Using Diagrams to Present Data[1]

Raw data often give so much detail that it is impossible to see the overall patterns. Data reduction clears away the detail and highlights the underlying patterns: it presents summarized results which are concise, but still give an accurate view of the original data. The reduction can be done in several ways, and in this chapter we describe alternative types of diagram.

After reading this chapter and doing the exercises you should be able to:

- outline the purpose of data reduction
- design tables of numerical data
- draw graphs to show the relationship between variables
- design pie charts
- draw a variety of bar charts
- draw pictograms and recognize their **limitations**
- use frequency distributions and tables
- draw histograms
- draw ogives and Lorenz curves for cumulative data

4.1 Summarizing data

4.1.1 Introduction

This chapter is based on the principle that there is a difference between data and information. Data are the raw numbers or facts which must be processed to give useful information. Thus 78, 64, 36, 70 and 52 are data which could be processed to give the information that the average mark of five students sitting an exam is 60%.

Imagine that you have spent a lot of effort collecting data and now want to communicate your findings to other people. This is done by **data presentation.** The purpose of data presentation is to show the characteristics of a set of data and highlight any important patterns. This can either be done numerically, or by using diagrams. The remainder of this chapter describes methods of presentation based on diagrams, while the next chapter discusses numerical presentations.

If you look around, there are countless examples of information presented in diagrams. Typically, newspaper articles describe a situation, and add summary diagrams to accompany the text. People find these diagrams attractive and are more likely to look at them than read the article (hence the saying, 'One picture is worth a thousand words'). The main benefit of diagrams is that people are good at recognizing patterns and can extract a lot of information in a short time.

In general, then, the success of a presentation can be judged by how easy it is to understand. A good presentation should make information clearer and allow us to see the overall picture, which would be missed if data were presented in any other form. Unfortunately, good presentations do not happen by chance but need careful planning. If you look at a diagram and cannot understand it, it is safe to assume that the presentation is poor; the fault is with the presenter rather than the viewer.

Sometimes, even when a presentation appears clear, closer examination may show that it does not give a true picture of the data. This may be a result of poor presentation, but sometimes results from a deliberate decision to present data in a form that is both misleading and dishonest. Advertisements are notorious for presenting data in a way that gives the desired impression, rather than accurately reflecting a situation. Likewise, politicians may be concerned with appearance rather than truth. The problem is that diagrams are a powerful means of presenting data, but they only give a summary. This summary can easily be misleading, either intentionally or by mistake. In this chapter we shall demonstrate good practice in data presentation and shall be rigorous in presenting results that are fair and honest.

IN SUMMARY

The aim of data presentation is to give an accurate summary of data. Here we concentrate on diagrammatic presentations. These have considerable impact, but need careful planning.

4.1.2 Data reduction

Provided they come in small quantities, most people can deal with numerical data. We can happily say, 'This building is 60 metres tall', 'A car can travel 40 miles on a gallon of petrol', 'An opinion poll shows one political party has 6% more support than another', and so on. Problems begin when there are a lot of data and we are swamped with detail. Suppose, for example, we know that weekly sales of a product in a shop over the past year are:

51 60 58 56 62 69 58 76 80 82 68 90 72

84 91 82 78 76 75 66 57 78 65 50 61 54

49 44 4145 38 28 37 40 42 22 25 26 21

30 32 30 32 3129 30 4145 44 47 53 54

If these data were given in a report, people would find it, at best, boring and would skip to more interesting material. They would ignore the figures, despite the fact that they could be important. To make the figures less daunting we could try including them in the text, but when there are a lot of numerical data this does not work. The figures above could only be described in the text of a report by saying, 'In the first week sales were 51 units, and then they rose by nine units in the second week, but in the third week they fell back to 58 units, and fell another two units in the fourth week...'. We need a more convenient way of presenting data.

The problem is that the raw data do not really tell us very much; we are simply swamped with detail and cannot see the wood for the trees. In most cases we are not interested in the small detail, but really want the overall picture. What we need, then, is a way of identifying general patterns in data and presenting a summary which allows these to be seen. This is the purpose of **data reduction.**

> The aim of data reduction is to give a simplified and accurate view of the data which shows the underlying patterns but does not overwhelm us with detail.

Thus the sequence of activities concerned with analysing data starts with data collection, then moves to data reduction, and finally to data presentation.

In practice, the distinction between data reduction and data presentation is not clear, and they are usually combined into a single activity.

Data reduction has a number of clear advantages:

- results are shown in a compact form

- results are easy to understand
- graphical or pictorial representations can be used
- overall patterns can be seen
- comparisons can be made between different sets of data
- quantitative measures can be used

Conversely, it has the disadvantages that:

- details of the original data are lost
- the process is irreversible

We mentioned in the last chapter that we are discussing the presentation of data after discussing their collection. We should say again that this is often the way things are organized in practice, but the way data will be presented should have an effect on the way they are collected. If, for example, results of a survey are to be shown as a graph, appropriate data could not be collected by asking an open-ended question like, 'Please give your comments on...'. If we want to present a summary of a company's financial position we need not collect data about every transaction that it made in the past few years.

It should also be clear that if you have a large quantity of data, processing them will always be done on a computer. However, the computer only manipulates the data and it plays no part in making decisions about the best analyses or how to present results. These decisions must be made by the person presenting the results.

IN SUMMARY

The detail given in raw data can be overwhelming and can obscure overall patterns. Data reduction simplifies the data and presents them so that underlying patterns can be seen.

Self-assessment questions

4.1 What is the difference between data and information?

4.2 Give five examples of misleading data presentation.

4.3 Why is data reduction necessary?

4.4 ‘Data reduction always gives a clear, detailed and accurate picture of the initial data.’ Is this statement true?

4.2 Diagrams for presenting data

4.2.1 Introduction

The purpose of data presentation is to summarize data and present them in a form which is more precise, but still gives an accurate view of the raw data. There are several ways in which data can be summarized in diagrams, and we shall classify the most important of these as:

- tables of numerical data
- graphs to show relationships between variables
- pie charts, bar charts and pictograms showing relative frequencies
- histograms which show relative frequencies of continuous data

The choice of best format is essentially a matter of personal judgement. There are, however, some guidelines that are largely common sense and include, where appropriate:

- select the most suitable format for the purpose
- present data fairly and honestly
- make sure any diagram is clear and easy to understand
- give each diagram a title
- state the source of data
- use consistent units and say what these units are
- label axes clearly and accurately
- put a clear scale on axes
- include totals, subtotals and any other useful summaries
- add notes to highlight reasons for unusual or atypical values.

It is worth mentioning that drawing diagrams for data presentation used to be quite time-consuming, but business graphics packages (such as Harvard Graphics, DrawPerfect, Corel Draw, Aldus FreeHand, Dr Halo and a whole range of equivalent packages) have made this task much simpler.

IN SUMMARY

Data can be presented in several ways, but the final choice is often a matter of opinion. Some guidelines for good practice can be given.

4.2.2 Tables

The easiest way of presenting numerical data is in a table. This is perhaps the most widely used method of data presentation (and has already been used several times in this book). Whenever you pick up a newspaper, magazine or report you are likely to see a number of tables. This is one of the easiest, and most effective, ways of presenting a lot of information, and spreadsheet packages make the design and manipulation of tables very easy.

The general features of a table can be seen in Table 4.1, which is a presentation of the data for sales given above.

Table 4.1

Week	*Quarter 1*	*Quarter 2*	*Quarter 3*	*Quarter 4*	*Total*
1	51	84	49	30	214
2	60	91	44	32	227
3	58	82	41	30	211
4	56	78	45	32	211
5	62	76	38	31	207
6	69	75	28	29	201
7	58	66	37	30	191
8	76	57	40	41	214
9	80	78	42	45	245
10	82	65	22	44	213
11	68	50	25	47	190
12	90	61	26	53	230
13	72	54	21	54	201
Totals	882	917	458	498	2 755

This gives some idea of the overall patterns so we can see, for example, that demand is higher in the first two quarters and lower in the second two. In this format, though, the table is still really a presentation of the raw data and it is difficult to get a feel for a typical week's sales; there is no indication of minimum or maximum sales; and so on. These defects would be even more noticeable if there were hundreds or thousands of observations. It would be useful to reduce the data and emphasize the patterns. The minimum sales are 21, so we might start by seeing how many weeks had sales in a range of, say, 20 to 29. If we count these, there are six weeks. Then we could count the number of observations in other ranges, as follows:

Range of sales	*Number of weeks*
20 to 29	6
30 to 39	8
40 to 49	10
50 to 59	9
60 to 69	7
70 to 79	6
80 to 89	4
90 to 99	2

This table shows how many values are in each range, and is called a **frequency table** (we shall return to these later in the chapter). The 'ranges' are usually referred to as **classes.** Then we can talk about the 'class of 20 to 29', where 20 is the lower class limit and 29 is the upper class limit and the class width is 29 - 20 = 9. We arbitrarily chose classes of 20 to 29, 30 to 39, and so on, but could have used any appropriate classes. It might be useful, for example, to choose the classes 17 to 32, 33 to 48, or any other convenient ones. The only constraint is that there should be enough classes to make any patterns clear, but not so many that they are obscured. If we felt that the eight classes used above were too many, we could redefine the classes to, say, the four shown in the following table. This table has also been given a title and a statement about the source of data.

