Have you ever wondered why VCE maths series are filled with questions that will never be asked on the exam?

_Nelson VCE Maths_ is the only VCE maths series where every question is exam practice.

**Student books**
- have a unique structure based on actual exam questions in a learning context
- use student performance data on exams to grade questions
- integrate continuous revision
- eliminate the need to buy a separate book of past exam questions

_Nelson VCE Maths_ is the only VCE maths series where every support product is FREE*.

**NelsonNetBooks**
- are web-based ebooks accompanying each printed textbook
- can be customised with voice recordings, highlighting, drawings, and links to websites
- allow fast easy-to-use navigation

**NelsonNet student website**
- contains worksheets that are write-in enabled PDFs and prior learning exercises

**NelsonNet teacher website***
- provides worked solutions to every question, for teachers to download and distribute to students
- includes curriculum grids, ExamView chapter quizzes and chapter PDFs of the student book

**ExamView***
- is exam and assessment software that can be downloaded from the NelsonNet teacher website
- includes a question bank of multiple choice and written response questions with answers
- allows you to quickly generate your own online or paper tests and practice exams
- allows you to make changes and add your own questions

*Please note that complimentary access to NelsonNet and the NelsonNetBook is only available to teachers who use the accompanying student textbook as a core educational resource in their classroom. Contact your sales representative for information about access codes and conditions.

And for the non-exam based VCE maths subject: _Nelson VCE Foundation Mathematics Units 1 and 2_
- written by Sue Thomson and Judy Binns, experts on teaching and writing for senior applied mathematics courses
- engaging students with interesting and relevant applications of mathematics, with an emphasis on numeracy and literacy
- featuring optional extension topics for students who wish to advance to the Further Mathematics pathway
Contents

About this series iii
About this book vi
About the author viii

FUNDAMENTALS

Computation and practical arithmetic 1
1.1 Scientific notation and orders of magnitude 2
Exam prep 1.1 4
Exam practice 1.1 5
1.2 Rounding and significant figures 7
Exam prep 1.2 10
Exam practice 1.2 10
1.3 Log scales 12
Exam prep 1.3 14
Exam practice 1.3 15
1.4 Percentages 16
Exam prep 1.4 25
Exam practice 1.4 26
1.5 Rates 28
Exam prep 1.5 31
Exam practice 1.5 31
1.6 The unitary method 33
Exam prep 1.6 35
Exam practice 1.6 36
Summary 38

Algebra and structure using CAS 41
2.1 The four operations in algebra 42
Exam prep 2.1 46
Exam practice 2.1 47
2.2 Expanding and simplifying algebraic expressions 48
Exam prep 2.2 50
Exam practice 2.2 51

2.3 Substituting 52
Exam prep 2.3 55
Exam practice 2.3 55
2.4 Solving equations 57
Exam prep 2.4 65
Exam practice 2.4 66
2.5 Solving simultaneous equations 69
Exam prep 2.5 72
Exam practice 2.5 72
Summary 74

Linear equations, graphs and models 77
3.1 Linear equations and graphs 78
Exam prep 3.1 83
Exam practice 3.1 84
3.2 The slope of a straight line 86
Exam prep 3.2 96
Exam practice 3.2 98
3.3 Linear models 100
Exam prep 3.3 104
Exam practice 3.3 105
3.4 Solving simultaneous equations graphically 107
Exam prep 3.4 113
Exam practice 3.4 115
3.5 Piecewise linear graphs 118
Exam prep 3.5 124
Exam practice 3.5 126
Summary 130
Exam practice

Are you under pressure to rush through course content to leave time to do sufficient exam practice at the end of the year?

*Nelson VCE Maths* eases the stress by integrating past exam and practice exam questions throughout the year.

Exam preparation

Do your students struggle with the maths workload and complain you are ‘going too fast’?

*Nelson VCE Maths* filters out the unnecessary busy work for students that usually fills VCE mathematics textbooks and targets the requirements of the exams, so students can concentrate on what matters.

Targeted theory

Do your students find it difficult to make sense of overly detailed and convoluted theory in textbooks?

