such package, and can be found on all university PCs. Minitab is run by clicking on

Start > All Programs > Statistical Software > Minitab 14 > Minitab 14

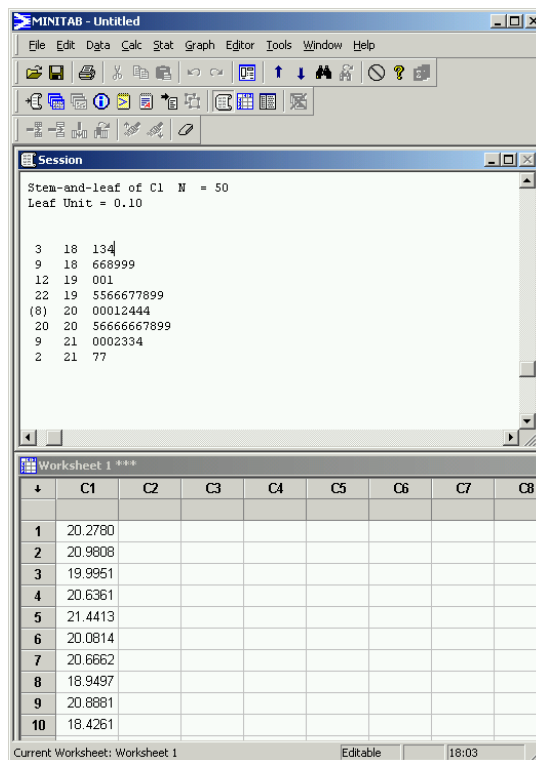
You will see two windows: a session window and a worksheet. Data are entered into columns labelled C1, C2, C3, etc in the worksheet. Suppose C1 contains some data. To obtain a stem and leaf plot of these data you would need to do the following:

Graph > Stem-and-Leaf...

This brings up the window below. You need to type in C1 under Variable and click OK. If you want you can choose the stem unit by entering a value in Increment first, otherwise the programme selects this for you.



This creates a stem and leaf plot in the session window:



It is easy to see some of the advantages of graphically presenting data. For example, here you can clearly see that the data are centred around a value in the low 20's and fall away on either side. From stem and leaf plots we can quickly and easily tell if the data are symmetric or asymmetric. We can see whether there are any **outliers**, that is, observations which are either much larger or much smaller than is typical of the data. We could perhaps even tell whether the data are **multi-modal**, that is to say, whether there are two or more peaks on the graph with a gap between them. If so, this could suggest that the sample contains data from two or more groups.

## 2.3 Bar Charts

**Bar charts** are a commonly-used and clear way of presenting categorical data or any ungrouped discrete frequency observations. As with stem and leaf plots, various computer packages allow you to produce these with relative ease. First let us work through the process of producing these by hand. This will enable you to get a clear idea of how these charts are constructed.

Constructing a bar chart is a 5 step process:

1. First decide what goes on each axis of the chart. By convention the variable being measured goes on the horizontal ( $x$ -axis) and the frequency goes on the vertical ( $y$ -axis).
2. Next decide on a numeric scale for the frequency axis. This axis represents the frequency in each category by its height. It must start at zero and include the largest frequency. It is common to extend the axis slightly above the largest value so you are not drawing to the edge of the graph.
3. Having decided on a range for the frequency axis we need to decide on a suitable number scale to label this axis. This should have sensible values, for example,  $0, 1, 2, \dots$ , or  $0, 10, 20, \dots$ , or other such values as make sense given the data.
4. Draw the axes and label them appropriately.
5. Draw a bar for each category. When drawing the bars it is essential to ensure the following:
  - the width of each bar is the same;
  - the bars are separated from each other by equally sized gaps.

Recall the example on students' modes of transport:

Student	Mode	Student	Mode	Student	Mode
1	Car	11	Walk	21	Walk
2	Walk	12	Walk	22	Metro
3	Car	13	Metro	23	Car
4	Walk	14	Bus	24	Car
5	Bus	15	Train	25	Car
6	Metro	16	Bike	26	Bus
7	Car	17	Bus	27	Car
8	Bike	18	Bike	28	Walk
9	Walk	19	Bike	29	Car
10	Car	20	Metro	30	Car