Table 4.2 Weekly sales of product

Range	*Number of weeks*
20 to 39	14
40 to 59	19
60 to 79	13
80 to 99	6

Source: Company weekly sales reports

These tables show one inevitable effect of data reduction: the more data are summarized, the more detail is lost. The last table, for example, shows the frequency of sales, but it gives no idea of the seasonal variations. Such loss of detail is acceptable if the table is easier to understand and still shows the required information, but is not acceptable if we need to know more detail.

Drawing tables needs a compromise between making them too long (where lots of details can be seen, but they are complicated with underlying patterns hidden) and too short (where underlying patterns are clear, but most details are lost). The number of classes, in particular, must be a subjective decision based on the use of the presentation, but a guideline would set a maximum number at about ten.

There is an almost limitless number of ways of drawing tables. Sometimes they are very simple, like a review of answers to the survey question, 'Did you read a Sunday newspaper last week?':

	Percentage of replies
Yes	76 %
No	14%
Don't know	10%

Sometimes tables are very complex. They can show a lot of information and may be the only realistic means of presentation. The example in Table 4.3 shows figures for crops grown in the UK during the 1980s.

Table 43 Main cereal crops grown in the United Kingdom

	1975-1977 average	*1984*	*1985*	*1986 forecast*
Wheat				
Area ('000 hectares)	1115(30.6)	1939 (48.2)	1902 (47.5)	1997 (49.8)
Harvest ('000 tonnes)	4 834 (33.3)	14 958 (56.4)	12 050 (53.8)	13 910 (56.9)
Yield (tonnes per hectare)	4.32	7.71	6.33	6.96
Barley				
Area ('000 hectares)	2313 (63.4)	1979 (49.2)	1966 (49.1)	1917 (47.8)
Harvest ('000 tonnes)	8897 (61.3)	11064(41.7)	9740 (43.5)	10 010 (41.0)
Yield (tonnes per hectare)	3.85	5.59	4.95	5.22
Oats				
Area ('000 hectares)	221 (6.1)	106 (2.6)	134 (3.3)	97 (2.4)
Harvest ('000 tonnes)	783 (5.4)	517(1.9)	615 (2.7)	505 (2.1)
Yield (tonnes per hectare)	3.54	4.89	4.59	5.16
Totals				
Area ('000 hectares)	3649	4024	4002	4011
Harvest ('000 tonnes)	14 514	26 539	22 405	24 425

Source: *Annual Review of Agriculture,* HMSO

Notes: Figures in brackets are percentages of annual totals

Rounding may make percentages not add to 100%

Droughts in the summers of 1975 and 1976 had an effect on yields in these years.

In common with most tables there are several ways in which this information could be presented and the format given is only one suggestion. If you are repeatedly presenting data over some period, it is a good idea to keep the same format so that direct comparisons can be made. Useful examples of this are given in government publications, such as *Annual Abstract of Statistics*, *Monthly Digest of Statistics*, *Social Trends* and *Economic Trends* which are published by the Central Statistical Office.

Unfortunately, complex tables need more interpretation and do not easily show patterns. This could be avoided by splitting the table into smaller self-contained tables. An alternative is to use a table of values as the first step in data reduction and then give summaries in some other form. Alternatives for this are described in the following sections.

IN SUMMARY

Tables are a widely used method of presenting numerical data. A well-designed table can show a lot of information and can be tailored to specific needs. A poorly designed table can obscure underlying patterns and lose details of the data.

4.2.3. Graphs

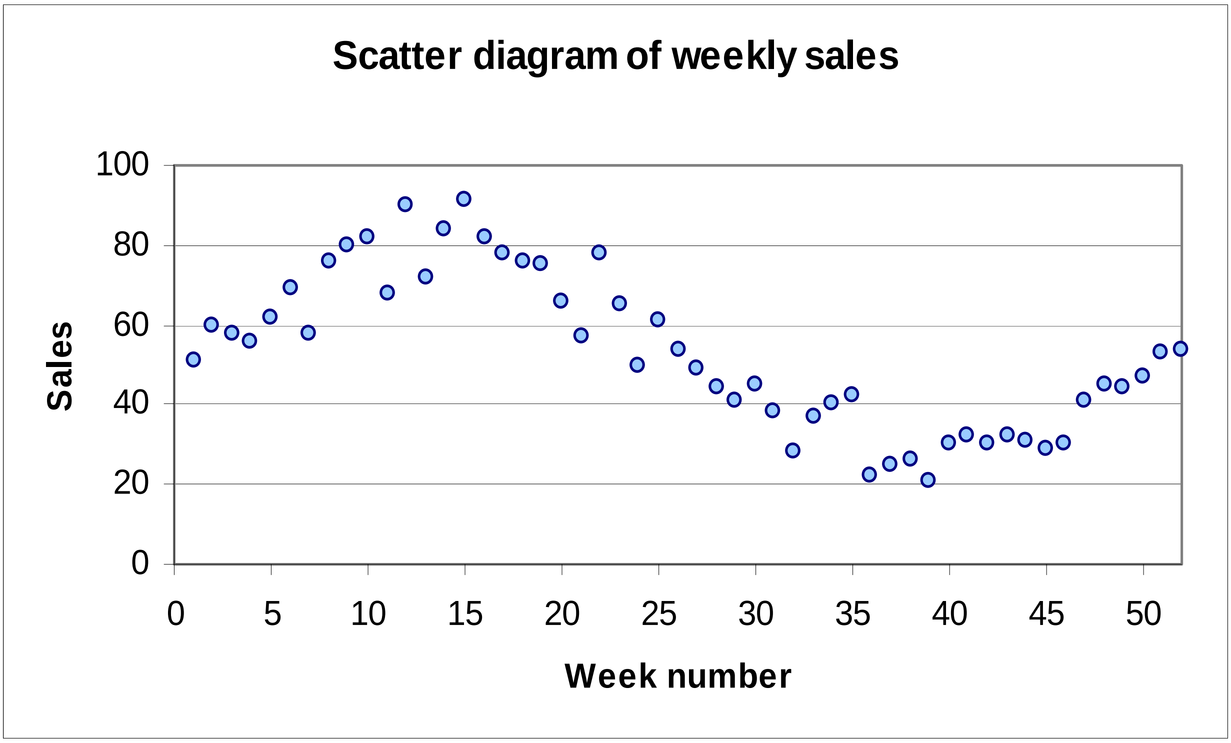
Tables are good at presenting a lot of information, but they do not necessarily highlight underlying patterns. These can be seen more clearly with some form of pictorial representation. Perhaps the most widely used of these are graphs, which we described in Chapter 2.

In essence, a graph shows the relationship between two variables on a pair of rectangular (or Cartesian) axes, where:

- the horizontal or *x* axis shows the variable that is responsible for a change (the independent variable)
- the vertical or *y* axis shows the variable that we are trying to explain (the dependent variable)

In some cases it is not obvious which is the dependent and which the independent variable. If we are plotting sales of ice cream against temperature, then clearly there is an independent variable (temperature) and a dependent variable (sales of ice cream). However, if we are plotting sales of ice cream against sales of sausages, then there is no such clear relationship. Then it is a matter of choice as to which way round to plot the axes.

Figure 4.1 Scatter diagram of weekly sales.



Graphs summarizing a set of raw data can be drawn in a number of ways. Returning to the weekly sales described earlier, we could start by plotting sales (the dependent variable that we are trying to explain) against the week (the independent variable that causes the changes). The simplest graph of this would just show the individual points in a **scatter diagram,** as illustrated in Figure 4.1.

This graph shows the general pattern, but this is made clearer if the points are joined, as shown in Figure 4.2. The sales clearly follow a seasonal cycle with peak sales around week 12 and lowest sales around week 38. There are small random variations away from this overall pattern, so the graph is not a smooth curve. Usually we are more interested in the smooth trend than the random variations, so we should emphasize this. Figure 4.3 shows individual points plotted around the smooth trend line.

The most common difficulty with graphs is the choice of scale for the v axis. We could redraw the graphs in Figures 4.1 - 4.3 with changed scales for the y axis, and the shape of the graph would vary considerably. Figure 4.4 shows a very stable pattern with only small variations from a low constant value. Figure 4.5 shows widely varying values, which are consistently high in the first half, and then almost zero in the second half. These two graphs actually show the same data as Figures 4.1 - 4.3, but with changed scales for the y axis.

As graphs give a very strong initial impact, the choice of scale for the axes is clearly important, with a bad choice giving a false view of the data. Although the choice of scale is largely subjective, some guidelines for good practice can be given:

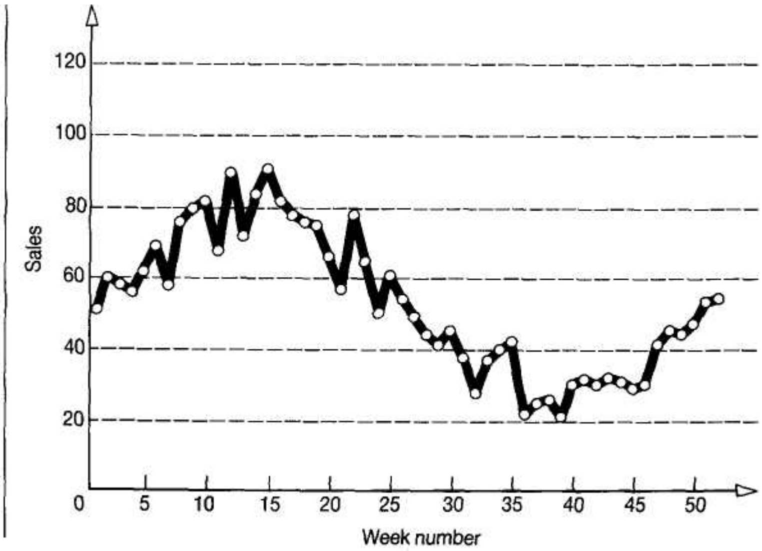


Figure 4.2 Graph of weekly sales.

