**Aims and Objectives**

This guide explains the different types of average (mean, median and mode). It details their use, how to calculate them, and when they can be used most effectively.

**Introduction**

The term average is used frequently in everyday life to express an amount that is typical for a group of people or things. For example, you may read in a newspaper that on average people watch 3 hours of television per day. We understand from the use of the term average that not everybody watches 3 hours of television each day, but that some watch more and some less. However, we realize from the use of the term average that the figure of 3 hours per day is a good indicator of the amount of TV watched in general.

Averages are useful because they:

- summarise a large amount of data into a single value; and
- indicate that there is some variability around this single value within the original data.

In everyday language most people have an inherent understanding of what the term average means. However, within the language of mathematics there are three different definitions of average known as the mean, median and mode.

The mean, median and mode are each calculated using different methods and when applied to the same set of original data they often result in different average values. It is important to understand what each of these mathematical measures of average tells you about the original data and consider which measure, the mean, median or mode, is the most appropriate to calculate should you wish to use an average value to describe a dataset.

**Part one: The Mean**

What is the mean?
The mean is the most commonly used mathematical measure of average and is generally what is being referred to when people use the term average in everyday language. The mean is calculated by totalling all the values in a dataset; this total is then divided by the number of values that make up the dataset.

For example, to find out the mean amount borrowed by 6 students in a tutorial group taking out a student loan in 1998/9, the amounts borrowed by each student have been collected. These 6 amounts form the dataset given in table 1.

<table>
<thead>
<tr>
<th>Category of Student</th>
<th>Amount of Loan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male – living at home</td>
<td>£1,170</td>
</tr>
<tr>
<td>Male – living in private accommodation</td>
<td>£1,890</td>
</tr>
<tr>
<td>Male – living in university accommodation</td>
<td>£1,530</td>
</tr>
<tr>
<td>Female – living at home</td>
<td>£1,160</td>
</tr>
<tr>
<td>Female – living in private accommodation</td>
<td>£1,870</td>
</tr>
<tr>
<td>Female – living in university accommodation</td>
<td>£1,520</td>
</tr>
<tr>
<td>Total</td>
<td>£9,140</td>
</tr>
</tbody>
</table>

(Table 1: Amounts borrowed by 6 students taking out a student loan in 1998/9)

In order to find the mean loan, the total amount borrowed (£9,140) is divided by the number of students (6) which equals £1,523.

**Formula for the mean**

Whilst it is not vital to know the mathematical formula for the calculation of the mean you may want to include it at some point in a report or dissertation.

The formula for the mean is written in the following way:

$$
\bar{X} = \frac{\sum X}{n}
$$

$\bar{X}$ is the symbol for the mean and is referred to as bar X (ex)

$\Sigma$ is the Greek symbol sigma and simply means sum or add up

$X$ refers to each of the individual values that make up the dataset

$n$ is the number of values that make up the dataset

Re-writing the equation in words results in “the mean is equal to the sum of the individual values in the dataset, divided by the number of values in the dataset”.

**When to use the mean**
The mean is a good measure of the average when a dataset contains values that are relatively evenly spread with no exceptionally high or low values – this was the case with the data on student loans given in table 1.

If a dataset contains one or two very high or very low values the mean will be less typical as it will be adversely influenced by these exceptional value(s). This can be seen in table 2, where the mean salary of 6 graduates who responded to a survey about salaries in their first jobs is calculated to be £23,995 (£143,970 divided by 6).

<table>
<thead>
<tr>
<th>Name</th>
<th>Job</th>
<th>Annual Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clare</td>
<td>Web Designer</td>
<td>£19,100</td>
</tr>
<tr>
<td>Navroop</td>
<td>Teacher (London)</td>
<td>£21,650</td>
</tr>
<tr>
<td>Mark</td>
<td>Teacher (Swansea)</td>
<td>£18,750</td>
</tr>
<tr>
<td>Pablo</td>
<td>Trainee Supermarket Manager</td>
<td>£22,400</td>
</tr>
<tr>
<td>Sara</td>
<td>Temping</td>
<td>£14,870</td>
</tr>
<tr>
<td>Steve</td>
<td>Merchant Banker</td>
<td>£47,200</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>£143,970</strong></td>
</tr>
</tbody>
</table>

(Table 2: Graduate starting salaries)

Examining the dataset shows that 5 of the 6 graduates earn less than the mean salary of £23,995 and it is Steve’s exceptionally high salary that produces the high mean value. In this example, the mean gives a misleading impression of the amount a typical graduate earns in their first job. For datasets containing extremely high or low values the median (see next section) is a better measure of the average value.

When not to use the mean

The mean is generally an inappropriate measure of average for data that are measured on ordinal scales. Ordinal data are rated according to a category where a higher score indicates a higher or better rank than a lower score. Ordinal data are frequently used in surveys that ask people to indicate preference. The final information is relative and the difference between the ranks is not equal. For example, in response to a question regarding the flavour of a new blend of coffee a score of 10 implies a better taste than a score of 1 but it does not mean that the flavour is ten times as good!

Part two: The Median

What is the median?

The median refers to the middle value in a dataset, when the values are arranged in order of magnitude from smallest to largest or vice-versa. When there are an odd number of values in the dataset the middle value is straightforward to find. When there are an equal number of values, the mid-point between the two central values is the median.
For example, if the prices of seven sandwiches bought on campus are placed in order the median will be the 4th price in the sequence:

£1.10, £1.26, £1.30, [£1.40], £1.45, £1.85, £2.00

When the six starting salaries from example 2 are placed in order of magnitude the median value lies half-way between the 3rd and 4th salaries:

£14,870, £18,750, £19,100, £21,650, £22,400, £47,200

The median value lies half-way between £19,100 and £21,650 and is £20,375 \( ((£19,100 + £21,650) ÷ 2) \).

**When to use the median**

The median is a good measure of the average value when the data include exceptionally high or low values because these have little influence on the outcome.

The median is the most suitable measure of average for data classified on an ordinal scale.

The median is also easy to calculate but this does not imply that it is an inferior measure to the mean – what is important is to use an appropriate measure to determine the average.

Another area where the median is useful is with frequency data. Frequency data give the numbers of people or things in particular categories. For example, the frequency distribution of shoe sizes for a sample of 21 women was collected and is summarised in table 3.