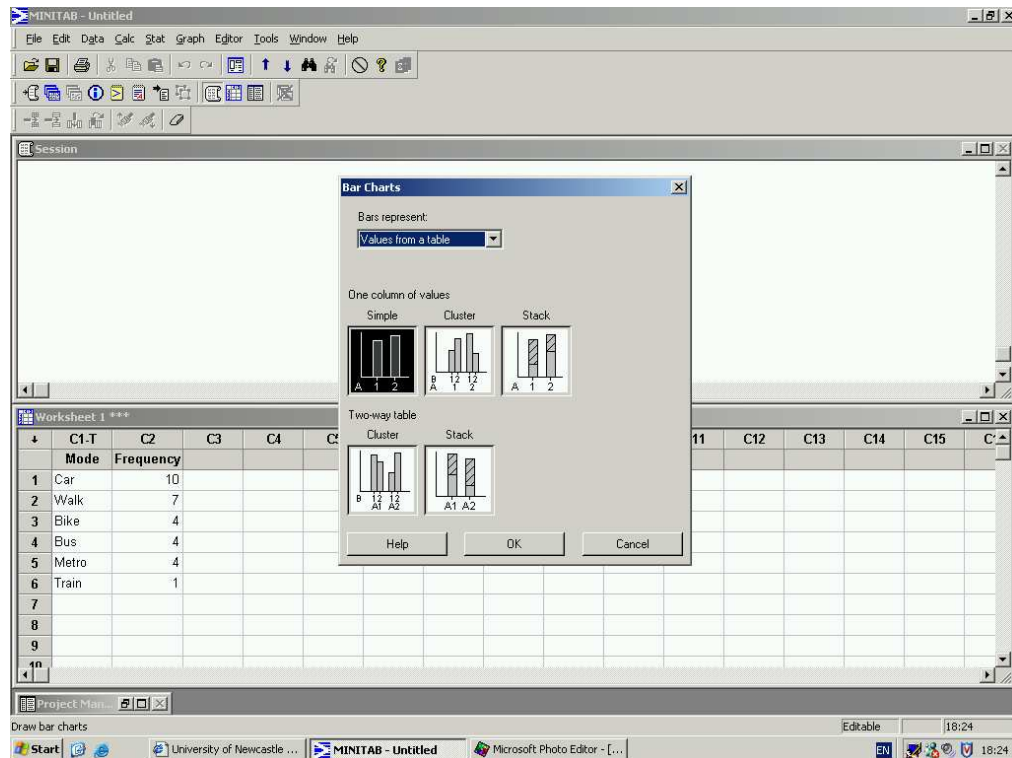
The first logical step is to put these into a frequency table, giving

Mode	Frequency
Car	10
Walk	7
Bike	4
Bus	4
Metro	4
Train	1
<b>Total</b>	30

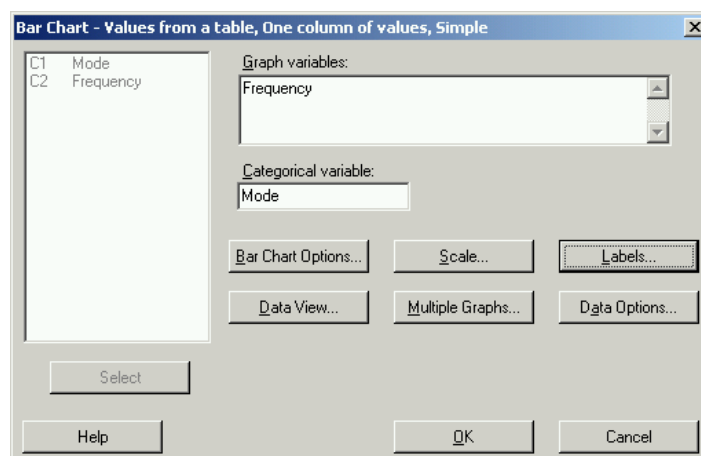
We can then present this information as a bar chart, following the five steps on the previous page:

Such graphs are easily drawn using Minitab:

1. First enter the data in the worksheet, either in summary format or as raw data, with column C1 containing the categories and the (raw or frequency) counts in column C2.
2. Graph > Bar Chart...