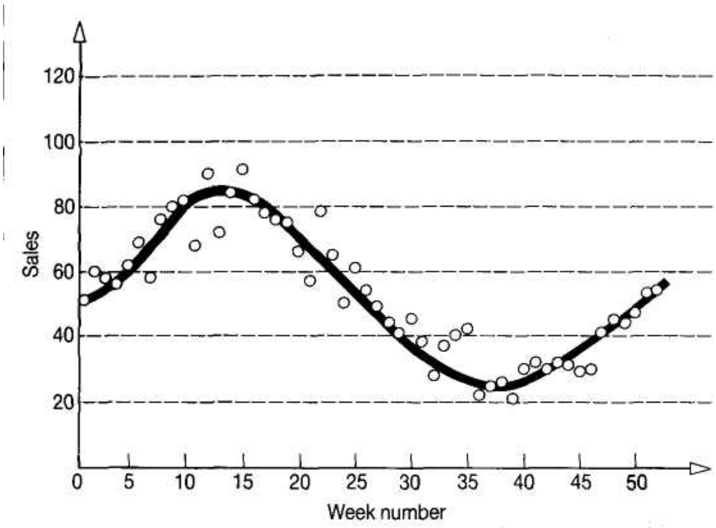


Figure 4.3 Smoothed graph of weekly sales.

- always label the axes clearly and accurately
- show the scales on both axes
- the maximum of the scale should be slightly above the maximum observation
- wherever possible the scale on axes should start at zero: if this cannot be done the scale must be shown clearly, perhaps with a zig-zag on the axis to indicate a break

- where appropriate, give the source of data

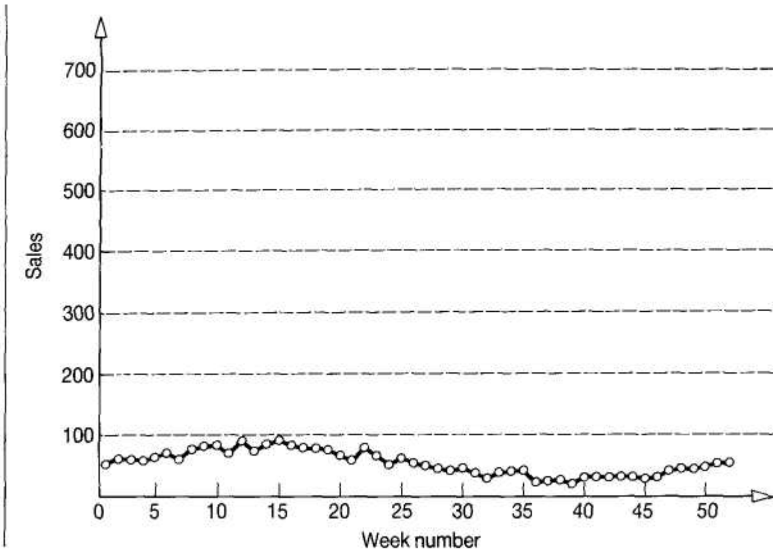


Figure 4.4 Graph of stable weekly sales.

- where appropriate, give the graph a title

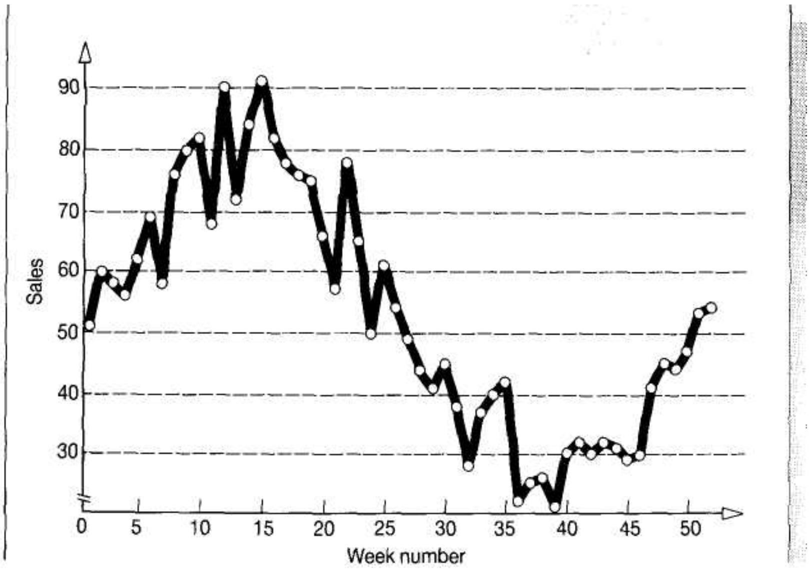


Figure 4.5 Graph of variable weekly sales.

One of the benefits of graphs is their ability to compare data by plotting several graphs on the same axes. Figure 4.6, for example, shows how the unit price of a basic commodity has varied each month over the past five years. Notice that the price axis does not go down to zero. The price differences are small and can be highlighted by using a narrower range for the v axis. This means, of course, that the axis must be clearly labelled.

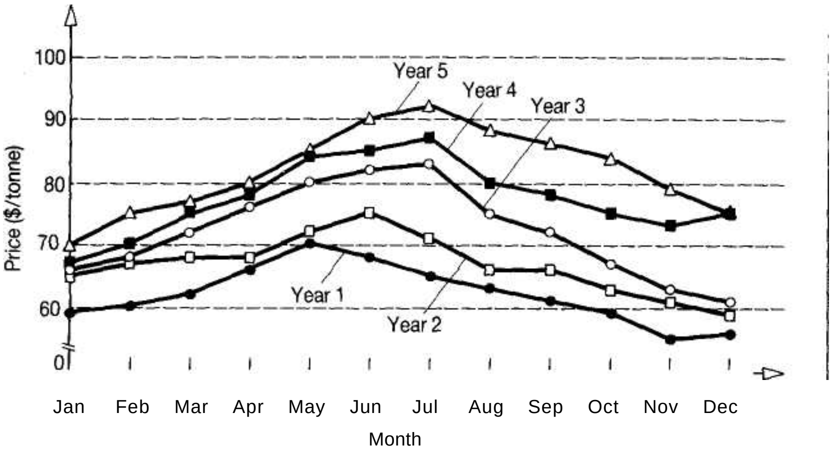


Figure 4.6 Price in $ per tonne of commodity by month. (Source: UN Digest)

WORKED EXAMPLE 4.1

Table 4.3 shows the quarterly profit reported by a company and the corresponding average price of its shares quoted on the London Stock Exchange. Draw a graph of these data.

Table 4.3

Year	1				2				3			
Quarter	1	2	3	4	1	2	3	4	1	2	3	4
Profit	12.1	12.2	11.6	10.8	13.0	13.6	11.9	11.7	14.2	14.5	12.5	13.0
Share price	122	129	89	92	132	135	101	104	154	156	125	136

Source: company reports and the *Financial Times*

Note: profits are in millions of pounds and share prices are in pence

Solution

The independent variable is the one that is responsible for changes; in this example it is the company profit. The dependent variable is the one that we are trying to explain; in this example it is the share price. A graph of these results is shown in Figure 4.7. The chosen scale highlights the linear relationship between profit and share price. As always, information must be carefully examined, and in this case inflation might have a significant effect on the results. If it is important to show the cyclical nature of the data, graphs could also be drawn of profit and share price against quarter.

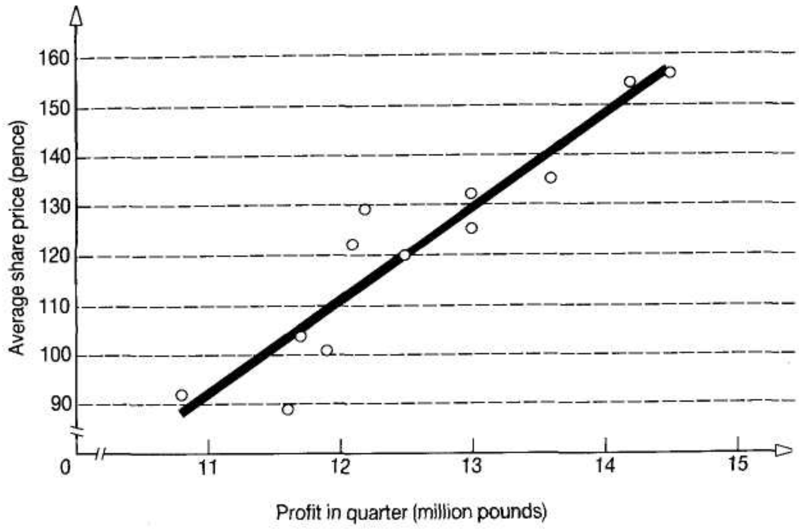


Figure 4.7 Graph of share price against sales for Worked Example 4.1.

IN SUMMARY

Graphs show clear relationships between two variables. Underlying patterns are easily identified and different sets of data can be compared. Care must be taken in choosing appropriate scales for the axes.

4.2.4 Pie charts

Graphs are good at showing relationships between two variables, but other methods of presenting data rely more directly on pictures. Pie charts are simple diagrams that are used for comparisons of limited amounts of information.

To draw a pie chart the data are first classified into distinct categories. Then a circle is drawn (the pie) which is divided into sectors, each of which represents one category. The area of each sector (and hence the angle at the centre of the circle) is proportional to the number of observations in the category.