*Nelson VCE Maths* provides clear and concise theory, so your students don’t get bogged down with wordy explanations.
CAS focus

Do your students struggle to find relevant instructions on how to use their CAS at the point of need?

*Nelson VCE Maths* has CAS instructions clearly labelled and located where students need them.

Detailed solutions

Do your students sometimes need more than just answers at the back of the book?

For selected questions, *Nelson VCE Maths* also provides worked solutions and identification of common errors from the VCAA examination reports (worked solutions to all questions can be downloaded from the NelsonNet teacher website).
About the author

Series editor and lead author Dirk Strasser taught VCE Mathematics for 12 years and has been a senior educational publisher for 18 years. He has published and co-written eight best-selling mathematics series, including Heinemann Maths Zone 7–10 and Pearson Mathematics 7–10.
CHAPTER 4

DATA DISTRIBUTIONS

4.1 Data and variables
- Data and variables
- Types of data and variables
- The range, median and mode

4.2 Tables and charts
- Frequency tables
- Bar charts

4.3 Histograms
- Continuous data
- Grouped discrete data
- Ungrouped discrete data
- Symmetric and skewed distributions
- Outliers
- Medians
- Range
- Using CAS: Constructing histograms for ungrouped data
- Using CAS: Constructing histograms for grouped data

4.4 Boxplots
- Quartiles
- The five-number summary
- Using CAS: Five-number summary
- IQR and outliers
- Boxplots
- Using CAS: Boxplots
- Comparing boxplots and histograms

4.5 Dot plots and stem plots
- Dot plots
- Stem plots

4.6 Back-to-back stem plots and parallel boxplots
- Back-to-back stem plots
- Parallel boxplots
- Using CAS: Constructing parallel boxplots
- Which display do you use?

4.7 The mean and standard deviation
- The mean
- The median vs the mean
- The standard deviation
- Using CAS: Finding the mean and standard deviation

Summary
Data and variables

We live in a world where data is central to our lives. Governments, businesses, scientists, sporting organisations, researchers and schools all make decisions by relying on data. Does smoking cause lung cancer? Let's look at the data. Is global warming occurring and is it caused by human activity? Let's look at the data. Is a country in danger of falling into recession? Let's look at the data. Without data, the big questions can't be answered.

The data on its own, however, is only part of the answer to the big questions. It is important to get good quality data, but it is also important to organise it in a way that allows us to make sense of it. We need to be able to put it in a form that allows us to interpret it. Without organisation, data is simply a meaningless collection of information.

Here is some data:

Responses by a group of ten students
- 1, 1, 1, 2, 3, 4, 2, 1, 2
- white, white, grey, white, red, blue, red, red, silver
- 4, 3, 5, 2, 1, 4, 6, 3, 4, 4
- 8, 3, 6, 1, 0, 0, 12, 7, 8, 0
- 3149, 3149, 3148, 3149, 3149, 3149, 3166, 3147, 3149, 3148

There's not much we can do with this. What are missing are the variables, the things about which we are recording information:
Reponses by a group of ten students

<table>
<thead>
<tr>
<th>Variables</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating given to latest Star Wars movie where 1 = great, 2 = okay, 3 = awful, 4 = didn't see it</td>
<td>1, 1, 1, 2, 3, 3, 4, 2, 1, 2</td>
</tr>
<tr>
<td>Colour of the first car seen this morning</td>
<td>white, white, grey, white, red, blue, red, red, silver</td>
</tr>
<tr>
<td>Number of mobile phones in home</td>
<td>4, 3, 5, 2, 1, 4, 6, 3, 4, 4</td>
</tr>
<tr>
<td>Age of oldest living pet (in years)</td>
<td>8, 3, 6, 1, 0, 0, 12, 7, 8, 0</td>
</tr>
<tr>
<td>Postcode of home address</td>
<td>3149, 3149, 3148, 3149, 3149, 3149, 3166, 3147, 3149, 3148</td>
</tr>
</tbody>
</table>

How you organise this data depends on what sort of variables are involved.

**Types of data and variables**

There are two types of variables.

**Categorical variables**

Categorical data is data that can be sorted into categories or groups. If there are no numbers involved, then the data is clearly categorical. If there are numbers involved, you need to ask the question ‘Is the number simply being used like a name?’ If it is, then the data is categorical.