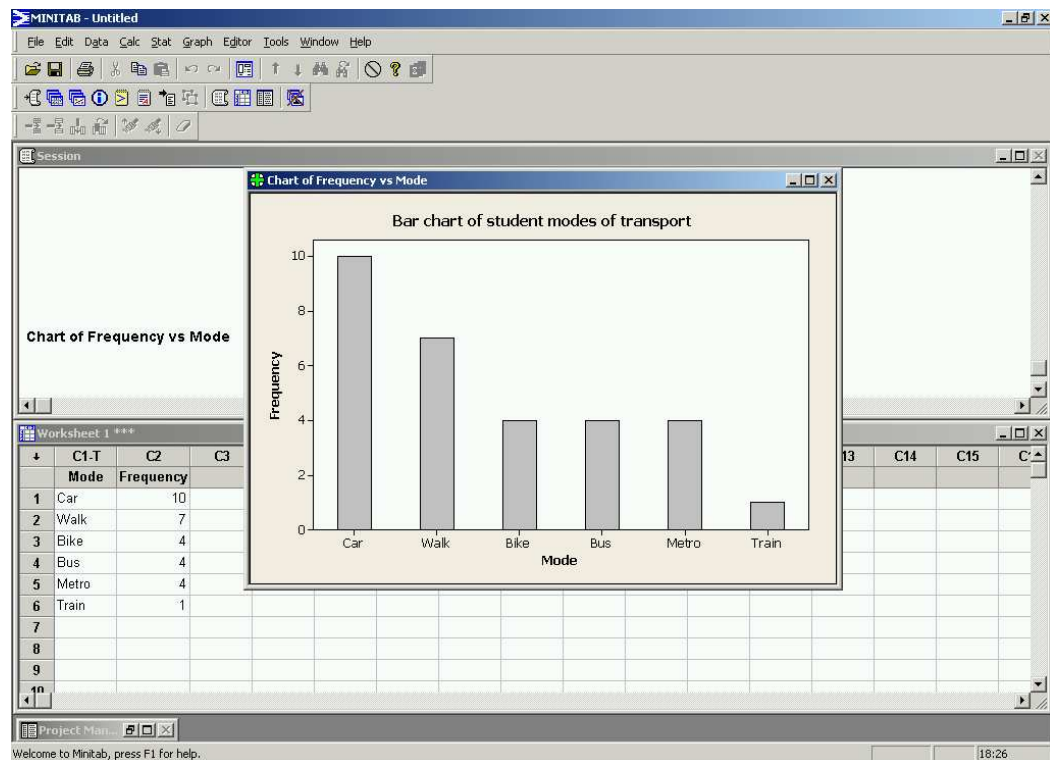
3. Select the appropriate data format (raw data or tabulated data), the columns containing the data, and the graph format.



4. When ready click OK.



This procedure produces the chart



This bar chart clearly shows that the most popular mode of transport is the car and that the metro, walking and cycling are all equally popular (in our small sample). Bar charts provide a simple method of quickly spotting simple patterns of popularity within a discrete data set.

## 2.4 Multiple Bar Charts

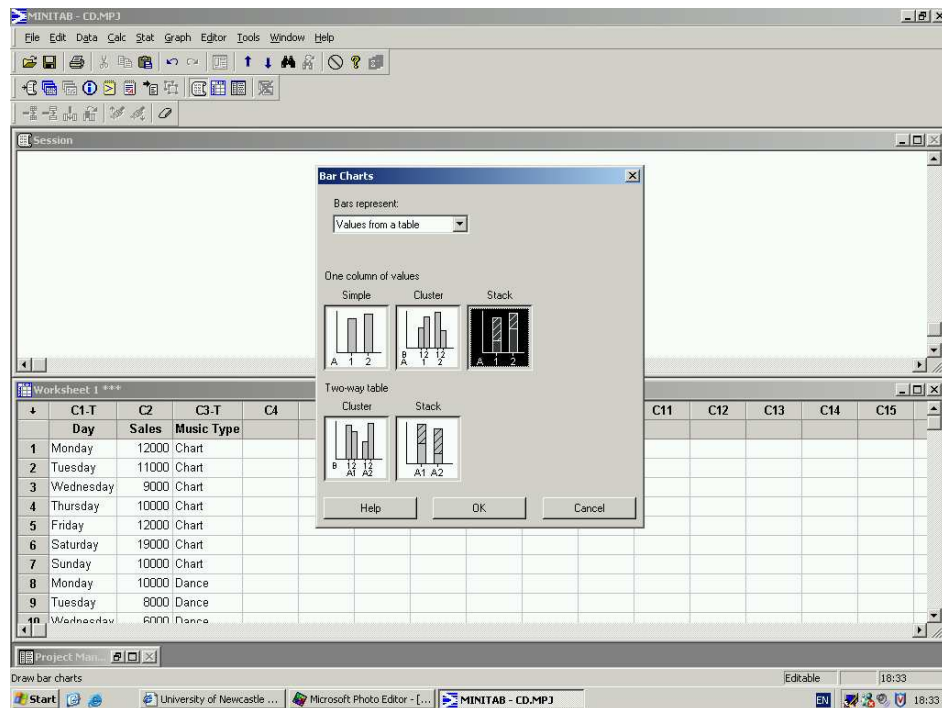
The data below gives the daily sales of CDs (in £) by music type for an independent retailer.