WORKED EXAMPLE 4.2

Sales in four regions are given in the following table. Draw a pie chart to represent these.

Region	*Sales*
North	25
South	10
East	45
West	25
Total	100

Solution

There are 360° in a circle, and these represent 100 observations. Therefore each observation is represented by an angle of 360/100 = 3.6° at the centre of the circle. Then the sales in the North region are represented by a sector with an angle of 25 x 3.6 = 90° at the centre of the circle; sales in the South region are represented by a sector with an angle of 10 x 3.6 = 36° at the centre, and so on. A basic pie chart for this is shown in Figure 4.8(a). The appearance of pie charts can be improved in several ways, and Figure 4.8(b) shows the same results with slices in the pie sorted into order, percentages calculated and a three-dimensional effect added.

Sometimes two pies can be linked, so Figure 4.8(c) shows the sales and profits from each region, with the results for the West region pulled out for emphasis.

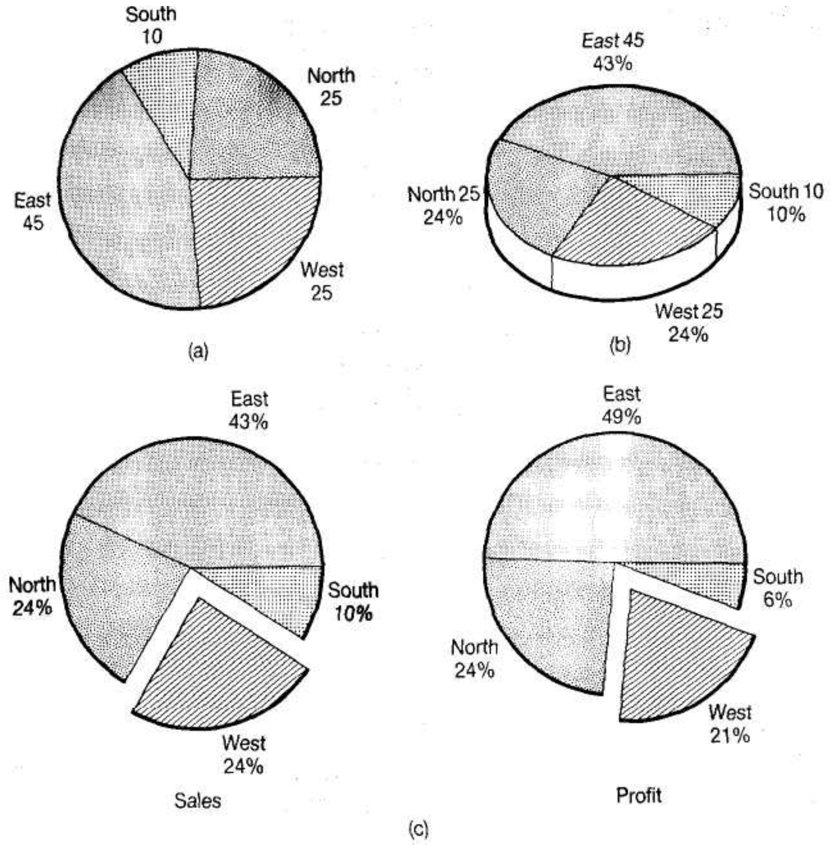


Figure 4.8 (a) Basic pie chart of sales for Worked Example 4.2 (b) Fuller pie chart of sales. (c) Associated pie charts.

Pie charts compare the relative number of observations in different categories They can be used for percentages, but really have little other use. They ai certainly only useful when there are a few categories, say four to eight, a beyond this they become too complicated and lose their impact.

IN SUMMARY

Pie charts represent the relative frequency of observations by the sectors < a circle. They can give considerable impact, but are only useful for sm< quantities of data.

4.1.1. Bar charts

Like pie charts, bar charts are diagrams that show the number of observations in different categories of data. This time, though, the numbers of observations are shown by lines or bars rather than sectors of a circle.

In a bar chart, each category of data is represented by a different bar, and the length of the bar is proportional to the number of observations. Bar charts are usually drawn vertically, but they can be horizontal, and there are many adjustments that enhance their appearance. One constant rule, however, is that the scale must start at zero; any attempt to save space or expand the vertical scale by omitting the lower parts of bars is simply confusing.

WORKED EXAMPLE 4.3

Draw a bar chart of the regional sales in Worked Example 4.2.

Solution

Using a simple format, where the length of each bar corresponds **to the number** of sales in a region, gives the result shown in Figure 4.9.

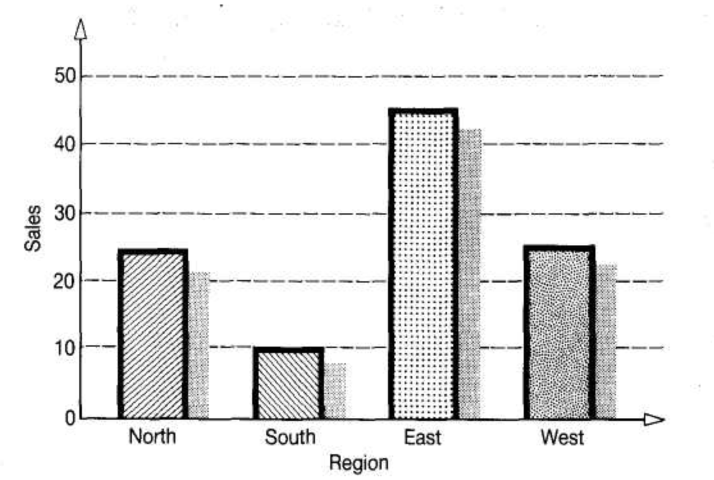


Figure 4.9 Bar chart of sales for Worked Example 4.3.

There are several different types of bar chart and the most appropriate is, again, a matter of choice. We should, however, remember that the purpose of diagrams is to present the characteristics of the data clearly; it is not necessarily to draw the prettiest picture. One particularly useful type of bar chart compares several sets of data, as illustrated in the following example.

WORKED EXAMPLE 4.4

There are five hospitals in a Health District, and they classify the number of beds in each hospital as follows.

	Hospital				
	Foothills	*General*	*Southern*	*Heathview*	*StJohn*
Maternity	24	38	6	0	0
Surgical	86	85	45	30	24
Medical	82	55	30	30	35
Psychiatric	25	22	30	65	76

Draw a bar chart to represent these data.

Solution

There are many possible formats for bar charts. Figure 4.10 shows a vertical form which has an added three-dimensional effect. This chart emphasizes the number of beds of each type, but if we wanted to highlight the relative sizes of the hospitals, we could 'stack' the bars to give the single bars shown in Figure 4.11.

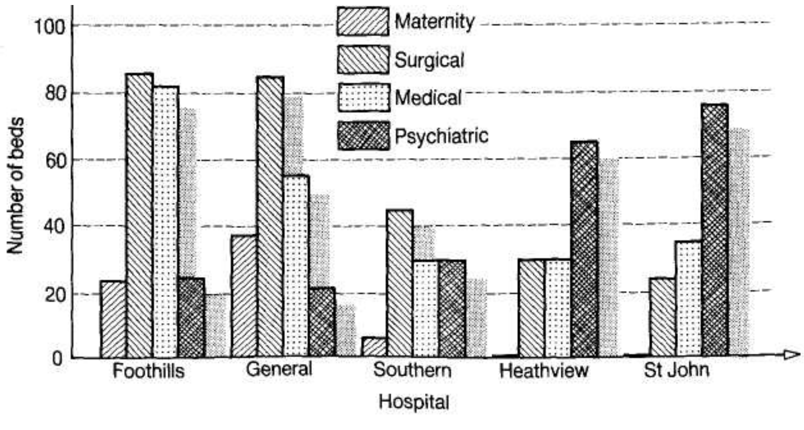


Figure 4.10 Number of beds in hospitals for Worked Example 4.4.

We could also represent these as percentages, as shown in Figure 4.12. There is an almost limitless variety of bar charts, and the most appropriate one to choose depends on the circumstances.

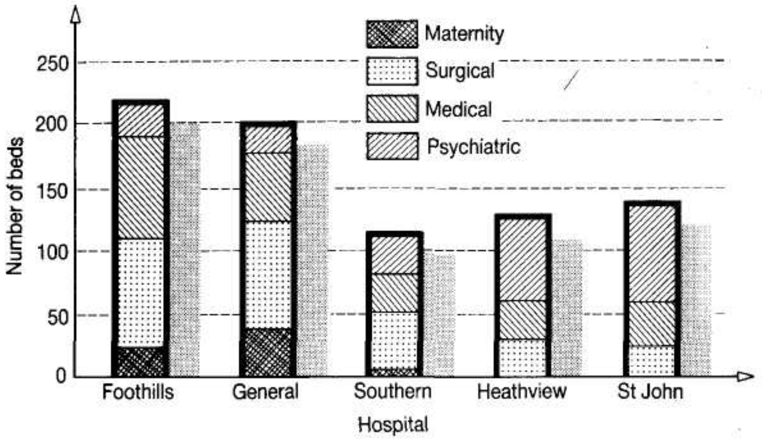


Figure 4.11 Emphasizing the number of beds in each hospital.