Categorical data can be divided into two sub-categories: nominal data and ordinal data.

<table>
<thead>
<tr>
<th>Categorical data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal data</td>
</tr>
<tr>
<td>This is categorical data that doesn’t have a natural order, even when numbers are involved.</td>
</tr>
</tbody>
</table>
| • Colour of first car seen this morning  
• Brand of mobile phone you use  
• Favourite vegetable  
• Postcode of home address  
• Number on basketball player uniform | • Ratings given to the latest Star Wars movie (1 = great, 2 = okay, 3 = awful, 4 = didn’t see it)  
• House numbers  
• Shoe sizes  
• Grades in a test (A, B, C, D, E)  
• Clothing sizes (small, medium, large, extra large) |

To decide whether categorical data is nominal or ordinal data, ask the question ‘Does it make sense to order the numbers?’ If it does, the data is ordinal. If it doesn’t, it’s nominal.

**Numerical variables**

Numerical data is data that can be either counted or measured to an increasing level of accuracy. The questions to ask are ‘Does this involve counting something?’ and ‘Could I measure this more accurately if I wanted to?’ If the answer to either of these questions is yes, then the data is numerical.

Numerical data can be divided into two sub-categories: discrete data and continuous data.
Numerical data

<table>
<thead>
<tr>
<th>Discrete data</th>
<th>Continuous data</th>
</tr>
</thead>
<tbody>
<tr>
<td>This is numerical data that involves counting something.</td>
<td>This is numerical data that could be measured to an increasing level of accuracy.</td>
</tr>
<tr>
<td>• The number of mobile phones in your home</td>
<td>• The age of oldest living pet</td>
</tr>
<tr>
<td>• The number of people who like Vegemite</td>
<td>• The time taken to run 100 metres</td>
</tr>
<tr>
<td>• The amount of money spent in a shop</td>
<td>• The height of players in a basketball team</td>
</tr>
<tr>
<td>• To decide whether numerical data is discrete or continuous, ask the two questions ‘Does it involve counting something?’ and ‘Could I measure this more accurately if I wanted to?’</td>
<td></td>
</tr>
</tbody>
</table>

Exam hack

Don’t fall into the trap of thinking that because numbers are involved, the data must be numerical!

Worked example 1

State whether the following variables are nominal categorical, ordinal categorical, discrete numerical, or continuous numerical.

a. Gender
b. Number of children
c. Finishing position in a 100 metre race
d. Height

**Working**

a. Gender is a category (male or female) with the order not being important. Nominal categorical

b. Number of children involves counting. Discrete numerical

b. Categories can be numbers. Each finishing position in a race is a category (1st place, 2nd place, etc. or 1, 2, 3, etc.) and the order is important. Ordinal categorical

d. Height is a measurable numerical value and can be measured to an increasing level of accuracy. Continuous numerical
The range, median and mode

We can identify some special features of a data set of numerical variables.

Range
- The range is a measure of the spread of the data.
- Range = largest value – smallest value

Median
- The median is a measure of the centre of the data.
- The median is the middle value when the data is ordered from smallest to largest.
- When there are two middle values, we add them and divide by 2 to find the median.

Mode
- The mode is the most frequently occurring value in the data set.
- Nominal data has a mode, even though it doesn’t have a range or a median.
- There can be more than one mode, but if every data value occurs the same number of times, then there is no mode.

Worked example 2

For the following set of data, find
a the median  
b the mode  
c the range.
30 28 26 31 34 35 32 33 21 25 28 32 32 35

Working

a 1 Order the data from the smallest to the largest.
2 Locate the position of the median.
3 Locate the 7th and 8th data values and state the median.

b State the number with the highest frequency.

c Identify the smallest and largest value.

Exam hack

The median does not necessarily have to be one of the data values.
Data and variables

**Prep 1**

**WORKED EXAMPLE 1**

State whether the following variables are nominal categorical, ordinal categorical, discrete numerical, or continuous numerical.