Day	Chart	Dance	Rest	Total
Monday	12000	10000	2700	24700
Tuesday	11000	8000	3000	22000
Wednesday	9000	6000	2000	17000
Thursday	10000	5000	2500	17500
Friday	12000	11000	3000	26000
Saturday	19000	12000	4000	35000
Sunday	10000	8000	2000	20000
<b>Total</b>	83000	60000	19200	162200

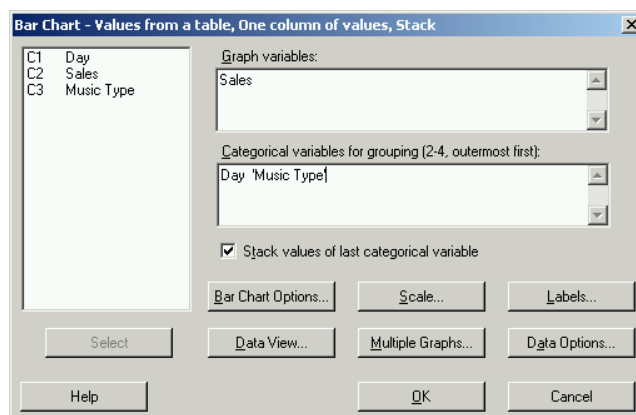
Bar charts could be drawn of total sales per music type in the week, or of total daily sales. It might be interesting to see daily sales broken down into music types. This can be done in a similar manner to the bar charts produced previously. The only difference is

that the height of the bars is dictated by the total daily sales, and each bar has segments representing each music type. This is done in Minitab as follows:

1. Enter the data into the worksheet, the types of music in columns and the days as rows.
2. Graph > Bar Chart...
3. Select the appropriate data format and the Stack graph format.

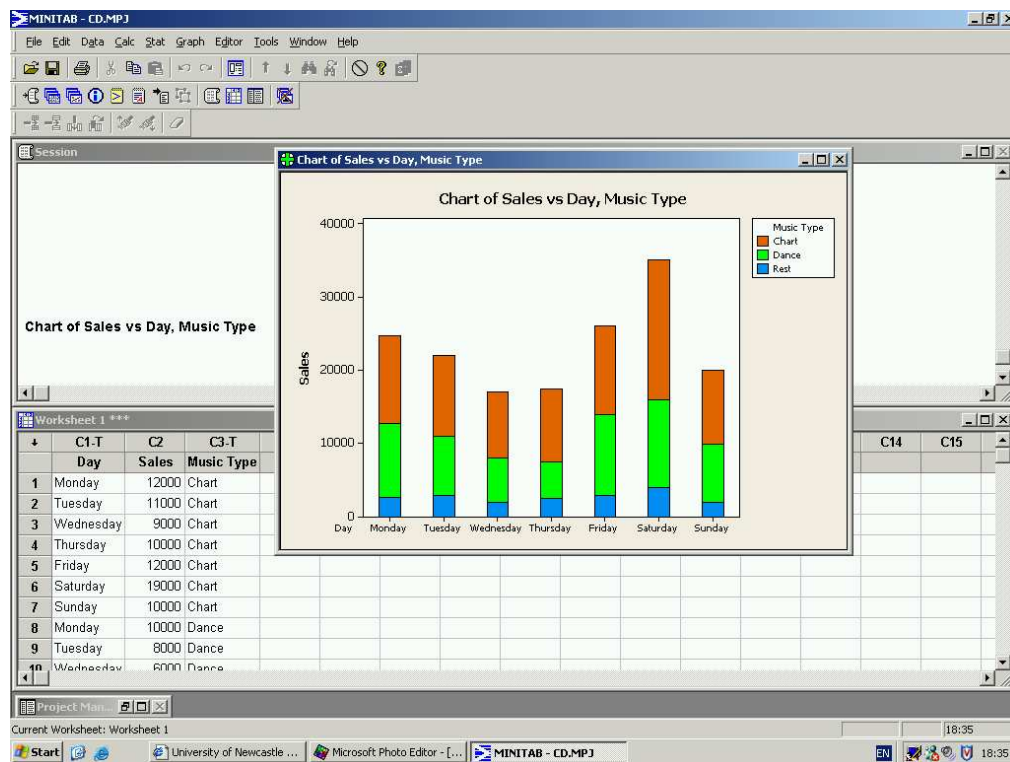


4. Click OK.
5. Enter the column containing the Sales data under Graph variables and the Day and Music Type in the grouping dialogue box.