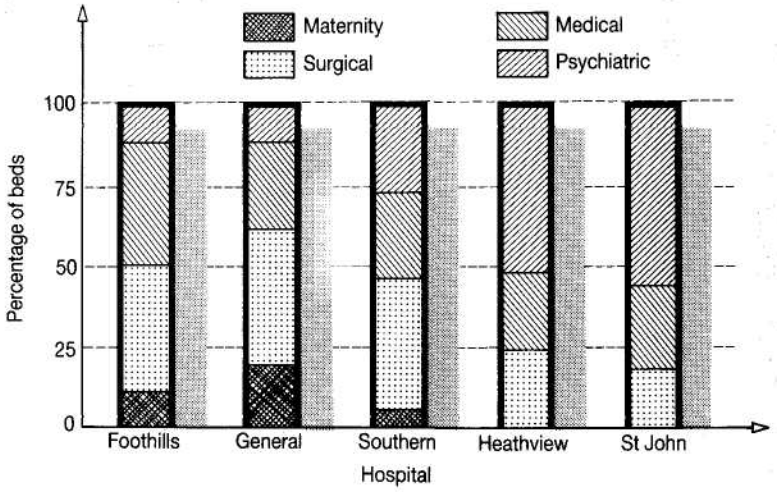


Figure 4.12 Percentage of beds.

IN SUMMARY

Bar charts can give flexible presentations. They use bars to represent categories, with the length of each bar proportional to the number of observations in the category.

4.2.6 Pictograms

These are similar to bar charts, except that the bars are replaced by sketches of the things being described. Thus the percentage of people owning cars might be represented as in Figure 4.13. In this pictogram, each 10% of people are represented by one car.

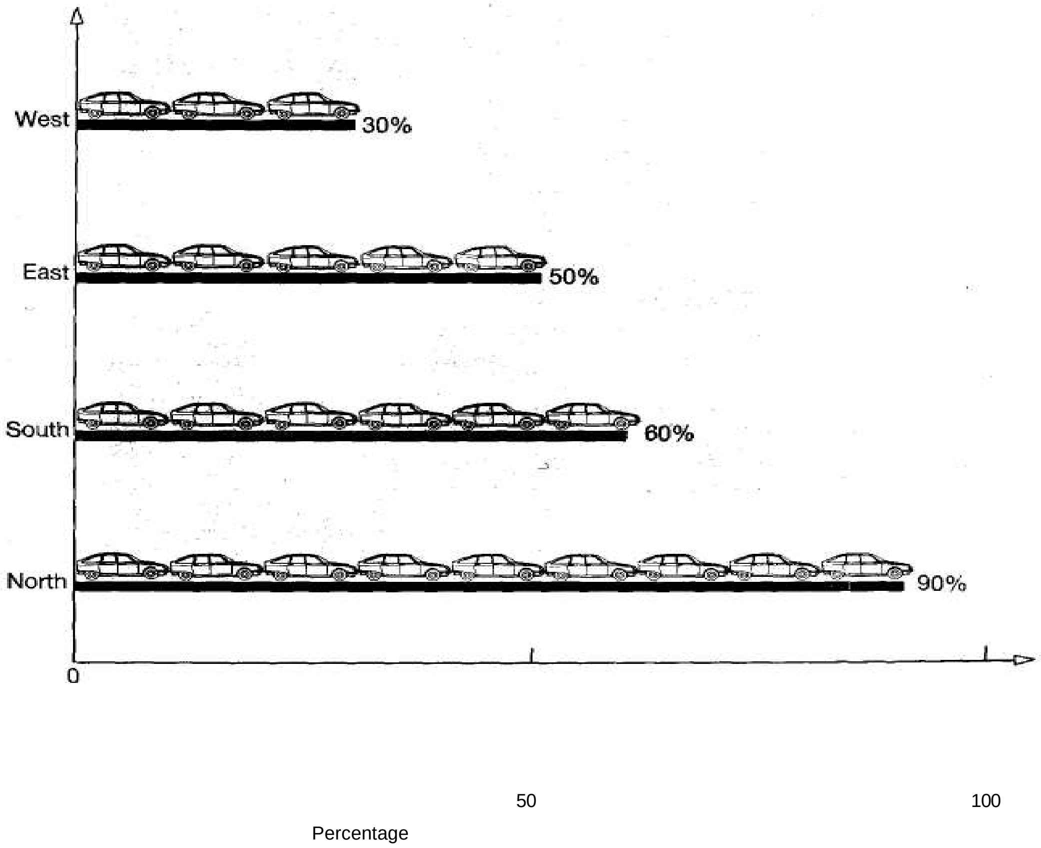


Figure 4.13 Pictogram showing percentage of people with cars.

Pictograms are very eye-catching and are, therefore, widely used in newspapers and magazines. They are not very accurate, but they are effective in giving general impressions. A problem arises with fractional values, such as 53% of people owning cars in Figure 4.13. This would be shown by a stack of 5.3 cars, and the 0.3 of a car clearly has little meaning. Nonetheless, the diagram would give the general impression of 'just over 50%'.

Pictograms should show different numbers of observations by different numbers of sketches, as shown in Figure 4.13. The wrong way to draw them is to make a single sketch bigger, as shown in Figure 4.14. The problem here is that we should be concentrating on the height of the sketches, but it is the area that has the immediate impact. If the number of observations is doubled, the sketch should also be doubled in height. Unfortunately, it is the area of the sketch that is noticed and this is increased by a factor of four. Figure 4.14 shows (as nuclear bomb mushroom clouds) the number of nuclear missiles built by 'Us' and 'Them'. All the figures are put on the graph to show that 'They' have just over twice as many missiles as 'Us', but it is the area of the graph that makes the impact, and this suggests a considerably greater difference.

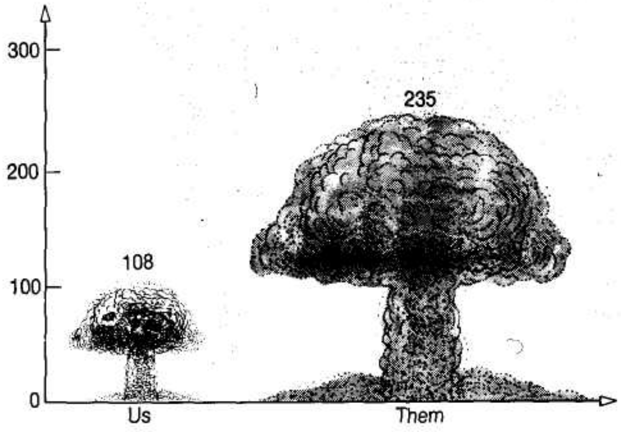


Figure 4.14 Poor pictogram showing the number of nuclear missiles.

IN SUMMARY

Pictograms replace the bars in bar charts by sketches. These can attract attention, but the results are not very accurate and need careful interpretation.

Self-assessment questions

4.5 What are the two main methods of presenting statistical data?

4.6 What are the advantages of using tables of data?

4.7 Why is it necessary to label the axes of graphs?

4.8 If you had a large quantity of numerical data what formats would you consider for their presentation?

4.9 'When using bar charts there is only one format that can be used for any set of data.' Is this statement true? Is it true of other methods of presenting data?

4.10 What are the main problems with pictograms?

4.3 Frequency distributions

Earlier in the chapter we showed that the number of observations in different classes could be drawn as a frequency table. In this section we are going to look at this idea in more detail.

Frequency tables

We have already met a frequency table of the form shown below. This divides weekly sales into a number of distinct classes and shows the number of weeks where demand fell in each class. The result is called a **frequency distribution.**

Class	*Number of weeks*
20 to 39	14
40 to 59	19
60 to 79	13
80 to 99	6

There are six observations in the highest class of sales, 80 to 99. Sometimes it is better to be less specific when defining classes, particularly if there are odd outlying values. If, for example, the data had included one observation of 120 it would be better to include this in the highest class, than create an additional class some distance from the others. Then we might define a class as '80 or more'. Similarly, it could be better to replace the precise '20 to 39' by the less precise '39 or fewer'.

When defining the boundaries between classes we must be sure there is no doubt about which class an observation is in. We would not, for example, have adjacent classes of '20 to 30' and '30 to 40', as a value of 30 could be in either one. To overcome this the classes are defined as '20 to 29' and '30 to 39'. This solution works if data are discrete (such as the number of sales) but is more difficult with continuous data. If, for example, we were classifying people by age we could not use classes '20 to 29' and '30 to 39', as this would leave no place for people who were 29.5. We must, therefore, describe the classes clearly and unambiguously, so that ages might be 'more than 20 and less than 30', and soon.

Most of the data described so far in this chapter have been discrete, but we should now move on and discuss continuous data in more detail. This is not a major step, as the comments we have made apply equally to discrete and continuous data. We can demonstrate this by drawing a frequency table of continuous data. The only thing we have to be careful about is defining the classes so that any observation can fall into one, and only one, class.

WORKED EXAMPLE 4.5

During a particular period the wages (in pounds) paid to 30 people have been recorded as follows:

202 457 310 176 480 277 87 391 325 120 554 94 362 221 274

145 240 437 404 398 361 144 429 216 282 153 470 303 338 209 Draw a frequency table of

these data.

Solution

The first decision concerns the number of classes. Although this is largely a subjective decision, the number should be carefully chosen. Too few classes (say, three) does not allow patterns to be highlighted; too many classes (say, 20) is confusing and too detailed. In this example we shall look for about six classes. Now we have to define ranges for our six classes. The range of wages is £87 to £554, and a suitable set of classes is:

'Less than £100', '£100 or more and less than £200', '£200 or more and less than £300', and so on

Notice that we are careful not to say 'more than £100 and less than £200', as someone might earn exactly £100 and would then not appear in any class.

Adding the number of observations in each class gives the following frequency table:

Class	*Frequency*
less than £100	2
£100 or more, but less than £200	5
£200 or more, but less than £300	8
£300 or more, but less than £400	9
£400 or more, but less than £500	5
£500 or more, but less than £600	1

This clearly shows the frequency distribution of wages, with more than half of people earning between £200 and £400.