Recording information on

- **a** the number of books in the house
- **b** the distance to the nearest train station from your home
- **c** your middle initial
- **d** the amount of time spent sleeping last night
- **e** the speed of an aeroplane just after take-off
- **f** the candidate a student is intending to vote for in the next election for school captain
- **g** numbers on the Matildas’ uniforms
- **h** opinion of chocolate on a scale of 1 to 5, where 1 is hate and 5 is love
- **i** salary in dollars
- **j** salary classified as high, medium or low
- **k** the number of people watching the Australian Tennis Open
- **l** the different titles of surfing magazines
- **m** the foot length of swimmers in a squad
- **n** the finishing places of 11 horses in a race
- **o** the number of levels in an office building
- **p** the state of birth of each person born in Australia

**Prep 2**

**WORKED EXAMPLE 2**

Find the range, median and mode of each of the following.

- **a** The ages of people (in years) working in a restaurant:
  
  42  21  18  35  19  18  27

- **b** The number of people buying coffees at a café on 6 consecutive mornings:
  
  39  33  35  38  56  41

- **c** The lowest maximum monthly temperature (°C) at a ski resort from April to November:
  
  3.0  -2.0  -1.6  -2.0  -1.8  -2.8  0.8  2.1
Data and variables

**Question 1**
What is the correct data classification for a person’s religion?

A. Ordinal categorical  
B. Discrete numerical  
C. Nominal categorical  
D. Continuous numerical  
E. None of the above

**Question 2**
What is the correct data classification for a hotel star rating?

A. Ordinal categorical  
B. Discrete numerical  
C. Nominal categorical  
D. Continuous numerical  
E. None of the above

**Question 3**
The marks for a spelling test of ten words were recorded as follows:
8 6 8 4 5 6 8 5 7 4 7 8 6 8 9
The median and mode respectively are

A. 6 and 8  
B. 7 and 6.6  
C. 8 and 7  
D. 7 and 8  
E. 7 and 7

**Question 4**
A survey was completed that collected the information about the heights of students within a cohort of Year 11 students. The type of data collected is best described as

A. nominal categorical data.  
B. ordinal numerical data.  
C. ordinal categorical data.  
D. discrete numerical data.  
E. continuous numerical data.

Use the following information to answer Questions 5 & 6.

A survey was conducted about the colour of family cars, with the following results:
grey grey white grey white red white blue red white red silver

**Question 5**
What is the mode of this data?

A. white  
B. grey  
C. red  
D. blue  
E. there is no mode
Question 6
What is the median of this data?
A white           B grey           C red
D blue            E there is no median

Question 7
The variables
region (city, urban, rural)
population density (number of people per square kilometre)
A are both categorical.
B are both numerical.
C are categorical and numerical respectively.
D are numerical and categorical respectively.
E are neither categorical nor numerical.

Use the following information to answer Questions 8 & 9.
The percentage investment returns of seven superannuation funds for the year 2002 were
-4.6%  -4.7%  2.9%  0.3%  -5.5%  -4.4%  -1.1%

Question 8
The median investment return is
A -4.7%  B -4.6%  C -4.5%  D -4.4%  E 0.3%

Question 9
The range of investment returns is
A 2.6%  B 3.5%  C 4.0%  D 5.5%  E 8.4%
Researchers conducted a survey of 403 school leavers who had recently entered the workforce. The aim was to determine whether the type of work they undertook was gender-related. Work type was classified as ‘trade’, ‘clerical’, ‘manual’ or ‘professional’.

In this survey, the variables

work type (trade, clerical, manual or professional)

and

gender (male or female)

are

A both categorical variables.
B both numerical variables.
C categorical and numerical variables respectively.
D numerical and categorical variables respectively.
E neither categorical nor numerical variables.

[VCAA 2002 1CQ1]

The level of water usage of 250 houses was rated in a survey as low, medium or high, and the size of the houses as small, standard or large.

The variables, level of water usage and size of house, as recorded in this survey, are

A both numerical variables.
B both categorical variables.
C neither numerical nor categorical variables.
D numerical and categorical variables respectively.
E categorical and numerical variables respectively.

[VCAA 2003 1CQ7]
Tables and charts

When you are dealing with a large number of data values, to see patterns or draw conclusions you need to organise and display the data in a manageable form. When choosing a display for your data, you must decide which one best shows what you wish to communicate.