<table>
<thead>
<tr>
<th>Shoe Size</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>

**Total** 21

*(Table 3: Frequency distribution of shoe sizes)*

A common mistake is to think that the median shoe size is 6 since this is the middle value in the first column. This is incorrect since it is the frequency information rather than the category (shoe size) that must be considered. There are 21 women in the sample so when the shoe sizes are arranged in order of magnitude the median shoe size will be placed 11th (mid-way) along that list. The 11th value in the frequency column corresponds with a shoe size of 5.
An alternative way of finding the median shoe size in this case is to re-write the data from the table to show each shoe size and the number of times it occurred in each category and use this to work out the median:

\[4, 4, 4, 4, 4, 5, 5, 5, 5, 5, [5], 6, 6, 6, 6, 6, 6, 7, 7, 8\]

Using this method it is easier to see that the median shoe size is 5.

**Part three: The Mode**

**What is the mode?**

The mode is the value that occurs with the greatest frequency in a dataset. It is representative or typical because it is the most common value. There may be more than one mode in a dataset if several values are equally common; alternatively there may be no mode. In example 3 (frequency distribution of shoe sizes), size 6 is the mode (or modal class), since this shoe size occurs the most frequently (7 times) in the sample.

**When to use the mode**

The mode is the only measure of average that can be used with nominal data. For example, late-night users of the library were classified by faculty as: 14% science students, 32% social science students, and 54% biological sciences students. No median or mean can be calculated but the mode is biological science students as students from this faculty were the most common.

**Part four: Calculating Averages Using Excel**

In addition to the manual methods described above, you can also calculate the mean, median and mode in Excel using special commands. The commands are entered into the formula bar towards the top of the spreadsheet and are preceded by \(=\). This use of \(=\) informs Excel that a calculation needs to be performed on the data. The corresponding cells in the spreadsheet show the result of the calculation. Example 4, below shows how Excel can be used to find the mean, median and mode of the student loan data originally given in example 1. Column A shows the different categories of student, whilst column B shows the amounts borrowed.

Example 4
To calculate a mean, Excel uses the term average and in the example spreadsheet, the command \( \text{=AVERAGE(B2:B7)} \) has been typed in cell D2. This command will automatically calculate the mean of the loans in cells 2 to 7 of column B. Excel performs the calculation instantly and the mean value of £1,523.33 is immediately shown in cell D2, however, the command used to perform the calculation remains displayed in the formula bar for as long as cell D2 is active (or highlighted) which is indicated by the box surrounding it.

In cell D4 the command for the mode has been entered \( \text{=MODE(B2:B7)} \); however, as there is no modal value in this dataset the result is given as \#N/A or \#VALUE.

In cell D6 the command for the median has been entered \( \text{=MEDIAN (B2:B7)} \)

These examples show the quick method of calculating averages using a cell range. Each of the commands can also be written out in a longer format with each of the different amounts of student loan entered as a separate value.

For example using the command \( \text{=MEDIAN(1170, 1890, 1530, 1160, 1870, 1520)} \) will produce an identical result to \( \text{=MEDIAN(B2:B7)} \). However, if the amount of one of the loans in column B is changed, the cell range method will automatically adjust the median, whereas the longer format will require manual adjustment of the command.
Working with percentages

Study guide

This guide is designed to help you use percentages in your work and daily life. It explains how to use percentages to compare values, how to quantify changes using percentages, and how to calculate the amount represented by a percentage increase or decrease.

Introduction

The term percentage or symbol % is used frequently in everyday language and life. For example, it is common to see sales with 20% discount, or restaurant bills with 10% service charges, as well as reports in newspapers discussing tax, unemployment and other values in percentage terms. Percentages are used in these ways as a simplified means of conveying size or scale or value.

Percentage means parts out of 100 and is the same as a fraction with a denominator (bottom) of 100. Therefore:

- 15% means 15 parts out of 100 and is the same as the fraction 15/100
- 87% means 87 parts out of 100 and is the same as the fraction 87/100

A further way of expressing parts out of 100 is using a decimal and so percentages can also be expressed as decimals:

- 15% is the same as 0.15 or 15/100
- 87% is the same as 0.87 or 87/100

There are three main ways in which percentages are frequently used:

- to enable data with different sample sizes or totals to be compared;
- to quantify the amount of change over time;
- to express an increase or reduction relative to initial size.

Using percentages to compare information

Whilst researching for an essay or dissertation you may come across many sources of data in tables, graphs or reports which you would like to incorporate into your work. However, this can be difficult if they do not share a common base line. Percentages are useful for comparing information where the sample sizes or totals are different. By converting different data to percentages you can readily compare them.

For example, the following table shows sales of CDs, LPs, cassettes and singles in 1996 and 1997.
Because the total music sales in 1996 and 1997 were different it is difficult to compare the data for the two years and to determine whether or not there was any notable change in the format in which music was bought. However, if the number of sales in each format (LP, CD etc.) is expressed as a percentage of the total sales, then it is easier to compare the data for the two years.

For example, the conversion from the actual number of CD sales in 1996 to a percentage can be done in the following way:

Determine the fraction of the total music sales represented by CDs in 1996:

**159.7 million out of a total of 286.6 million sales were of CDs**

Re-write this as a fraction: 159.7/286.6

Convert the fraction to a decimal by dividing **159.7 by 286.6**

\[
\frac{159.7}{286.6} = 0.5572
\]

Convert the decimal to a percentage by multiplying by 100

\[
0.5572 \times 100 = 55.72
\]

The result indicates that sales of CDs accounted for 55.72% of all music sales in 1996.

These four stages can be summarised by the following formula: \((159.7/289.6)\times(100/1)\)

This process can be repeated for all the music sales in 1996 and 1997 to produce table 2:

<table>
<thead>
<tr>
<th>Year</th>
<th>LPs</th>
<th>Cassettes</th>
<th>CDs</th>
<th>Singles</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>2.4</td>
<td>46.2</td>
<td>159.7</td>
<td>73.3</td>
<td>286.6</td>
</tr>
<tr>
<td>1997</td>
<td>2.5</td>
<td>36.6</td>
<td>158.6</td>
<td>67.0</td>
<td>284.9</td>
</tr>
</tbody>
</table>

Converting the music sales to percentages makes it easier to compare the sales according to the music format. For example, table 2 shows that in 1996 16.1% (or 16.1 out of every 100 music sales) was in the form of cassettes, but in 1997 this had fallen to 12.8%.

**Calculating percentage change**
Percentages are also very useful if you wish to quantify change. This is because they provide a result in the form of parts per hundred that is usually more readily understandable and comparable than when the information is presented as raw values.

Using the information presented in table 1 the decrease in cassette sales between 1996 and 1997 can be calculated as a percentage.