Frequency distributions show the actual number of observations in each class. A useful extension is a **percentage frequency distribution,** which shows the percentage of observations in each class. The results are presented in exactly the same way as in standard frequency tables. The data in Worked Example 4.5 can be shown in the following percentage frequency distribution:

Class	*Frequency*	*Percentage frequency*
Less than £100	2	6.7
£100 or more, but less than £200	5	16.7
£200 or more, but less than £300	8	26.7
£300 or more, but less than £400	9	30.0
£400 or more, but less than £500	5	16.7
£500 or more, but less than £600	1	3.3

Another useful extension of frequency distributions looks at **cumulative frequencies.** Instead of recording the number of observations in a class, cumulative frequency distributions add all observations in lower classes. In the last table there were 2 observations in the first class, 5 in the second class and 8 in the third. The cumulative frequency distribution would show 2 observations in the first class, 2 + 5 = 7 in the second class and 2 + 5 + 8 = 15 in the third. We could also extend this into a **cumulative percentage frequency distribution,** as shown in Table 4.4.

Table 4.4

Class	*Frequency*	*Cumulative frequency*	*Percentage frequency*	*Cumulative percentage frequency*
Less than £100	*2*	*2*	6.7	6.7
£100 or more, but less than £200	*5*	7	16.7	23.3
£200 or more, but less than £300	8	15	26.7	50.0
£300 or more, but less than £400	9	24	30.0	80.0
£400 or more, but less than £500	5	29	16.7	96.7
£500 or more, but less than £600	1	30	3.3	100.0

The calculations for such tables are best done in the order: frequency distribution percentage frequency distribution cumulative frequency distribution cumulative percentage frequency distribution

We should also note that spreadsheets make such>calculations very easy.

WORKED EXAMPLE 4.6

Construct a table showing the frequency, cumulative frequency, percentage frequency and cumulative percentage frequency for the following discrete data:

150 141 158 147 132 153 176 162 180 165 174 133 129 119 103 188

190 165 157 146 161 130 122 169 159 152 173 148 154 171

Solution

We start by defining suitable classes, and as we do not know the purpose of the data, any suitable ones can be suggested. We shall note that the data are discrete and arbitrarily use 100 to 109,110 to 119, 120 to 129, and so on. The results are shown in Table 4.5. The cumulative percentage frequency distribution does not add up to 100% because of rounding.

Table 4.4

Class	*Frequency*	*Cumulative frequency*	*Percentage frequency*	*Cumulative percentage frequency*
100 to 109	1	1	3.3	3.3
110 to 119	1	2	3.3	6.6
120 to 129	2	4	6.7	13.3
130 to 139	3	7	10.0	23.3
140 to 149	4	11	13.3	36.6
150 to 159	7	18	23.3	59.9
160 to 169	5	23	16.7	76.6
170 to 179	4	27	13.3	89.9
180 to 189	2	29	6.7	96.6
190 to 199	1	30	3.3	99.9

IN SUMMARY

Frequency tables show the number of observations that fall into different classes. They can be used for both continuous and discrete data, and can be extended to show percentage frequency distributions and cumulative distributions.

4.3.2. Histograms

Histograms are diagrams of frequency distribution for continuous data. In appearance they are similar to bar charts, but there are some important differences. The most important difference is that histograms are only used for continuous data, so the horizontal axis has a continuous scale. Bars are drawn on this scale, so their width, as well as their height, has a definite meaning. This is an important point: in bar charts it is only the height of the bar that is important, but in histograms it is both the width and the height, or in effect the area. We can show this by drawing a histogram of the continuous data for wages shown above.

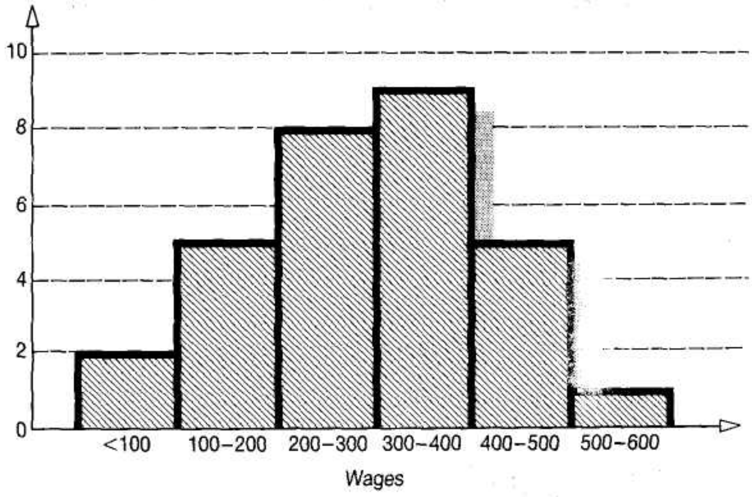


Figure 4.15 Histogram of wages.

In Figure 4.15 each class is the same width, so the areas are determined by the height of the bars. In effect, this is the same as a bar chart. Suppose, though, that the classes were of different widths. If we doubled the width of one class, we would have to halve its height to maintain the same area. This is demonstrated in the following worked example.

WORKED EXAMPLE 4.7

Draw a histogram of the following data:

Class	*Frequency*
Less than 10	8
10 or more, but less than 20	10
20 or more, but less than 30	16
30 or more, but less than 40	15
40 or more, but less than 50	11
50 or more, but less than 60	4
60 or more, but less than 70	2
70 or more, but less than 80	1
80 or more, but less than 90	1

Solution

Using the classes given we can draw the histogram shown in Figure 4.16. Because this diagram has a long tail with only eight observations in the last four classes, we might be tempted to combine these into one class with eight observations and then draw the histogram in Figure 4.17. This would be wrong, however. We cannot change the horizontal scale, so the single class would be four times as wide as the other classes. Making the last bar four units wide and eight units high would imply that it represents 32 observations instead of eight. As the area represents the number of observations the single last box should be four units wide and, therefore, two units high, as shown in Figure 4.18.

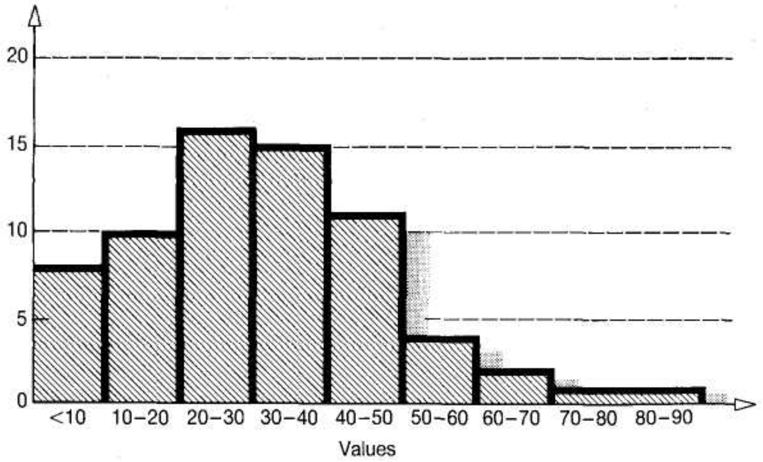


Figure 4.16 Histogram of value for Worked Example 4.7

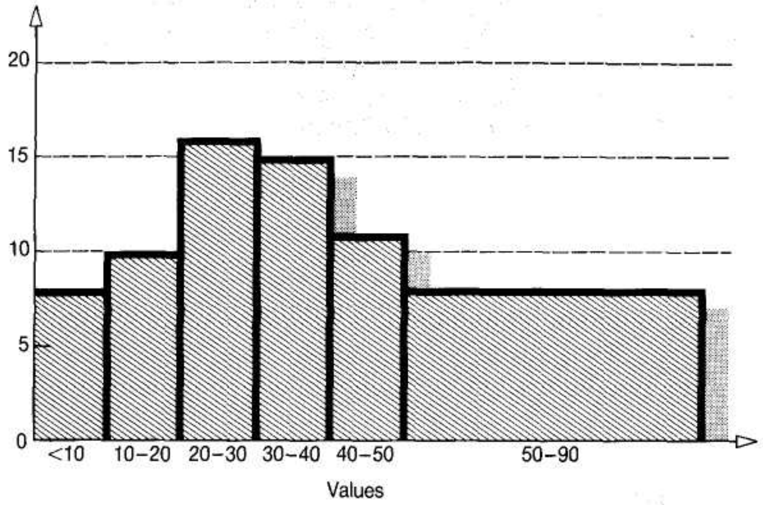


Figure 4.17 Incorrect histogram combining last few classes.

One problem with histograms occurs with open-ended categories. How, for example, do we deal with classes containing values 'greater than 20'? The answer (apart from avoiding such definitions wherever possible) is to make assumptions about the upper limit. By examining the data wexcan suggest an appropriate upper limit whereby 'greater than 20' might be interpreted as 'greater than 20 and less than 22'. Another consistent problem is that the shape of a histogram depends to a large extent on the way that classes are defined.