6. Click OK.

The Minitab worksheet and chart this produces are shown below:



These types of charts are particularly good for presenting such financial information or illustrating any breakdown of data over time – for example, the number of new cars sold by month and model.

## 2.5 Histograms

Bar charts have their limitations; for example, they cannot be used to present continuous data. When dealing with continuous random variables a different kind of graph is required. This is called a **histogram**. At first sight these look similar to bar charts. There are, however, two critical differences:

- the horizontal ( $x$ -axis) is a continuous scale. As a result of this there are *no gaps between the bars* (unless there are no observations within a class interval);
- the height of the rectangle is only proportional to the frequency if the class intervals are all equal. With histograms it is the *area* of the rectangle that is proportional to their frequency.

Initially we will only consider histograms with equal class intervals. Those with uneven class intervals require more careful thought.

Producing a histogram is much like producing a bar chart and in many respects can be considered to be the next stage after producing a grouped frequency table. In reality,

it is often best to produce a frequency table first which collects all the data together in an ordered format. Once we have the frequency table, the process is very similar to drawing a bar chart.

1. Find the maximum frequency and draw the vertical ( $y$ -axis) from zero to this value, including a sensible numeric scale.
2. The range of the horizontal ( $x$ -axis) needs to include not only the full range of observations but also the full range of the class intervals from the frequency table.
3. Draw a bar for each group in your frequency table. These should be the same width and touch each other (unless there are no data in one particular class).

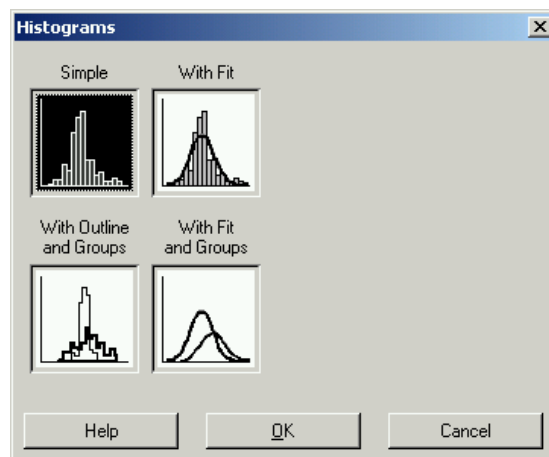
The frequency table for the data on service times for a telephone call centre (Section 1.3.2) was

Service time	Frequency
$175 \leq \text{time} < 180$	1
$180 \leq \text{time} < 185$	3
$185 \leq \text{time} < 190$	3
$190 \leq \text{time} < 195$	6
$195 \leq \text{time} < 200$	10
$200 \leq \text{time} < 205$	12
$205 \leq \text{time} < 210$	8
$210 \leq \text{time} < 215$	3
$215 \leq \text{time} < 220$	3
$220 \leq \text{time} < 225$	1
<b>Total</b>	50

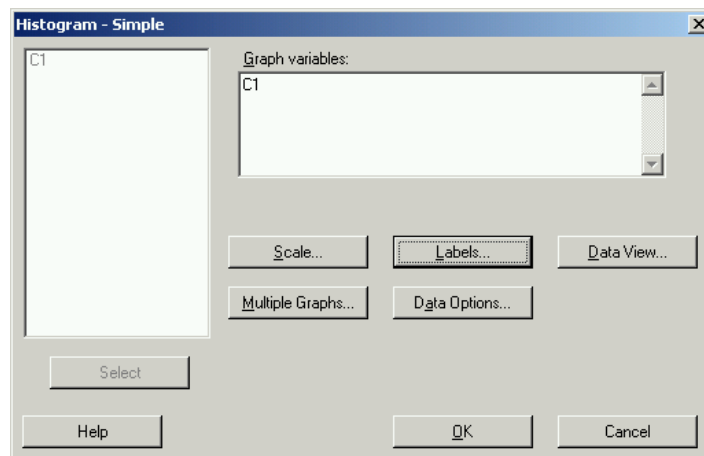
The histogram for these data is:

Normally, as with stem and leaf plots and bar charts, we would get Minitab to do this for us.