Determine the difference between the 1996 and 1997 cassette sales:

46.2 million (1996 sales) – 36.6 million (1997 sales) = 9.6 million

Express this difference as a fraction of the starting value (the 1996 sales of 46.2 million) and multiply the result by 100: 

\[
\frac{9,600,000}{46,200,000} \times 100 = 20.78\%
\]

This means that there was a fall of 20.78% in the number of cassette sales in 1997 as compared to 1996.

When calculating the percentage change between two values it is important to determine the correct base from which to calculate the percentage change (i.e. the appropriate starting value). This is because the percentage change from a low number to a higher number is not the same as the percentage change from the same higher number to the same lower number. For example:

The percentage increase from 50 to 75 = \( \frac{25}{50} \times 100 = 50\% \)

However, the percentage decrease from 75 to 50 = \( \frac{25}{75} \times 100 = 33.3\% \)

**Calculating percentage increase or reductions**

In the previous example, the percentage change between two values was calculated. The reverse process is where the actual amount represented by a particular percentage is to be calculated. Typical examples of the use of percentages in this way are in shop sales, where the prices have been reduced by a certain percentage, or bank charges where interest rates are expressed as a percentage. Again the use of the term percentage expresses the amount in terms of parts per hundred. For example, a bank interest rate of 6% is the same as saying the interest on each £1.00 invested is *6 parts out of 100* or 6p and a sale reduction of 12% means that an item will be reduced by 12p for every £1.00 it originally cost.

There are two ways of calculating the new value of the item following a percentage increase or reduction. For example you may wish to calculate the sale price of a TV normally priced at £180.00 that has been reduced by 15%.

The first method is to calculate the amount that the item has been reduced by:

\[
15\% \text{ of } £180 = \frac{15}{100} \times £180 = £27
\]
This amount is then subtracted from the original price:

£180 – £27 = £153

The second method is to calculate the percentage of the original price you are now paying for the item:

100% (full price) - 15% (discount) = 85% (discounted price)

This value is multiplied by the original price:

0.85 x £180 = £153 (this uses the decimal format explained above)

The result (£153) is the same whichever method you use.

You can use the same methods to calculate the value of percentage increases. For example, if you had savings of £112.72 a year ago that you put in an account earning 4.3% interest, you can calculate the current value of your savings. Again there are two methods.

Either calculate the actual amount by which your savings have increased:

4.3% of £112.72 = 4.3 ÷ 100 x £112.72 = £4.85

and add this to the original value of the savings:

£112.72 + £4.85 = £117.57

or calculate the percentage of your original savings you now have:

100% + 4.3% = 104.3%

and multiply this by the original value of your savings:

1.043 x £112.72 = £117.57 (this uses the decimal format explained on page 1)

Again the two methods produce the same result.

Summary

This guide has shown three main uses of percentages and how to perform calculations for each of these. Percentages are related to fractions and decimals and express values in the common currency of parts per hundred. Expressing values in the form of percentages in your work will enable you to readily compare information from different sources, quantify change over time and find the amount by which something has increased or decreased following a percentage change. Whilst you will come across data that is already presented as percentages remember that you can also convert data from
tables and graphs that you find in books, journals and other publications into percentages in order to make useful comparisons.
Measures of variability: the range, inter-quartile range and standard deviation

Aims and Objectives

This guide outlines three methods used to summarise the variability in a dataset. It will help you identify which measure is most appropriate to use for a particular set of data. Examples are also given of the use of these measures and how the standard deviation can be calculated using Excel.

Introduction

Measures of average such as the median and mean represent the typical value for a dataset. Within the dataset the actual values usually differ from one another and from the average value itself. The extent to which the median and mean are good representatives of the values in the original dataset depends upon the variability or dispersion in the original data. Datasets are said to have high dispersion when they contain values considerably higher and lower than the mean value.

In figure 1 the number of different sized tutorial groups in semester 1 and semester 2 are presented. In both semesters the mean and median tutorial group size is 5 students, however the groups in semester 2 show more dispersion (or variability in size) than those in semester 1.

Dispersion within a dataset can be measured or described in several ways including the range, inter-quartile range and standard deviation.

The Range

The range is the most obvious measure of dispersion and is the difference between the lowest and highest values in a dataset. In figure 1, the size of the largest semester 1 tutorial group is 6 students and the size of the smallest group is 4 students, resulting in a range of 2 (6 - 4). In semester 2, the largest tutorial group size is 7 students and the smallest tutorial group contains 3 students, therefore the range is 4 (7 - 3).

- The range is simple to compute and is useful when you wish to evaluate the whole of a dataset.
The range is useful for showing the spread within a dataset and for comparing the spread between similar datasets.

An example of the use of the range to compare spread within datasets is provided in table 1. The scores of individual students in the examination and coursework component of a module are shown.

<table>
<thead>
<tr>
<th>Student</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coursework mark</td>
<td>27</td>
<td>44</td>
<td>39</td>
<td>23</td>
<td>41</td>
<td>48</td>
<td>37</td>
<td>34</td>
<td>40</td>
<td>43</td>
<td>30</td>
<td>43</td>
<td>29</td>
<td>27</td>
</tr>
<tr>
<td>Examination mark</td>
<td>12</td>
<td>47</td>
<td>26</td>
<td>28</td>
<td>30</td>
<td>45</td>
<td>35</td>
<td>35</td>
<td>41</td>
<td>39</td>
<td>32</td>
<td>25</td>
<td>18</td>
<td>30</td>
</tr>
</tbody>
</table>

To find the range in marks the highest and lowest values need to be found from the table. The highest coursework mark was 48 and the lowest was 27 giving a range of 21. In the examination, the highest mark was 45 and the lowest 12 producing a range of 33. This indicates that there was wider variation in the students’ performance in the examination than in the coursework for this module.

Since the range is based solely on the two most extreme values within the dataset, if one of these is either exceptionally high or low (sometimes referred to as outlier) it will result in a range that is not typical of the variability within the dataset. For example, imagine in the above example that one student failed to hand in any coursework and was awarded a mark of zero, however they sat the exam and scored 40. The range for the coursework marks would now become 48 (48-0), rather than 21, however the new range is not typical of the dataset as a whole and is distorted by the outlier in the coursework marks. In order to reduce the problems caused by outliers in a dataset, the inter-quartile range is often calculated instead of the range.

The Inter-quartile Range

The inter-quartile range is a measure that indicates the extent to which the central 50% of values within the dataset are dispersed. It is based upon, and related to, the median.

