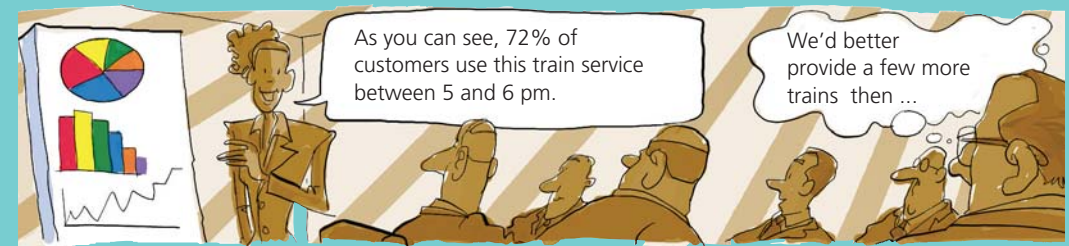


D2 Handling data

This unit will show you how to:

- ▶▶ Discuss how data relate to a problem.
- ▶▶ Identify possible sources of data, including primary and secondary sources.
- ▶▶ Find summary values that represent the raw data.
- ▶▶ Select, construct and modify suitable graphical representation.
- ▶▶ Identify key features present in the data.
- ▶▶ Have a basic understanding of correlation. Compare two or more distributions and make inferences.
- ▶▶ Communicate interpretations and results of a statistical enquiry.
- ▶▶ Solve increasingly demanding problems and evaluate solutions.
- ▶▶ Present a concise, reasoned argument, using symbols, diagrams, graphs and related explanatory text.



Displaying data can make it easier to understand.

Before you start

You should know how to ...

- 1 Draw bar charts and line graphs.
- 2 Calculate the mean, median, mode and range.

Check in

- 1 Draw a bar chart to represent this data:

Pet	Frequency
Cat	5
Dog	7
Rabbit	3

- 2 Find the mean, median, mode and range for this set of data:
3 5 4 8 8 7 1

D2.1 Planning data collection

This spread will show you how to:

- ▶ Discuss how data relate to a problem that can be addressed by statistical methods.
- ▶ Identify possible sources of primary and secondary data.
- ▶ Plan how to collect the data.

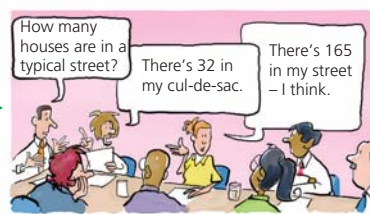
KEYWORDS

Statistical methods
Primary data
Secondary data

Benford is a company that produce house numbers. They are having a problem at their warehouse!



Some digits seem to occur more often than others. The company needs to do some research ...



First they need to collect some data.



They consider using primary data.



They finally opt for primary data.



They consider using secondary data.

Benford have planned how they are going to collect the data. They still need to work out how they are going to record it.

- ▶ When you discuss a problem that can be addressed by statistical methods there may be related questions you need to explore.
- ▶ You need to decide what data to collect and identify possible sources.
- ▶ You need to plan how you are going to collect and record the data.

Exercise D2.1

- 1 A survey of house numbers in five streets produced the following data.

Street	A	B	C	D	E
Even numbers	2 to 34	2 to 60	2 to 38	2 to 52	2 to 44
Odd numbers	1 to 33	1 to 45	1 to 55	1 to 51	1 to 59

Note:

Sometimes streets are curved so that there are more houses on one side than the other.

- a Do you think this data is typical? Give reasons for your answer.
 - b Using the data for street C complete a tally chart to show the distribution of all the digits that will be needed to number each house.
 - c Using the data for street C complete a tally chart to show the distribution of the **first** digit needed for each house.
 - d Compare the two tally charts and comment on your results.
 - e Do you think that the other streets will show similar results? Give a reason for your answer.
- 2
- a Choose one of streets A, B, D or E from the table in question 1 and draw two tally charts to show the distribution of:
 - i all digits
 - ii the first digits of house numbers.
 - b Compare the two tally charts and comment on your results.
 - c Compare your observations on house number digits for this street with your observations for street C. What do you notice?
- 3 A hospital radio station can only broadcast 15 minutes of sport each Saturday. They can do this either in:
- i three 5-minute slots
 - ii a 10-minute slot and a 5-minute slot, or
 - iii one 15-minute slot.
- The sporting news they transmit is from the game played by the local football team.
- a If they want to be on air when a goal is scored, which of the three options should they choose?
 - b Identify what data you would need to collect and where you might collect this data.