1. Enter the data in column C1 of the worksheet. For illustrative purposes I have randomly generated 500 observations in this column.
2. Graph > Histogram...
3. Select the Simple graph format.



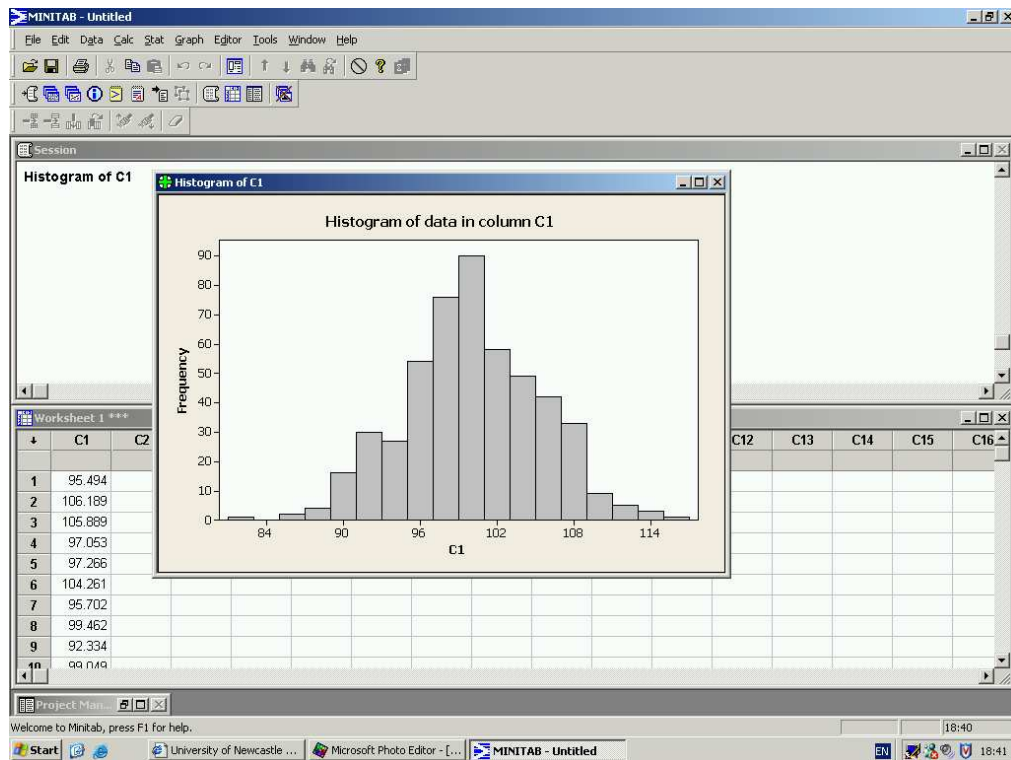
4. Select C1 under Graph variables.



Note: various advanced options are available e.g. a title can be added by clicking Labels

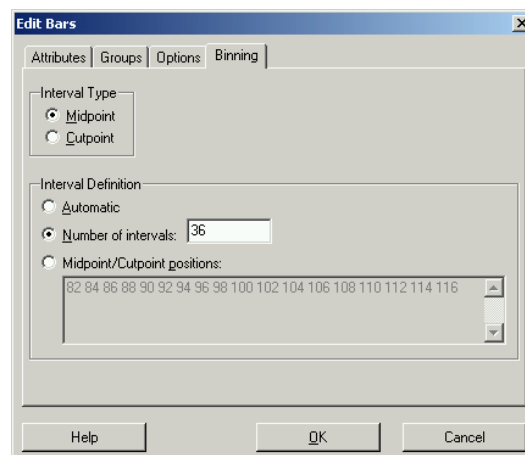
5. When happy with your choices click OK.