In the same way that the median divides a dataset into two halves, it can be further divided into quarters by identifying the upper and lower quartiles. The lower quartile is found one quarter of the way along a dataset when the values have been arranged in order of magnitude; the upper quartile is found three quarters along the dataset. Therefore, the upper quartile lies half way between the median and the highest value in the dataset whilst the lower quartile lies halfway between the median and the lowest value in the dataset. The inter-quartile range is found by subtracting the lower quartile from the upper quartile.

For example, the examination marks for 20 students following a particular module are arranged in order of magnitude.
The median lies at the mid-point between the two central values (10th and 11th) = half-way between 60 and 62 = 61

The lower quartile lies at the mid-point between the 5th and 6th values = half-way between 52 and 53 = 52.5

The upper quartile lies at the mid-point between the 15th and 16th values = half-way between 70 and 71 = 70.5

The inter-quartile range for this dataset is therefore 70.5 - 52.5 = 18 whereas the range is: 80 - 43 = 37.

The inter-quartile range provides a clearer picture of the overall dataset by removing/ignoring the outlying values.

Like the range however, the inter-quartile range is a measure of dispersion that is based upon only two values from the dataset. Statistically, the standard deviation is a more powerful measure of dispersion because it takes into account every value in the dataset. The standard deviation is explored in the next section of this guide.

**Calculating the Inter-quartile range using Excel**

The method Excel uses to calculate quartiles is not commonly used and tends to produce unusual results particularly when the dataset contains only a few values. For this reason you may be best to calculate the inter-quartile range by hand.

**The Standard Deviation**

The standard deviation is a measure that summarises the amount by which every value within a dataset varies from the mean. Effectively it indicates how tightly the values in the dataset are bunched around the mean value. It is the most robust and widely used measure of dispersion since, unlike the range and inter-quartile range, it takes into account every variable in the dataset. When the values in a dataset are pretty tightly bunched together the standard deviation is small. When the values are spread apart the standard deviation will be relatively large. The standard deviation is usually presented in conjunction with the mean and is measured in the same units.

In many datasets the values deviate from the mean value due to chance and such datasets are said to display a normal distribution. In a dataset with a normal
distribution most of the values are clustered around the mean while relatively few values tend to be extremely high or extremely low. Many natural phenomena display a normal distribution.

For datasets that have a normal distribution the standard deviation can be used to determine the proportion of values that lie within a particular range of the mean value. For such distributions it is always the case that 68% of values are less than one standard deviation (1SD) away from the mean value, that 95% of values are less than two standard deviations (2SD) away from the mean and that 99% of values are less than three standard deviations (3SD) away from the mean. Figure 3 shows this concept in diagrammatical form.

If the mean of a dataset is 25 and its standard deviation is 1.6, then

- 68% of the values in the dataset will lie between \textbf{MEAN-1SD} (25-1.6=23.4) and \textbf{MEAN+1SD} (25+1.6=26.6)
- 99% of the values will lie between \textbf{MEAN-3SD} (25-4.8=20.2) and \textbf{MEAN+3SD} (25+4.8=29.8).

If the dataset had the same mean of 25 but a larger standard deviation (for example, 2.3) it would indicate that the values were more dispersed. The frequency distribution for a dispersed dataset would still show a normal distribution but when plotted on a graph the shape of the curve will be flatter as in figure 4.
Population and sample standard deviations

There are two different calculations for the Standard Deviation. Which formula you use depends upon whether the values in your dataset represent an entire population or whether they form a sample of a larger population. For example, if all student users of the library were asked how many books they had borrowed in the past month then the entire population has been studied since all the students have been asked. In such cases the population standard deviation should be used. Sometimes it is not possible to find information about an entire population and it might be more realistic to ask a sample of 150 students about their library borrowing and use these results to estimate library borrowing habits for the entire population of students. In such cases the sample standard deviation should be used.

Formulae for the standard deviation

Whilst it is not necessary to learn the formula for calculating the standard deviation, there may be times when you wish to include it in a report or dissertation.

The standard deviation of an entire population is known as $\sigma$ (sigma) and is calculated using:

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

Where $x$ represents each value in the population, $\mu$ is the mean value of the population, $\Sigma$ is the summation (or total), and $N$ is the number of values in the population.

The standard deviation of a sample is known as $S$ and is calculated using:

$$S = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$
Where $x$ represents each value in the population, $\bar{x}$ is the mean value of the sample, $\Sigma$ is the summation (or total), and $n-1$ is the number of values in the sample minus 1.

**Calculating the standard deviation using Excel**

Excel has functions to calculate the population and sample standard deviations. The appropriate commands are entered into the formula bar towards the top of the spreadsheet and the corresponding cells in the spreadsheet are updated to show the result.

For an example of calculating the population standard deviation, imagine you wish to know how fuel-efficient a new car that you have just purchased is. You calculate how many kilometres you have done per litre on your first five trips. This information is presented as column A of the spreadsheet (figure 5). As you have only made 5 trips you do not have any further information and you are therefore measuring the whole population at this point in time. The command to find the population standard deviation in Excel is $=\text{STDEVP}(\text{VALUES})$ and in this case the command is $=\text{STDEVP}(A2:A6)$ which gives an answer of 0.49.

Basing your results on the population standard deviation and assuming that your first 5 trips in your new car have been typical of your usual journeys, you can be 99% confident that your new car will do between $14.75$ ($\text{MEAN}-3\text{SD}$) and $17.69$ ($\text{MEAN}+3\text{SD}$) kilometres per litre.

The same data can be used to demonstrate how to calculate the sample standard deviation in Excel. In this case, imagine that the data in column A represent the kilometres per litre found for a sample of 5 new cars tested by the manufacturer. The population standard deviation is calculated using $=\text{STDEV}(\text{VALUES})$ and in this case the command is $=\text{STDEV}(A2:A6)$ which produces an answer of 0.55.

The sample standard deviation will always be greater than the population standard deviation when they are calculated for the same dataset. This is because the formula for
the sample standard deviation has to take into account the possibility of there being more variation in the true population than has been measured in the sample.

Based on their sample of 5 cars, and therefore using the sample standard deviation, the manufacturers could state with 99% confidence that similar cars will do between 14.57 (\text{MEAN-3SD}) and 17.87 (\text{MEAN+3SD}) kilometres per litre.

These examples show the quick method of calculating standard deviations using a cell range. Each of the commands can also be written out in a longer format with the individual kilometres/litre entered.

For example entering: \text{=STDEV(16.13,16.40,15.81,17.07,15.69)} produces an identical result to \text{=STDEV(A2:A6)}. However, if one of the values in column A was found to be incorrect and adjusted, the cell range method would automatically update the calculation of the standard deviation whereas the longer format will require manual adjustment of the command.