D2.2 Displaying data

This spread will show you how to:

- ▶▶ Gather data from specified secondary sources.
- ▶▶ Construct pie charts for categorical data.
- ▶▶ Draw stem-and-leaf diagrams.

KEYWORDS

Pie chart
Stem-and-leaf diagram

Frank surveyed the houses in his street and recorded the first digit of each house number. He collated his results in a table.

First digit	1	2	3	4	5	6	7	8	9
Frequency	12	11	11	9	6	2	1	1	1

Frank constructed a pie chart to represent the data.

- 1 First find the **frequency total**.
In the table, this is 54.
- 2 Then find each angle as a **fraction of 360°**.
So for first digits 1, 2, ...
 $\frac{12}{54} \times 360^\circ = 80^\circ$, $\frac{11}{54} \times 360^\circ = \dots$

Frank found that he would need more of the lower digits to number each house in his street.

Benford needed to know if Frank's street was typical.

They collected data on the number of houses there were in each of 27 streets.

52 62 44 47 59 72 109 64 78
52 84 70 68 66 48 60 52 105
56 80 81 66 58 58 74 50 106

They drew a **stem-and-leaf diagram** to represent the data.

To draw a stem-and-leaf diagram:

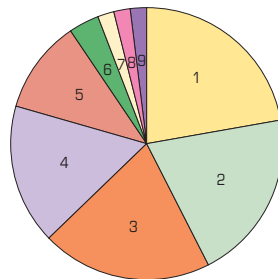
- 1 Look at the data and decide what the **stem** should be. Most of the data are tens and units numbers.
- 2 The units are the leaves added to the stem. Organise the leaves in numerical order.
- 3 Write a key.

10	5 6 9
9	
8	0 1 4
7	0 2 4 8
6	0 2 4 6 6 8
5	0 2 2 2 6 8 8 9
4	4 7 8

Key: $5 \mid 2$ means 52

The diagram suggests that most streets have 50 to 70 houses.

- ▶ A stem-and-leaf diagram shows all the original data and gives you an overall picture of the trend.



Exercise D2.2

- 1 Here is part of the data from Exercise D2.1 on page 149, relating to house numbers on a street.

Street	Even numbers	Odd numbers
C	2 to 38	1 to 55

Use the data for street C to draw pie charts to show:

- a the distribution of all the digits
- b the distribution of the first digit of each house number.

- 2 The doors of the houses in street C are painted in the following colours:

Colour	Blue	Green	Red	White	Yellow
Frequency	12	8	6	18	3

Draw a pie chart to represent these data.

- 3 Gemma surveyed 27 streets in her area, counting the number of houses in each street. They were

47 54 68 39 32 57 54 44 60
56 53 48 62 38 58 38 46 58
36 48 62 54 64 66 42 44 50

Draw a stem-and-leaf diagram to represent these data.

- 4 The data shows the total points scored by the teams in the premier league division for the 2001/2002 season.

87 80 77 71 66 64 53 50 50 46 45 45 44 44 43 40 40 36 30 28

Draw a stem-and-leaf diagram to represent these data.

- 5 Draw a back to back stem-and-leaf diagram to show the overall points scored by Division One and Division Two in the Nationwide league for the 2001/2002 season.

Division One

99 89 86 77 76 75 75 72 67 66 66 64
60 59 55 54 53 51 50 50 49 49 48 26

Division Two

90 84 83 83 80 78 73 71 70 64 64 63
59 58 57 56 55 52 50 49 44 44 43 34

A **back-to-back** stem-and-leaf diagram looks like this:

Boys		Girls	
9 5 2	8	0 5 9	
9 9 3 3 1	7	1 2 4 6 6 7	
9 5 4 4 4 2	6	2 3 3 8 9	
8 3 2 2 1	5	0 4 5 5 6	
7 5 0	4	2 3 7	
8 4 1	3	0 2	

You can use it to **compare** two data sets.