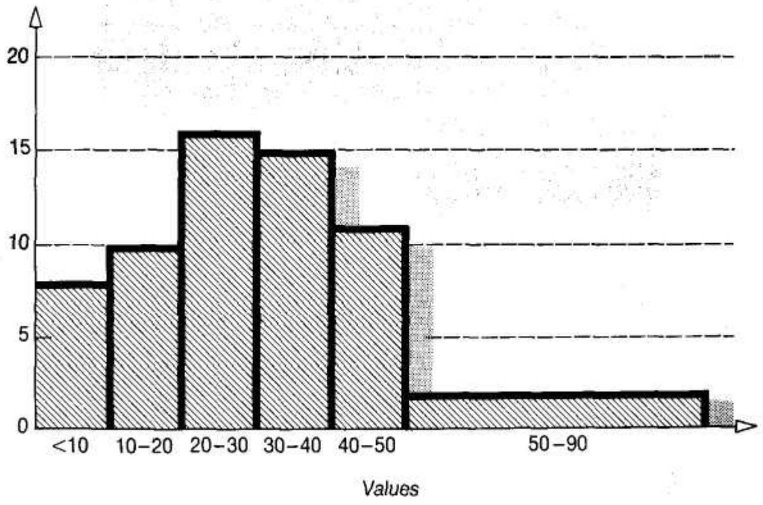


Figure 4.18 Correct histogram combining last few classes.

Overall, the use of histograms can be quite difficult. This is especially relevant as the less precise bar charts often give better -looking results with less effort, so there is some advantage in simply using bar charts rather than histograms. The main point in favour of histograms is that they allow further statistical analyses and they are, therefore, of considerable practical use.

IN SUMMARY

Histograms are diagrams of frequency distributions for continuous data. They represent frequencies by areas and are useful in further analyses. It is sometimes difficult to draw a reasonable histogram.

4.3.3 Summary charts

Earlier in this section we described how **tables can be drawn to show cumulative** frequency distributions, like the one below: ^

Class	*Frequency*	*Cumulative frequency*
100 or less	22	22
150 or less, but more than 100	44	66
200 or less, but more than 150	79	145
250 or less, but more than 200	96	241
300 or less, but more than 250	44	285
350 or less, but more than 300	15	300

This kind of result can be drawn on a graph relating cumulative frequency to class. The resulting curve has the scale of classes across the *x* axis and the cumulative frequency up the v axis and is called an **ogive.** We can start drawing an ogive of the data above by plotting the point (100,22) to show that 22 observations are in the class '100 or less'. The next point is (150,66) which shows that 66 observations are 150 or less, then the point (200,145) shows 145 observations are 150 or less, and so on. Plotting these points and joining them gives the result shown in Figure 4.19. We should note that ogives are always drawn vertically, and they are usually an elongated S shape.

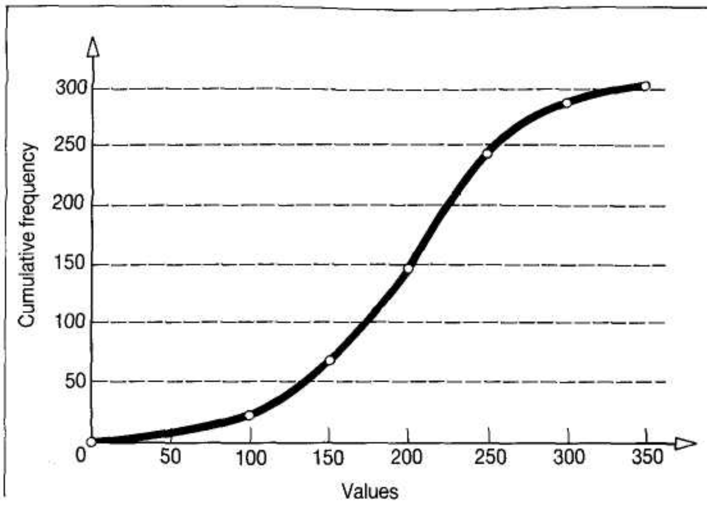


Figure 4.19 Ogive of observations.

A specific extension to ogives is a **Lorenz curve.** This is primarily used to describe the distribution of income or wealth among a population. It is a graph of cumulative percentage wealth, income or some other appropriate measure against cumulative percentage of the population.

WORKED EXAMPLE 4.8

Annual tax returns suggest that the percentages of a country's total wealth owned by various percentages of the population are as shown in the following table:

Percentage of population	*Percentage of wealth before tax*	*Percentage of wealth after tax*
45	5	15
20	10	15
15	15	15
10	10	15
5	15	15
3	25	15
2	20	10

(a) Draw a Lorenz curve of the wealth before tax.

(b) Draw a Lorenz curve of the wealth after tax. What conclusion can be drawn from these curves?

Solution

(a) A Lorenz curve plots the cumulative percentage of^ population against the corresponding cumulative percentage of wealth, so we first have to find these values, as shown in the following table:

Percentage of population	*Cumulative percentage of population*	*Percentage of wealth before tax*	*Cumulative percentage of wealth*
45	45	5	5
20	65	10	15
15	80	15	30 -
10	90	10	40
5	95	15	55
3	98 v	25	80
2	100	20	100

Now we plot these cumulative figures on a graph. The first point is (45,5), the second point is (65,15) and so on, as shown in Figure 4.20. A diagonal line has also been added to emphasize the shape of the Lorenz curve.

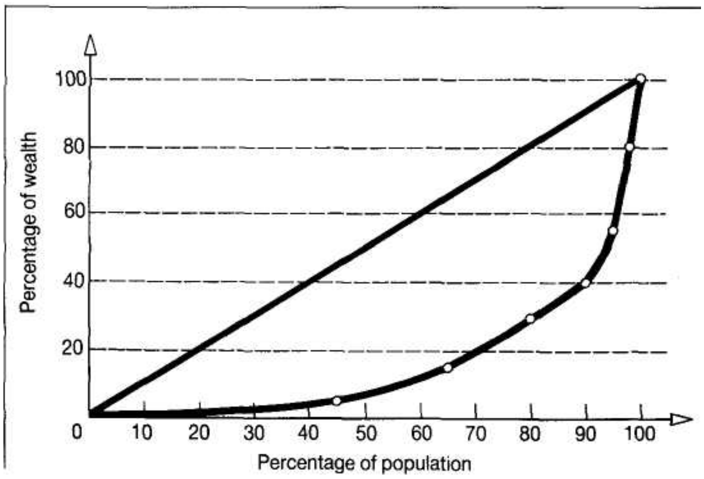


Figure 4.20 Lorenz curve for wealth before tax.

(b) Repeating the calculations for the after-tax wealth gives the following values:

Percentage of population	*Cumulative percentage of population*	*Percentage of wealth before tax*	*Cumulative percentage of wealth*
45	45	15	15
20	65	15	30
15	80	15	45
10	90	15	60
5	95	15	75
3	98	15	90
2	100	10	100

This gives the Lorenz curve shown in Figure 4.21.

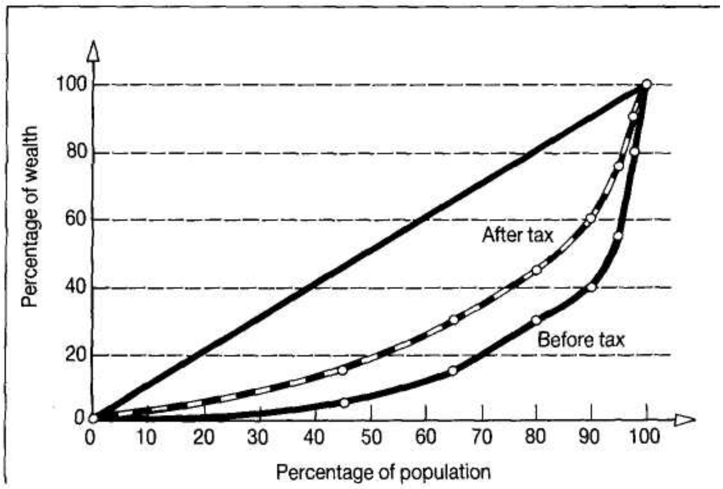


Figure 4.21 Lorenz curve for wealth after tax.

If the distribution of wealth is perfectly fair, we would get the diagonal straight line connecting the origin to the point (100, 100). If the graph is significantly below this, the distribution of wealth is unequal, and the further from the straight line the less equal is the distribution. The Lorenz curve for after-tax wealth is considerably closer to the straight line, and this shows that taxes have had an effect in redistributing wealth.

IN SUMMARY

Ogives are graphs of cumulative frequency against class. One extension of these is the Lorenz curve, which shows the distribution of income or wealth among a population. It can also be used for related measures, such as the effect of taxation.

Self-assessment questions

4.11 What is a frequency distribution?

4.12 What is the difference between a frequency distribution, a percentage frequency distribution, a cumulative distribution and a cumulative percentage distribution?

4.13 'In bar charts and histograms the height of the bar shows the number of observations in each class.' Is this statement correct?

4.14 If two classes of equal width are combined into one for a histogram, how high is the resulting bar?

4.15 What is the purpose of an ogive?

4.16 'A fair Lorenz curve should be a straight line connecting points (0, 0) and (100,100).' Is this statement true?

CHAPTER REVIEW

Once they have been collected, data must be processed to give useful information. This chapter considered one aspect of this processing, by describing how data can be summarized in diagrams. In particular it:

- discussed the purpose of data reduction as showing the overall pattern of data without getting bogged down in the details
- described alternative formats for data presentation, and suggested that the best depends on the purpose of the presentation, but is largely a matter of personal judgement

- designed tables of numerical data (which can show a lot of information but do not emphasize overall patterns)
- drew graphs to show relationships between variables
- drew pie charts, bar charts and pictograms to show relative frequencies
- described frequency tables and distributions, including percentage anc cumulative distributions
- drew histograms for continuous data
- drew ogives and Lorenz graphs

problems

4.1 Find some recent trade statistics published by the government and present these in different ways to emphasize different features. Discuss which

^ formats are fairest and which are most misleading.