\textbf{Summary}

The range, inter-quartile range and standard deviation are all measures that indicate the amount of variability within a dataset. The range is the simplest measure of variability to calculate but can be misleading if the dataset contains extreme values. The inter-quartile range reduces this problem by considering the variability within the middle 50% of the dataset. The standard deviation is the most robust measure of variability since it takes into account a measure of how every value in the dataset varies from the mean. However, care must be taken when calculating the standard deviation to consider whether the entire population or a sample is being examined and to use the appropriate formula.
Presenting numerical data

Aims and Objectives

This guide offers practical advice on how to incorporate numerical information into essays, reports, dissertations, posters and presentations. The guide outlines the role of text, tables, graphs and charts as formats for presenting numerical data. It focuses on issues that should be addressed when presenting numerical data for different audiences and highlights ways that will maximise the impact of such data and ensure that they are easy to read and interpret.

Introduction

It is likely that there will be occasions when you have numerical information that you want to include in your work, for example figures and other statistics from secondary sources (such as books, journal articles or newspaper reports); the results of experiments; or data that you have collected and analysed as part of a project or dissertation. Such information can be used to illustrate an argument or convey complex or detailed information in a concise manner.

There are three main methods of presenting such information:

- it can be incorporated into the main body of text;
- it can be presented separately as a table; or
- it can be used to construct a graph or chart.

Determining which of these methods is the most appropriate depends upon the amount of data you are dealing with and their complexity. The choice about whether to use text, tables or graphs requires careful consideration if you are to ensure that your reader or audience understands your argument and is not left struggling to interpret data that are poorly presented or in an inappropriate format. It is crucial to remember that when using a table or graph the associated text should describe what the data reveal about the topic; you should not need to describe the information again in words.

Including numbers in the main body of text

Numbers are most effective in the main body of the text of an essay, report or dissertation when there are only two values to compare.

For example:

86% of male students said they regularly ate breakfast compared to 62% of female students.

If you are discussing three or more numbers, including them within the main body of text does not facilitate comprehension or comparison and it is often more useful to use a table incorporated within the text.

For example:
53% of male students said that they always ate breakfast, 33% said that they usually did, and 14% said that they never ate breakfast.

Is more clearly expressed as:

Male students said they ate breakfast:

- Always 53%
- Usually 33%
- Never 14%

In order to help the reader compare the numbers it is also useful to list them according to their magnitude (e.g. from large to small) unless there is a particular pattern or trend in the data that you want to highlight.

In general, numbers are usually given as digits rather than spelt out in the text, e.g. 400 rather than four hundred. However, in some academic journals the convention is to spell out whole numbers between one and ten and use values for all other numbers - so you may wish to find out what the usual practice is within your own discipline.

**Presenting numbers in tables**

Tables are used to present numerical data in a wide variety of publications from newspapers, journals and textbooks to the sides of grocery packets. They are the format in which most numerical data are initially stored and analysed and are likely to be the means you use to organise data collected during experiments and dissertation research. However, when writing up your work you will have to make a decision about whether a table is the best way of presenting the data, or if it would be easier to understand if you were to use a graph or chart.

This section of the guide identifies the appropriate uses of tables, and discusses some design issues for constructing clear tables which are easy to interpret. The points covered in this guide apply equally to primary data that you have collected yourself, and to data that you have found in secondary sources and which you wish to include in your work. The latter may already be presented as a table in the original work but you do not have to reproduce it exactly. It may be that you only require an extract from the table to support your argument, or that the design of the table could be improved, or that you wish to merge information from two different publications. There is no problem in doing any of these as long as you ensure that you reference the original source of the data in your table.

**When to use tables**

Tables are an effective way of presenting data:

- when you wish to show how a single category of information varies when measured at different points (in time or space). For example, a table would be an appropriate way of showing how the category unemployment rate varies between different countries in the EU (different points in space);
when the dataset contains relatively few numbers. This is because it is very hard for a reader to assimilate and interpret many numbers in a table. In particular, avoid the use of complex tables in talks and presentations when the audience will have a relatively short time to take in the information and little or no opportunity to review it at a later stage;

- when the precise value is crucial to your argument and a graph would not convey the same level of precision. For example, when it is important that the reader knows that the result was 2.48 and not 2.45;

- when you don’t wish the presence of one or two very high or low numbers to detract from the message contained in the rest of the dataset. For example if you are presenting information about the annual profits of an organisation and don’t want the underlying variability from one year to the next to be swamped by a large loss in a particular year.

**Table design**

In order to ensure that your table is clear and easy to interpret there are a number of design issues that need to be considered. These are listed below:

- Since tables consist of rows and columns of information it is important to consider how the data are arranged between the two. Most people find it easier to identify patterns in numerical data by reading down a column rather than across a row. This means that you should plan your row and column categories to ensure that the patterns you wish to highlight are revealed in the columns. It is also easier to interpret the data if they arranged according to their magnitude so there is numerical progression down the columns, although this may not always be possible.

- If there are several columns or categories of information a table can appear complex and become hard to read. It also becomes more difficult to list the data by magnitude since the order that applies to one column may not be the same for others. In such cases you need to decide which column contains the most important trend and this should be used to structure the table. If the columns are equally important it is often better to include two or more simple tables rather than using a single more complex one.

- Numbers in tables should be presented in their most simple format. This may mean rounding up values to avoid the use of decimal places, stating the units (e.g. £4.6 million rather than £4,600,000) or using scientific notation (e.g. 6.315 x 10^-2 rather than 0.06315).

- All tables should be presented with a title that contains enough detail that a reader can understand the content without needing to consult the accompanying text. There should also be information about the source of the data being used; this may be a reference to a book or journal, or could indicate that the data are results from an experiment carried out on a particular date.

- Where more than one table is being presented it is standard practice to give each one a unique reference number, and in larger pieces of work, such as dissertations, a list of tables with their page number is usually provided in addition to the contents page.

- The formatting of the table should not resemble a spreadsheet where each entry is bounded by a box since this makes it difficult to read across rows or down
columns. However, the design of the table should help the reader interpret the data and so the use of lines and/or bold text to separate headings from the body of data, or highlighting/shading specific rows or may be effective. Avoid large gaps between columns since this also makes it difficult to read along a row.