D2.3 Interpreting data

This spread will show you how to:

- ▶ Interpret tables, graphs and diagrams.
- ▶ Draw inferences to support or cast doubt on initial conjectures.

KEYWORDS

Interpret Infer
Mean Median
Mode Range
Stem-and-leaf diagram

Benford want to analyse the results of their house number survey.
They can calculate statistics from their stem-and-leaf diagram.

Median: the 27 pieces of data are arranged in order so just count to find the middle (14th) value:
The median is 64 houses.

Mode: Just scan the rows to find the most frequent number:
The mode is 52 houses.

Mean: Add up all the values (1821) and divide by the number of values (27): $1821 \div 27 = 67.4$
The mean is 67.4 houses.

Range: Subtract the smallest number (44) from the largest (109):
The range is 65.

Benford need to interpret the statistics that they have calculated.

The part of the stem with most leaves is 50, and the mode supports this.

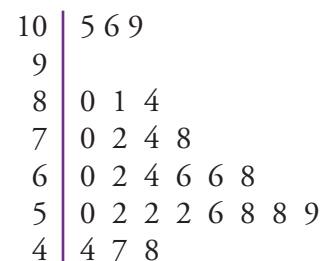
5 | 0 2 2 2 6 8 8 9

However, the median and mean are both in the sixties, suggesting a slightly higher average value for house numbers.

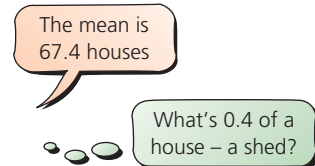
The shape of the diagram shows that most of the data is in the fifties and sixties. All three averages are within these values.

The range is 65, suggesting that there is a lot of variation in the number of houses in each street.

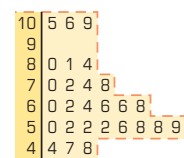
Three streets contained more than 100 houses.
Benford did not have enough data to decide whether streets with over 100 houses were **extreme** values, or were likely to be **typical**.



Key:
5 | 2 means 52



The mean is not necessarily a data value.



Exercise D2.3

1 The masses, in milligrams, of a sample of pebbles are:

169 178 164 182 194 204 186 192 201
182 164 175 179 173 182 172 180 171
168 172 169 185 182 200 166 198 170

- a Draw a stem-and-leaf diagram to represent these data.
- b Find the range, mean, median and mode of these data.
- c Comment on the shape of the diagram and your results in b.



2 The winning times, in hours, minutes and seconds, of the New York marathon from 1976 to 1992 for men are given in this set of data.

2 : 10 : 10 2 : 11 : 29 2 : 12 : 12 2 : 11 : 42 2 : 09 : 41
2 : 08 : 13 2 : 09 : 29 2 : 08 : 59 2 : 14 : 53 2 : 11 : 34
2 : 11 : 06 2 : 11 : 01 2 : 08 : 20 2 : 08 : 01 2 : 12 : 39
2 : 09 : 24 2 : 09 : 29

- a Draw a stem-and-leaf diagram to represent these data.
- b Find the range, mean, median and mode of these data.
- c Comment on the shape of the diagram and your results in b.

3 The data shows the total points scored by the football teams in Division 1 and Division 2 at the end of the 1980/1981 season.

Division One
60 56 53 52 51 50 50 48 44 43 42
39 38 37 36 35 35 35 35 33 32 19

Division Two
66 53 50 50 48 45 45 43 43 42 42
40 40 39 39 38 38 38 36 36 30 23

- a Draw a back-to-back stem-and-leaf diagram to represent these data.
- b Find the mean, median and range of these data sets.
- c Comment on the shape of the diagram and your results in b.



D2.4 Scatter graphs

This spread will show you how to:

- ▶ Construct scatter graphs.
- ▶ Identify scatter graphs.
- ▶ Develop basic understanding of correlation.

KEYWORDS

Scatter graph
Variable
Correlation
Trend

The table gives data about the number of vehicles wheel-clamped and towed away by London police.

All figures are given in thousands to the nearest thousand.