These instructions produce the following histogram:

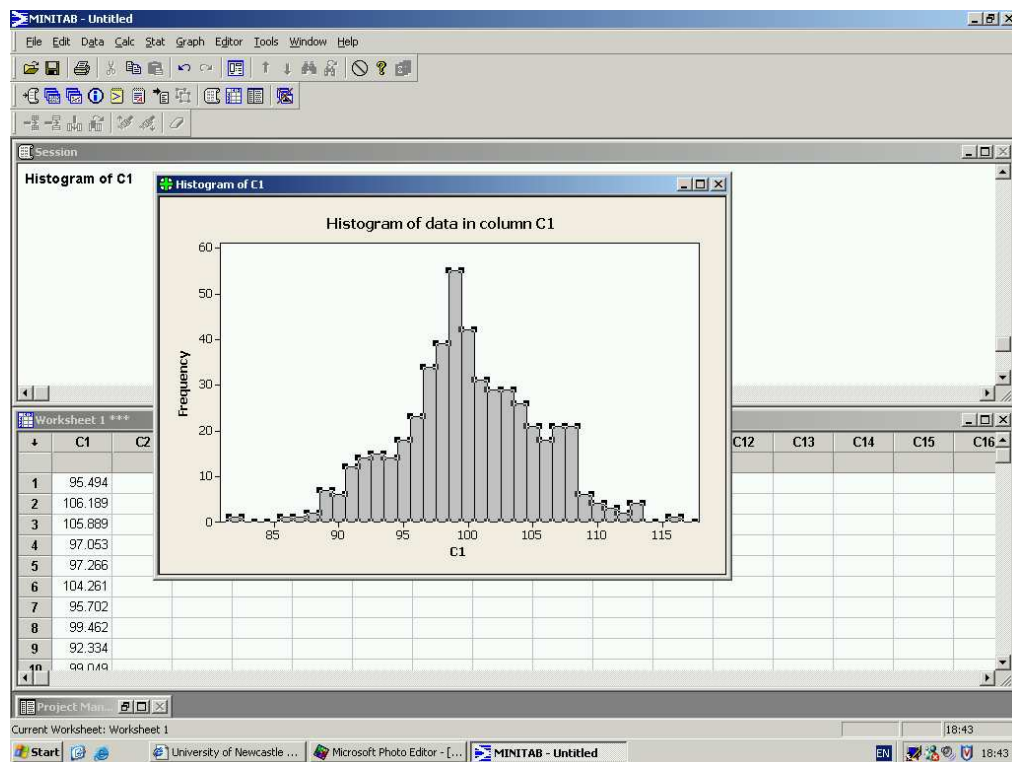


The histogram produced can be amended by **right-clicking** on the graph. For example, the intervals used in the histogram can be changed or, more simply, the number of intervals using **Edit bars > Binning**.

We can double the number of intervals (from 18 to 36 intervals) using the **Binning** dialogue box