Examples of poor and better practice in the presentation of data in tables

Poor example

<table>
<thead>
<tr>
<th>Region</th>
<th>% adults taking a holiday</th>
</tr>
</thead>
<tbody>
<tr>
<td>East Anglia</td>
<td>60</td>
</tr>
<tr>
<td>East Midlands</td>
<td>64</td>
</tr>
<tr>
<td>Greater London</td>
<td>56</td>
</tr>
<tr>
<td>Humberside and Yorkshire</td>
<td>64</td>
</tr>
<tr>
<td>North</td>
<td>54</td>
</tr>
<tr>
<td>North West</td>
<td>59</td>
</tr>
<tr>
<td>South East</td>
<td>60</td>
</tr>
<tr>
<td>South West</td>
<td>81</td>
</tr>
<tr>
<td>West Midlands</td>
<td>56</td>
</tr>
</tbody>
</table>

(Adaptation of Table 1 (below), originating from Regional Trends (2000), Issue 35, Section 8, Table 8.16, Crown Copyright).

This is a poor example because:

- the table lacks a title
- the source of the information is not provided
- row titles straddle two lines
- each cell is bounded as if in a spreadsheet
- the alphabetical listing of regions results in a non-numerical ordering of data down the columns

Better example

Table 1: Regional Differences in the percentage of adults taking a holiday in 1998, from Regional Trends (2000), Issue 35, Section 8, Table 8.16, Crown Copyright.
Graphs are a good means of describing, exploring or summarising numerical data because the use of a visual image can simplify complex information and help to highlight patterns and trends in the data. They are a particularly effective way of presenting a large amount of data but can also be used instead of a table to present smaller datasets. There are many different graph types to choose from and a critical issue is to ensure that the graph type selected is the most appropriate for the data. Having done this, it is then essential to ensure that the design and presentation of the graph help the reader or audience interpret the data.

A summary of the types of data that can be presented in the most common types of graphs is provided below and this is followed by some general guidelines for designing readily understandable graphs. There is more detailed information on the uses and good design of particular types of graph in the companion study guides covering bar charts, histograms, pie charts, line graphs and scatter plots available from the Student Learning Centre.

### Types of Graph

#### Bar charts

Bar charts are one of the most commonly used types of graph and are used to display and compare the number, frequency or other measure (e.g. mean) for different discrete categories or groups. The graph is constructed such that the heights or lengths of the different bars are proportional to the size of the category they represent. Since the x-axis (the horizontal axis) represents the different categories it has no scale. The y-axis (the vertical axis) does have a scale and this indicates the units of measurement. The bars can be drawn either vertically or horizontally depending upon the number of categories and length or complexity of the category labels. There are various ways in which bar charts can be constructed and this makes them a very flexible chart type. For example, if there is more than one set of values for each category then grouped or component bar charts can be used to display the data. Further details about each of these different types of bar chart can be found in the associated study guide Bar Charts.
Histograms

Histograms are a special form of bar chart where the data represent continuous rather than discrete categories. For example, a histogram could be used to present details of the average number of hours exercise carried out by people of different ages because age is a continuous rather than a discrete category. However, because a continuous category may have a large number of possible values the data are often grouped to reduce the number of data points. For example, instead of drawing a bar for each individual age between 0 and 65, the data could be grouped into a series of continuous age ranges such as 16-24, 25-34, 35-44 etc. Unlike a bar chart, in a histogram both the x- and y-axes have a scale. This means that it is the area of the bar that is proportional to the size of the category represented and not just its height. Further information on constructing histograms is available in the associated study guide Histograms.

Pie charts

Pie charts are a visual way of displaying how the total data are distributed between different categories. The example here shows the proportional distribution of visitors
between different types of tourist attractions. Similar uses of a pie chart would be to show the percentage of the total votes received by each party in an election. Pie charts should only be used for displaying nominal data (i.e. data that are classed into different categories). They are generally best for showing information grouped into a small number of categories and are a graphical way of displaying data that might otherwise be presented as a simple table. The study guide Pie Charts gives more details about designing pie charts and using them to compare data.

![Pie chart showing distribution of visitors among the most popular types of tourist attractions in Great Britain, 1991.](image)


### Line graphs

Line graphs are usually used to show time series data – that is how one or more variables vary over a continuous period of time. Typical examples of the types of data that can be presented using line graphs are monthly rainfall and annual unemployment rates. Line graphs are particularly useful for identifying patterns and trends in the data such as seasonal effects, large changes and turning points. As well as time series data, line graphs can also be appropriate for displaying data that are measured over other continuous variables such as distance. For example, a line graph could be used to show how pollution levels vary with increasing distance from a source, or how the level of a chemical varies with depth of soil. However, it is important to consider whether the data have been collected at sufficiently regular intervals so that estimates made for a point lying half-way along the line between two successive measurements would be reasonable. In a line graph the x-axis represents the continuous variable (for example year or distance from the initial measurement) whilst the y-axis has a scale and indicates the measurement. Several data series can be plotted on the same line chart and this is particularly useful for analysing and comparing the trends in different datasets.
Scatter plots

Scatter plots are used to show the relationship between pairs of quantitative measurements made for the same object or individual. For example, a scatter plot could be used to present information about the examination and coursework marks for each of the students in a class. In the example here, the paired measurements are the age and height of children in 1837. In a scatter plot a dot represents each individual or object (child in this case) and is located with reference to the x-axis and y-axis, each of which represent one of the two measurements. By analysing the pattern of dots that make up a scatter plot it is possible to identify whether there is any systematic or causal relationship between the two measurements. For example, in this case it is clear from the upward trending pattern of dots that children's height increases with age. Regression lines can also be added to the graph and used to decide whether the relationship between the two sets of measurements can be explained or if it is due to chance.

Good graph design

Although there are many different types of graph, there are a number of elements that are common to the majority of them such as axes. This section provides some general
guidelines to help you design your graph and ensure that you apply these elements in a way that will help the reader or audience interpret the data you are presenting.

**Components of a graph**

The different components of a graph are identified in the diagram on the next page and this is followed by a description that highlights some of the specific design and presentation issues related to each component.

**Chart area**

The chart area defines the boundary of all the elements related to the graph including the plot itself and any headings and explanatory text. It emphasises that these elements need to be considered together and that they are separate from the surrounding text. The boundary of the chart area can be imaginary rather than defined by a frame.

**Plot area**

The plot area is the region containing the data. It is bounded by the x- and y-axes to the bottom and left side. The frame can be completed by drawing around the top and right sides too, but this is not essential.

**The x-axis**

The x-axis is the horizontal line that defines the base of the plot area. Depending upon which type of graph is being considered different locations on the x-axis represent either different categories (such as years) or different positions along a numerical scale (such as temperature or income). Details are placed just below the x-axis and an axis label is usually provided to clarify the units of measurement. However, if the category details are mentioned somewhere else such as in the title of the graph, or are very obvious (such as years) then it is not necessary to include an axis label.