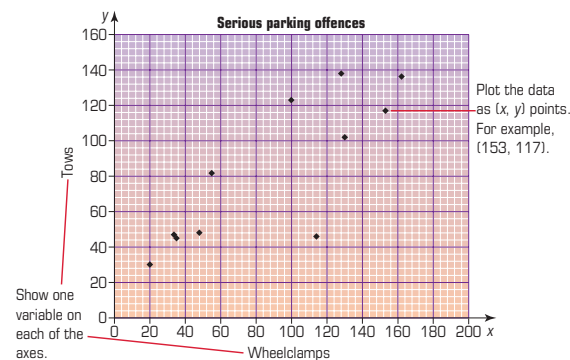
Year	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
Wheel-clamps	44	35	34	114	130	153	163	128	100	55	21
Tows	48	45	47	46	102	117	136	138	123	81	31

Each column in the table shows a pair of variables, linked by year, for example (34, 47) in 1986.

1986
34
47

▶ You can show data containing two linked variables on a scatter graph.

The graph shows that generally as the number of wheel-clamps increased, so did the number of vehicles towed away. The **trend** is generally increasing. There is a relationship, or **correlation**, between wheel-clamps and tows.



▶ You use a scatter graph to show if there is any **correlation** between two variables.

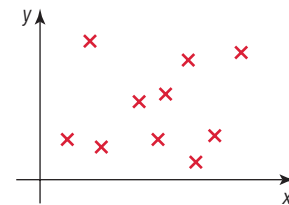
When the trend is increasing, there is **positive correlation** between the variables.



When the trend is decreasing, there is **negative correlation** between the variables.



When there is no apparent trend, there is **no correlation** between the variables.



Exercise D2.4

- 1 The data shows the goals scored for and the goals scored against each team in the premier division in 2001/2002.

For	79	67	87	74	53	66	48	46	49	55	46	35	36	38	45	44	29	41	33	30
Against	36	30	45	52	37	38	57	47	53	51	54	47	44	49	57	62	51	64	63	64

- a Draw a scatter graph to display this data.
- b Comment on any trend shown by your graph.

- 2 The results of two tests X and Y, taken by 12 students are given in the table.

Test X	16	12	18	15	6	14	18	4	16	8	13	16
Test Y	14	13	15	15	8	12	19	6	17	5	11	19

- a Draw a scatter diagram to show these results.
- b What can you say about the performance of the students in each of the tests X and Y?

- 3 Josh was training hard to improve his 100 m sprint times. He was timed, to the nearest tenth of a second, each Sunday morning over a period of ten weeks.

Week	1	2	3	4	5	6	7	8	9	10
Time (s)	34.6	33.9	33.0	32.2	31.5	31.8	31.0	30.6	30.2	29.6

- a Draw a scatter diagram to show these results.
- b Comment on the correlation shown by your graph.
- c Use your graph to comment on what Josh's 100 m time might be after 50 weeks training. Is your answer sensible?

- 4 The table shows how long, in seconds, it took eight people to spell their name backwards.

Name	Hugh	Helen	Harry	Hannah	Hamish	Horatio	Heather	Henrietta
Number of letters in name	4	5	5	6	6	7	7	9
Time in (seconds)	3.3	5.4	4.9	3.6	6.2	7.8	7.6	9.6

- a Plot a scatter diagram of these data.
- b One of the points plotted does not seem to fit with the rest of the data. Circle this point and suggest a reason why it may not follow the trend.

D2.5 Comparing data

This spread will show you how to:

- ▶▶ Select statistics most appropriate to the problem.
- ▶▶ Compare two or more distributions using appropriate statistics.
- ▶▶ Analyse the effect that minimal changes in data can have on graphs and statistical measures.

KEYWORDS

Mean Range
Median Variation

The table gives data on vehicles wheel-clamped and towed away by London police, as shown on page 154. All figures are given in thousands, to the nearest thousand.

Year	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
Wheel-clamps	44	35	34	114	130	153	163	128	100	55	21
Tows	48	45	47	46	102	117	136	138	123	81	31

You can use statistics to compare the data.