4.2 A question in a survey gets the answer 'Yes' from 47% of men and 38% of women, 'No' from 32% of men and 53% of women, and 'Do not know' from the remainder. How could you present this result effectively?

43 The number of students taking a course in the past ten years is summarized in the following table:

Year	1	2	3	4	5	6	7	8	9	10
Male *Female*	21 4	22 6	20 3	18 5	28 12	26 16	29 14	30 19	32 17	29 25

Use a selection of graphical methods to summarize these data. Which do you think is the best?

4.4 Table 4.6 shows the quarterly profit reported by a company and the corresponding average price of its shares quoted on the London Stock Exchange. Devise suitable formats for presenting these data.

Table 4.6

Year	1				2				3			
Quarter	1	2	3	4	1	2	3	4	1	2	3	4
Profit	36	45	56	55	48	55	62	68	65	65	69	74
Share price	137	145	160	162	160	163	166	172	165	170	175	182

Source: company reports and the *Financial Times*

Note: profits are in millions of pounds; share prices are in pence

4.5 The number of people employed by Testel Electronics over the past ten years is as follows:

Year	1	2	3	4	5	6	7	8	9	10
Number	24	27	29	34	38	42	46	51	60	67

Design suitable ways of presenting these data>

4.6 Four regions of Yorkshire classify companies according to primary, manufacturing, transport, retail and service. The number of companies operating in each region in each category is shown below. Draw a number of bar charts to represent these data. Are bar charts the most appropriate format here?

	Industry type				
	Primary	*Manufacturing*	*Transport*	*Retail*	*Service*
Daleside	143	38	10	87	46
Twendale	134	89	15	73	39
Underhill	72	jV 67	11	165	55
Perithorp	54	41	23	287	89

4.7 The average wages of 45 people have been recorded as follows:

221 254 83 320 367 450 292 161 216 410 380 355 502 144 362 112 387 324 576 156 295 77 391 324 126 154 94 350 239 263 276 232 467 413 472 361 132 429 310 272 408 480 253 338 217

Draw a frequency table, percentage frequency and cumulative frequency table of these data. How could the data be presented in charts?

4.8 Draw a histogram of the following data:

Class	*Frequency*
Less than 100	120
100 or more, but less than 200	185
200 or more, but less than 300	285
300 or more, but less than 400	260
400 or more, but less than 500	205
500 or more, but less than 600	150
600 or more, but less than 700	75
700 or more, but less than 800	35
800 or more, but less than 900	15

Class	*Frequency*
Less than 2.5	0
2.5 or more, but less than 4.5	26
4.5 or more, but less than 6.5	40
6.5 or more, but less than 8.5	61
8.5 or more, but less than 10.5	75
10.5 or more, but less than 12.5	69
12.5 or more, but less than 14.5	55
14.5 or more, but less than 16.5	38
16.5 or more, but less than 18.5	15
18.5 or more	0

How could the last (a) two (b) three classes be combined?

4.9 Draw an ogive of the data in Problem 4.8.

4.10 Present the following data in a number of appropriate formats:

4.11 The wealth of a population is described in the following frequencj distribution. Draw Lorenz curves to represent this. Draw other appropriate graphs to represent the data

Percentage of people	5	10	15	20	20	15	10	5
Percentage of wealth before tax	20	1	3	6	15	20	20	15
Perce ntage of wealth after tax	15	3	6	10	16	20	20	10

Computer exercises

4.1 The following table shows last year's total production and profits (in consistent units) from six factories:

Factory	A	B	C	D	E	F
Production *Profit*	125 202	53 93	227 501	36 57	215 413	163 296

Use a graphics package to present this data in a number of formats.

(Note: The most appropriate software for this is some form of business graphics package, but alternatives include word processing, desktop publishing, statistical analysis packages and spreadsheets.)

4.2 The following data have been collected by a company. Put them into a spreadsheet, and hence reduce, manipulate and present them.

245 487 123 012 159 751 222 035 487 655 197 655 458 766 123 453 493 444 123 537 254 514 324 215 367 557 330 204 506 804 941 354 226 870 652 458 425 248 560 510 234 542 671 874 710 702 701 540 360 654 323 410 405 531 489 695 409 375 521 624 357 678 809 901 567 481 246 027 310 679 548 227 150 600 845 521 777 304 286 220 667 111 485 266 472 700 705 466 591 398 367 331 458 466 571 489 257 100 874 577

Now pass the data to a graphics package and describe them using suitable diagrams. Write a report on your findings.

4.3 The printout in Figure 4.22 uses Minitab to present a set of data (Cl). This gives a simplified version of a histogram and a frequency distribution. Use a statistical package to get equivalent results.

4.4 Figure 4.23 shows part of a spreadsheet. In this, a frequency table is drawn for a block of data. Use a spreadsheet to duplicate the results given and extend the analysis.

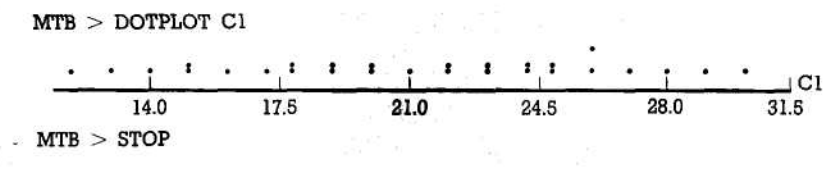
4.5 You are asked to prepare a report on the distance that people travel to get to work. Design a questionnaire to collect expected travel times for a large group of people. Now use this questionnaire to collect a set of real data. Use appropriate software to reduce and analyse the data. Now prepare your report in two formats:

- a written report
- the overhead slides to accompany a presentation to clients

```
MTB > SET Cl
DATA        >  12  13 18 24 28 17 25 18 14 30
DATA        > 29 15 15 16 19 20 20 21 22 23
DATA        > 24 26 22  19 25 24 23 27 END
MTB > HISTOGRAM Cl
Histogram of Cl    N = 28

                 Midpoint    Count
                       12        1
                       14        2
                       16        3
                       18        3
                       20        4
                       22        3
                       24        5
                       26        3
                       28        2
                       30
```

2



```
MTB > DOTPLOT C1
MTB > STOP
```

Figure 4.22 Example of Minitab printout.

	A	B	C	D	E	F	G	H
1	**DATA**							
2	5	4	1	2				
3	4	5	5	2				
4	6	7	6	4				
5	7	1	4	8				
6	1	5	3	7				
7	1	4	8	6				
8	2	8	7	6				
9	4	2	4	3				
10	5	4	2	4				
11	7	8	6	4				
12	8	6	4	5				
13	8	4	3	1				
14	9	8	7	2				
15	4	1	9	9		**CLASS FREQUENCY**		
16	5	2	5	7		**0**	**0**	
17	4	7	1	6		**2**	**16**	
18	5	9	2	4		**4**	**22**	
19	5	5	5	5		**6**	**22**	
20	4	1	8	3		**8**	**16**	
21	5	3	4	5		**10**	**4**	

Case study

High Acclaim Importers

The finance director of High Acclaim Importers was giving a summary of company business to a selected group of shareholders. He asked Jim Bowlers to collect some data from company records for his presentation.

At first Jim had been worried by the amount of detail available. The company seemed to keep enormous amounts of data on all aspects of its operations. These data ranged from transaction records in a computerized data base, to subjective management views which were never written down. The finance director had told Jim to give him some concise figure that could be used on overhead slides.

Jim did a conscientious job of collecting data and he looked pleased as he approached the finance director. As he handed over the results (Table 4.7), Jim explained: 'Some of our most important trading results are shown in this table. We trade in four regions, so for movements between each of these I have recorded seven key facts. The following table shows the number of units shipped (in hundreds), the average income per unit (in pounds sterling), the percentage gross profit, the percentage return on investment, a measure (between 1 and 5) of trading difficulty, the number of finance administrators employed in each area, and the number of agents. I thought you could make a slide of this and use it as a focus during your presentation.'

Table 4.7

From	*To*			
	Africa	*America*	*Asia*	*Europe*
Africa	105, 45,12, 4,4,15,4	85, 75,14, 7, 3, 20, 3	25, 60,15, 8, 3, 12, 2	160,80,13, 7, 2, 25, 4
America	45, 75,12, 3,4,15, 3	255,120,15, 9,1, 45, 5	60, 95, 8, 2, 2, 35, 6	345,115,10, 7,1, 65, 5
Asia	85, 70, 8, 4, 5, 20, 4	334,145,10, 5, 2, 55, 6	265, 85, 8, 3, 2, 65, 7	405,125,8, 3,2,70,8
Europe	100, 80,10, 5. 4, 30, 3	425,120,12, 8, 1, 70, 7	380,105, 9, 4, 2, 45, 5	555,140,10, 6, 110, 8

The finance director looked at the figures for a few minutes and then asked for some details on how trade had changed over the past ten years. Jim replied that in general terms the volume of trade had risen by 1.5, 3, 2.5, 2.5, 1, 1, 2.5, 3.5, 3 and 2.5% respectively in each of the last ten years, while the average price had risen by 4, 4.5, 5.5, 7, 3.5, 4.5, 6, 5.5, 5 and 5% respectively.

The finance director looked up from his figures and said: '1 am not sure these figures will have much impact on our shareholders. 1 was hoping for something a bit briefer and with a bit more impact. Could you give me the figures in a revised format by this afternoon?'

Your problem is to help Jim Bowlers to put the figures into a suitable format for presentation to shareholders.