**The y-axis**

The y-axis is the vertical line that usually defines the left side of the plot area, but if more than one variable is being plotted on the graph then the vertical lines on both the left and right sides of the plot area may be used as y-axes. The y-axis always has a
numerical scale and is used to show values such as counts, frequencies or percentages. Intervals on the scale are marked by numbers and tick marks, indicating the major divisions, to the left of the y-axis. Like the x-axis, the y-axis usually has a label that provides details of the units of measurement. The label is often written vertically to follow the line of the y-axis but can instead be placed just above the top of the y-axis. In order to best highlight a trend in the data, it may be necessary to start the y-axis scale at a point other than zero. In such cases the starting value on the y-axis should be clearly labelled and the readers’ attention drawn to the non-zero start by breaking the y-axis just below the first value as shown in the example opposite.

Gridlines

Gridlines are the vertical and horizontal lines placed within the plot area to help read values from the graph. The gridlines should be subtle and not detract from the data. In the case of simple graphs it is not always necessary to include them. Gridlines are usually drawn at regular intervals based on the major divisions of the y-axis scale.

Title

All graphs should include a title that summarises what the graph shows. The title should identify what is being described (e.g. speeding offences detected by automatic cameras) and the units of measurements (e.g. percentages, total number, frequency). The title may be placed within the chart area, as in the example above, or above or below the chart.

Reference to source of information

If the graph you are presenting is based on data from another publication then you should acknowledge the source of the original data somewhere within the chart area or title. If however, the graph is based on data that you have collected yourself then there is usually no need to provide details of the information source since this is usually clear from the accompanying text.

Other design considerations

Use of colour and shading
In bar charts, histograms, and pie charts, shading and colour are often used to distinguish the areas representing different categories. The choice of which colour combinations and shading patterns to use is ultimately a personal matter but there are some general points that will help ensure your chart is easy to interpret.

If there is a natural order or sequence to the data you are plotting then the colours or shading patterns used to represent the categories should reflect this so that the colours of adjacent areas grade from dark tones to light tones or vice versa.

In the example below therefore, A and D show a better use of shading than B. The darkest colours should also be placed at the bottom of the column otherwise they dominate the bar and can give the impression that it is top heavy. For example compare A and C which show the same colour sequence but in reverse order. Similar optical effects can occur with the use of some shading patterns. For example the patterns used in E are too busy and make distinguishing the different areas of the bar difficult. In F, the use of vertical lines results in the apparent lengthening of the bottom area of the bar whilst the horizontal lines have the effect of shortening the upper area of the bar when in fact the two areas are equal in size.

Examples of different shading effects in histogram bars

3D effects

Many PC-based graphing packages offer the possibility of producing graphs and charts with a 3D effect. However, in general the use of 3D makes it much more difficult to interpret the data presented in chart or graph because the false depth and perspective that are added to the chart make reading and comparing values extremely difficult. This can be seen in the example given here where it is difficult to judge the heights of the columns and the use of perspective results in some bars covering parts of others.
Similar problems occur with the use of 3D in pie charts which make it difficult to judge the size of the slices. Even when they represent similar sized categories, those slices at the rear of the chart automatically appear smaller than those at the front due to the false perspective.

**Summary**

This guide has outlined the role of text and tables as formats for presenting numerical data and summarised the most common graph types and the types of data that they can be used to present. The guide has focussed on providing general design and presentation advice for making numerical data easy to read and interpret.

**Bar charts**

**Aims and Objectives**

This guide explains what bar charts are and outlines the different ways in which they can be used to present data. It also provides some design tips to ensure that when you use bar charts to present data they are clear and easy to interpret.

**What is a bar chart?**

Bar charts are a type of graph that are used to display and compare the number, frequency or other measure (e.g. mean) for different discrete categories of data. In the example below, which shows the percentage of the British population who attended different types of cultural events during 1999-2000, the types of event are the discrete categories of data.
Bar charts are one of the most commonly used types of graph because they are simple to create and very easy to interpret. They are also a flexible chart type and there are several variations of the standard bar chart including horizontal bar charts, grouped or component charts, and stacked bar charts.

The chart is constructed such that the lengths of the different bars are proportional to the size of the category they represent. The x-axis represents the different categories and so has no scale. In order to emphasise the fact that the categories are discrete, a gap is left between the bars on the x-axis. The y-axis does have a scale and this indicates the units of measurement.

**What types of data can be displayed using a bar chart?**

Bar charts are useful for displaying data that are classified into nominal or ordinal categories. Nominal data are categorised according to descriptive or qualitative information such as county of birth, or subject studied at university. Ordinal data are similar but the different categories can also be ranked, for example in a survey people may be asked to say whether they thought something was very poor, poor, fair, good or very good.

With nominal data, arranging the categories so that the bars grade sequentially from the largest category to the smallest category helps the reader to interpret the data. However, this is not appropriate for ordinal data because the categories already have an obvious sequence. Bar charts are also useful for displaying data that include categories with negative values, because it is possible to position the bars below and above the x-axis.
Different types of bar charts

Horizontal bar charts

Bar charts are normally drawn so that the bars are vertical which means that the taller the bar, the larger the category. However, it is also possible to draw bar charts so that the bars are horizontal which means that the longer the bar, the larger the category. This is a particularly effective way of presenting data when the different categories have long titles that would be difficult to include below a vertical bar, or when there are a large number of different categories and there is insufficient space to fit all the columns required for a vertical bar chart across the page.
Data sourced from Table 6.7.2 (Section 6.7: Dental Hygiene Behaviou) Adult Dental Health Survey, 1998, pp. 331

Note that in Excel a chart in which the bars are presented vertically is referred to as a column chart, whilst a chart with horizontal bars is called a bar chart.

**Grouped bar charts**

Grouped bar charts are a way of showing information about different sub-groups of the main categories. In the example below, a grouped bar chart is used to show the different schemes (sub-groups) by which different categories of household materials are recycled.

A separate bar represents each of the sub-groups (e.g. civic amenity sites) and these are usually coloured or shaded differently to distinguish between them. In such cases, a legend or key is usually provided to indicate what sub-group each of the shadings/colours represent. The legend can be placed in the plot area or may be located below the chart.