	Mean	Median	Range
Wheelclamps	88.8	100	142
Tows	83.1	81	107

The two means and medians suggest that there are more vehicles wheel-clamped than towed away.

The two ranges suggest that there is greater variation in numbers of vehicles clamped than towed away.

When people deal with large amounts of data, sometimes data can get lost or distorted.

- ▶ Minimal changes in data can affect graphs and statistical measures.

If data had been lost for 1987 to 1990, then the calculated statistics would be:

	Mean	Median	Range
Wheelclamps	59.6	44	107
Tows	73.3	48	107

The median number of wheel-clamps is more typical of the data.

The mean is distorted by the large values in 1991 and 1992.

These statistics suggest that on average more vehicles are towed away than clamped, and that there is no difference in variation.

- ▶ The average you choose for a data set is the one that should best represent all the data.

Statistics summarise the original data, or **raw data**.

When you calculate statistics, you lose some of the information in the raw data.

- ▶ Statistics can help compare data, but the original values are also important.

Exercise D2.5

- 1 For each of the following sets of data write down, with reasons:
 - i which average, mean, median or mode, is the most appropriate to find, and
 - ii find this average.
 - a 14.6 19.3 12.0 15.7 31.7
 - b 2003 2005 2008 2011 2006 2003 2008 2003
 - c 101 102 102 102 108 108 108 109
 - d 49 44 44 47 38 36 44 44 49 44 43
- 2
 - a Find the mean, median, mode and range of this set of data.
7 7 24 5 2
 - b One number is added to the data set. What could that number be if
 - i the mode remains the same
 - ii the median remains the same
 - iii the mean remains the same
 - iv the mean is increased by 2?
 - c One number is removed from the data set. What could that number be if
 - i the range and median remain unchanged
 - ii the mean is increased by 1?
- 3
 - a Calculate the mean, median, mode and range of each of these two data sets.
Set A: 0, 99, 99, 100, 100, 100, 100, 100, 101, 101, 200
Set B: 0, 0, 99, 99, 100, 100, 100, 101, 101, 200, 200
 - b Compare your results. What do the statistics suggest about the two data sets?
- 4 Find two sets of numbers that have the same mean, median, mode and range. (Do not use the data in question 3.)
- 5 Find data sets with the following properties
 - a Set A: Whatever number you remove, the mode remains the same.
 - b Set B: When you remove one number, the mean stays the same.
 - c Set C: When you remove one number, the mean is half the original value.
 - d Set D: When you remove a number, the mean and range remain the same.

- 6 The table shows grouped data of heights of a sample of Year 8 pupils.

Height, h cm	$135 \leq h < 145$	$145 \leq h < 150$	$150 \leq h < 155$	$155 \leq h < 170$
Frequency	2	4	6	3

Find two different sets of raw data that:

- a could be grouped in this frequency table
- b have the same mean, median and range.

D2.6 Statistical reports

This spread will show you how to:

- ▶ Communicate interpretations and results of a statistical enquiry.
- ▶ Select tables, graphs and diagrams to support findings.

KEYWORDS

Line graph
Bar chart
Statistical enquiry

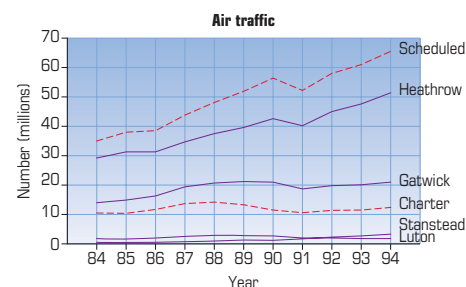
Sharon conducted a survey into how air traffic had changed from 1984 to 1994.

The table shows the numbers of air terminal passengers in millions.