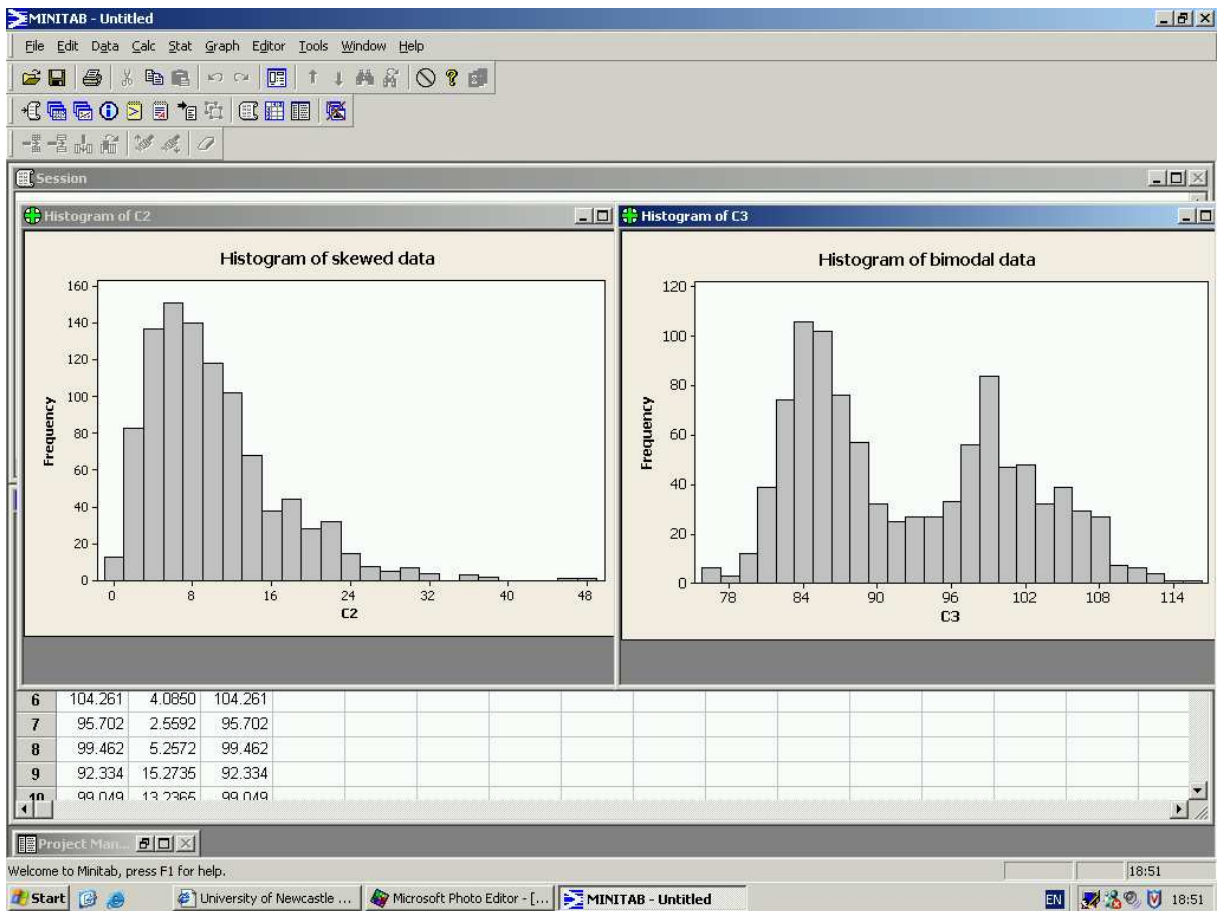


This changes the histogram to



Histograms are useful tools in data analysis. They are easy to produce in Minitab for large data sets and provide a clear visual representation of the data. Using histograms, it is easy to spot the **modal** or most popular class in the data, i.e. the one with the highest peak. It is also easy to spot simple patterns in the data. Is the frequency distribution symmetric, as the histograms produced above, or is it skewed to one side like the left-hand histogram in the following graphic?

Histograms also allow us to make early judgements as to whether all our data come from the same population. Consider the right-hand histogram in the graphic below. It clearly contains two separate modes (peaks), each of which has its own symmetric pattern of data. This clearly suggests that the data come from two separate populations, one centred around 85 with a narrow spread and one centred around 100 with a wider spread. In real situations it is unlikely that the difference would be as dramatic, unless you had a poor sampling method. However, the drawing of histograms is often the first stage of a more complex analysis.



Finally, when drawing histograms be aware of observations on variables which have boundaries on their ranges. For example heights, weights, times to complete tasks etc. can not take negative values so there is a lower limit at zero. Computer programs do not automatically know this. You should make sure that the lower limit of the first class interval is not negative in such cases.



## 2.6 Exercises

1. The following table shows the weight (in kilograms) of 50 sacks of potatoes leaving a farm shop.

10.41	10.06	9.38	11.36	9.65
11.24	10.58	8.55	10.47	8.22
9.36	9.63	10.33	10.05	11.57
11.36	10.82	8.93	10.08	9.53
10.05	11.30	11.01	9.72	10.67
9.91	10.26	10.67	10.21	8.18
8.70	9.49	10.98	10.01	9.92
9.27	11.69	9.66	9.52	10.40
10.61	8.83	10.11	10.37	9.73
10.72	10.63	12.86	10.62	10.26

Display these data in a stem and leaf plot. Note the number of decimal places and adjust accordingly. State clearly both the stem and leaf units.

2. A market researcher asked 650 students what their favourite daily newspaper was. The results are summarised in the frequency table below. Represent these data in an appropriate graphical manner.

The Times	140
The Sun	200
The Sport	50
The Guardian	120
The Financial Times	20
The Mirror	80
The Daily Mail	10
The Independent	30

3. Produce a histogram for the data on length of mobile phone calls from the exercises in Chapter 1 (listed again below) and comment on it.

281.4837	293.4027	306.5106	286.6464	298.4445
312.7291	327.7353	311.5926	314.8501	303.3484
270.7399	293.9364	310.9137	346.4497	304.6044
304.1124	320.7182	283.6594	337.5806	259.6408
305.4378	317.9180	289.5667	286.9626	300.5140
278.3108	300.1725	292.6725	312.9645	302.5770
293.2735	267.5344	326.9056	257.7226	285.9805
299.6535	293.9145	303.9191	323.7993	263.5242
281.1613	306.9344	310.2583	301.6963	313.9611
314.8500	292.0031	302.4314	267.9781	292.0917