Grouped bar charts can be used to show several sub-groups of each category but care needs to be taken to ensure that the chart does not contain too much information making it complicated to read and interpret. Grouped bar charts can be drawn as both horizontal or vertical charts depending upon the nature of the data to be presented.
Stacked bar charts

Stacked bar charts are similar to grouped bar charts in that they are used to display information about the sub-groups that make up the different categories. In stacked bar charts the bars representing the sub-groups are placed on top of each other to make a single column, or side by side to make a single bar. The overall height or length of the bar shows the total size of the category whilst different colours or shadings are used to indicate the relative contribution of the different sub-groups.

Example of a stacked bar chart.

Stacked bar charts can also be used to show the percentage contribution different sub-groups contribute to each separate category. In this case the bars representing the individual categories are all the same size. The information could also be presented in a series of pie charts.
Histograms

Aims and Objectives

This guide describes what histograms are and the types of data that they can be used to present. It outlines some of the problems in drawing accurate histograms in Excel and suggests some ways of overcoming these.

What are histograms?

Histograms are a special form of bar chart where the data represent continuous rather than discrete categories. This means that in a histogram there are no gaps between the columns representing the different categories. In the example below, a histogram has been used to show the average height of children of different ages in 1837. A histogram is used because age is a continuous rather than a discrete category.

In a bar chart the length of the bar indicates the size of the category, but in a histogram it is the area of the bar that is proportional to the size of the category. This difference is due to the fact that in a histogram both the x-axis and y-axis have a scale, whereas in a bar chart only the y-axis has a scale.


Grouping data in a histogram

A continuous category, such as age, may have a large number of possible values and this could result in complex histogram with so many columns that it becomes difficult to interpret the information. For this reason the data in a histogram are often grouped to reduce the number of categories. For example, instead of drawing a bar for each individual age from 16 onwards, the data in the histogram below have been grouped into a series of continuous age ranges: 16-24, 25-34 etc.

However, when grouping the original data, it is important to remember that in histograms the size of the category is represented by the area of the bars and not their length. A common error when constructing histograms is to overlook this relationship and this can produce a distorted view of the data.

This usually occurs if the data have been grouped into uneven sized categories, for example if the age ranges were 0-10, 11-15, 16-21, each would represent a different number of years (10, 5, 6) and therefore the corresponding bars in the histogram would have to have different widths to maintain the relationship between area and category size.

In the example above, the bar representing the age range 16-24 is slightly narrower than that for the age ranges 24-34, 35-44 etc. because it includes only 9 years whereas the others include 10. In the same way the bar representing ages over 75 is broader than the other bars since it represents an open-ended category.

**Histograms and Excel**

The spreadsheet package Excel does not include histograms amongst its standard chart types. It is however, possible to draw basic histograms using Excel by selecting either the column or bar chart types. By default these chart types include a gap between the columns representing each category but this can be removed, in order that adjacent columns butt onto one another, resulting in the chart appearing as a histogram.

This is achieved in the following way:

1. Highlight any of the columns in the graph by clicking on it with the mouse;
2. Select **Format selected data series** from the main menu bar. This will bring up a dialogue box titled **Format data series**;
3. In the dialogue box select the **Options** heading;
4. Adjust the gap size setting to **0**.
Unfortunately, Excel does not include the ability to alter the width of the columns or bars drawn for each category, and therefore it is not suitable for drawing histograms for grouped data when the categories represent groups of different sizes.
Pie charts

Aims and Objectives

This guide explains what a pie chart is, outlines the types of data that can be presented using a pie chart and provides some design tips to ensure that when you use pie charts to present data they are clear and easy to interpret.

What is a pie chart?

A pie chart is a circular graph that shows the relative contribution that different categories contribute to an overall total. A wedge of the circle represents each category's contribution, such that the graph resembles a pie that has been cut into different sized slices. Every 1% contribution that a category contributes to the total corresponds to a slice with an angle of 3.6 degrees.

What data can be presented using a pie chart?

- Pie charts are a visual way of displaying data that might otherwise be given in a small table.
- Pie charts are useful for displaying data that are classified into nominal or ordinal categories. Nominal data are categorised according to descriptive or qualitative information such as county of birth or type of pet owned. Ordinal data are similar but the different categories can also be ranked, for example in a survey people may be asked to say whether they classed something as very poor, poor, fair, good, very good.
- Pie charts are generally used to show percentage or proportional data and usually the percentage represented by each category is provided next to the corresponding slice of pie.
Pie charts are good for displaying data for around 6 categories or fewer. When there are more categories it is difficult for the eye to distinguish between the relative sizes of the different sectors and so the chart becomes difficult to interpret.

**Design issues**

Pie charts provide a good visual representation of the data when the categories show some variation in size. When there are several similar-sized categories, a pie chart can end up looking cluttered and it may be difficult to interpret the data. In such cases consider whether a table would present the information more effectively.

It is usual for the different sectors of the pie chart to be arranged clockwise in order of magnitude. If there is a slice that does not contain a unique category of data but summarises several, for example “other types” or “other answers”, then even if it is not the smallest category it is usual to display it last in order that it does not detract from the named categories of interest.

It is helpful to colour or shade the different slices so that they grade from dark to light tones as you move from the first to the last slice.

**Comparing pie charts**

Two or more pie charts can be used to compare two sets of data where the categories are the same or similar but there is a change in another variable, such as time or age. In such cases it is helpful to maintain the same ordering and colouring of slices in the second pie chart as the first in order to facilitate comparison. This may mean that in the second pie chart the slices are no longer presented in order of magnitude. If the two charts represent different sized totals show this by making the areas of the pie chart proportional to the totals they each represent.
3D effects and exploding pie charts

Avoid design elements that detract from the message you are trying to convey such as 3D effects and exploding slices. These can produce optical effects that make it hard to compare different categories.

For example, in the 3D-exploding pie chart below, the slice representing Alton Towers appears quite a bit larger than that for Pleasureland although there is actually only a 1% difference between the two categories. Also, both slices for Chessington World and Legoland represent 10% but again they appear different sizes due to the 3D perspective. In this example separating (exploding) the different wedges of the pie chart also adds to the difficulty of interpreting the data and estimating the relative size of different sectors.

There are however occasions when the use of an exploding pie chart can enhance the presentation of the data. For example if you wish to highlight information in one category/wedge in particular, or when additional information is provided about a particular category as is shown in the example below.
Distribution of visitors to different types of tourist attractions in Britain, 1999

- Theme Parks: 38%
  - Blackpool Pleasure Beach (47%)
  - Alton Towers (17%)
  - Pleasureland, Southport (16%)
  - Chessington World of Adventures (10%)
  - Legoland, Windsor (10%)
- Museums & Galleries: 37%
- Historic Houses & Monuments: 16%
- Wildlife Parks & Zoos: 9%