	Heathrow Airport	Gatwick Airport	Luton Airport	Stanstead Airport	Scheduled flight	Charter flight
1984	29.2	14.0	1.8	0.5	35.0	10.5
1985	31.3	14.9	1.6	0.5	38.0	10.4
1986	31.3	16.3	2.0	0.5	38.5	11.7
1987	34.7	19.4	2.6	0.7	43.8	13.7
1988	37.5	20.7	2.8	1.0	48.1	14.2
1989	39.6	21.2	2.8	1.3	51.9	13.3
1990	42.6	21.0	2.7	1.2	56.4	11.5
1991	40.2	18.7	2.0	1.7	52.2	10.6
1992	45.0	19.8	2.0	2.3	58.0	11.4
1993	47.6	20.1	1.8	2.7	61.0	11.5
1994	51.4	21.0	1.8	3.3	65.5	12.4

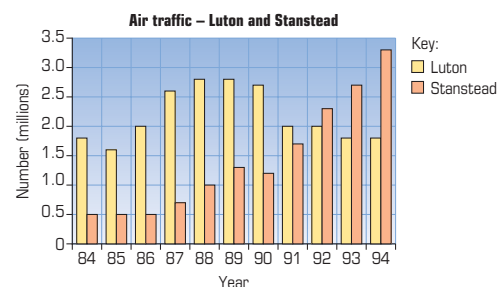
Sharon used statistics from the Department of Transport, which is secondary data. You should always state where the data has come from.

Sharon drew a line graph to show changes in passenger numbers at the airports.



- ▶ Generally numbers were increasing.
- ▶ There was a drop in numbers in 1991 (an effect of the Gulf War).
- ▶ Stanstead had a rapid increase in passenger numbers after 1991 (a new terminal opened).

Sharon drew a **multiple bar chart** to show passenger numbers at Luton and Stanstead.



- ▶ The bar chart shows clearly how passenger numbers at Stanstead had overtaken passenger numbers at Luton.

In a statistical report you should

- ▶ Explain how the data was collected
- ▶ Analyse the data
- ▶ Use graphs to highlight key points
- ▶ Look to see if there are other related questions to explore.

Exercise D2.6

- 1
 - a Use the air traffic data on the opposite page to draw a line graph showing how scheduled flights and charter flights changed over the 11-year period.
 - b Comment on the changes in scheduled and charter flights from 1984–1994.
- 2 The numbers of people injured in road traffic accidents in a London borough over a three-year period sorted into male and female drivers are given in the table.

	Pedestrians	Car drivers	Car passengers	Pedal cyclists	Motorcyclists	Bus passengers	Goods vehicles	Other
Male Drivers	400	1200	550	250	550	200	100	100
Female Drivers	200	1000	325	50	50	10	20	20

- a Draw a multiple bar graph to analyse these data.
 - b Write a short report on the differences between the vehicle types that are involved in accidents by men and women.
- 3 The table shows the number of road casualties in a London borough by month of the year and by age.

Month	Under 15	16–39 years	40–59 years	Over 60
January	35	130	120	50
February	55	180	150	35
March	65	135	135	60
April	45	115	125	45
May	60	150	130	50
June	45	135	115	45
July	60	130	120	40
August	60	140	120	50
September	70	160	140	45
October	60	160	155	60
November	50	180	170	70
December	30	180	130	65

- a Draw a diagram to represent these data.
- b Why do you think that the number of casualties aged under 15 gets lower through the Autumn while for other age groups it gets higher?
- c Write a short report about the monthly variations in the number of road casualties for the different age groups. Suggest reasons for the monthly variations in the number of road casualties.

D2 Summary

You should know how to ...

- 1 Communicate interpretations and results of a statistical enquiry using selected tables, graphs and diagrams in support.
- 2 Present a concise, reasoned argument, using symbols, diagrams, graphs and related explanatory text.

Check out

- 1 The heights of a class of girls are measured to the nearest cm.

160 154 174 151 161 168 166 161 164 166
154 175 160 168 167 165 155 171 167 172
163 169 169 154 153 168 165 168 155

- a Draw a stem and leaf diagram to represent these data.

You may need to expand your stem :

17
17
16 and so on
and split the units digits 0–4 and 5–9 on the stem.

- b Find the mean, median and range of these data.
- c Comment on the shape of your diagram and your results in (b).

- 2 The Tables show the number of tests taken by students at two driving schools before they passed.

Number of tests	1	2	3	4	5
Frequency (school x)	28	12	5	3	2

Number of tests	1	2	3	4	5
Frequency (school y)	45	30	21	4	0

Calculate statistics and draw diagrams to comment on the statement:

Students in school x did better than students in school y.