**Frequency tables**

*Frequency tables* can be used to display both categorical and numerical data. The data values are listed in one column and the corresponding frequencies are displayed in a frequency column.

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Tally</th>
<th>Frequency ($f$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Response 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Response 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Response 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Set up a frequency table for a coffee-shop owner who wants to record the first 20 types of hot drinks that he sells one morning, where the drinks were cappuccino, latte, tea, latte, latte, cappuccino, tea, latte, latte, cappuccino, cappuccino, latte, tea, latte, latte, cappuccino, cappuccino.

**Working**

Set up a three-column table. Tally and record the frequency.

<table>
<thead>
<tr>
<th>Hot drink</th>
<th>Tally</th>
<th>Frequency ($f$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cappuccino</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latte</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tea</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>20</td>
</tr>
</tbody>
</table>

Always add up the frequency column to make sure that you have included all the data.
When data is discrete with a large number of values or is continuous, it is not practical to list each individual number in a frequency table. For example, if you have discrete values ranging between 0 and 99, you would need 100 rows in a frequency table.

A **class interval** describes a range of values such as 0–9, meaning any number from 0 to 9 is assigned to that class interval. For discrete values ranging between 0 and 99, using class intervals of size 10 (0–9, 10–19, etc.) requires a frequency table with 10 rows.

When data has been put into class intervals, the **modal class** is the interval that occurs most frequently.

When numerical data is continuous or has a large number of values, numbers need to be grouped into class intervals for the frequency table to form a **grouped frequency table**.

Class intervals must all be the same size.

In general, there should be between 5 and 12 class intervals.
Construct a grouped frequency table for the following scores using class intervals of size 20 (0–19, 20–39…) and find the modal class.

45  78  67  68  59  32  12  99  45  58  56  69
78  16  67  65  51  50  43  22  70  35  66  43
19  21  77  80  89  56  54  61  68  74

**Working**

1. Construct a table with correct variable name and class intervals. The first class interval needs to contain the lowest score but does not have to start with that lowest score.

<table>
<thead>
<tr>
<th>Score</th>
<th>Tally</th>
<th>Frequency (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20–39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40–59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60–79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80–99</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The first score is 45, so place a line in the tally column for class interval 40–59. Continuing in this way, place a tally line for each score. Count the tallies and record their number in the frequency column. Make sure the frequency column adds up to the total number of scores (34).

<table>
<thead>
<tr>
<th>Score</th>
<th>Tally</th>
<th>Frequency (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–19</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>20–39</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>40–59</td>
<td></td>
<td>11</td>
</tr>
<tr>
<td>60–79</td>
<td></td>
<td>13</td>
</tr>
<tr>
<td>80–99</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>34</td>
</tr>
</tbody>
</table>

2. Find the interval that occurs most frequently.

The modal class is 60–79.

**Bar charts**

**Bar charts** help us to see patterns when dealing with categorical variables. Categories can be represented on the horizontal or vertical axis, with their corresponding frequency on the other axis.
This bar chart shows the number of songs downloaded from the Internet over 7 months by a group of teenagers.

Use the graph to answer the following questions.

a In which month were most songs downloaded from the Internet?

b How many songs were downloaded over the 7-month period?

c What percentage of songs were downloaded in March?

**Working**

a The longest bar is January with a frequency of 300. The most songs were downloaded in January.

b Add up the frequency for each month. \[300 + 200 + 100 + 250 + 150 + 200 + 50 = 1250\] Therefore 1250 songs were downloaded over the 7-month period.

c Use the formula \[
\frac{\text{downloads in March}}{\text{total downloads}} \times 100\%\]

\[
\frac{100}{1250} \times 100\% = 8\%, \text{ therefore 8\% of the downloads occurred in March.}\]
### Prep 1

**WORKED EXAMPLE 3**

For the following frequency table, construct a bar chart with the categories on the horizontal axis. Write an explanation of the data displayed on the graph.

<table>
<thead>
<tr>
<th>Hot drink</th>
<th>Tally</th>
<th>Frequency (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cappuccino</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latte</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tea</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>

### Prep 2

**WORKED EXAMPLE 4**

Construct a grouped frequency table for each set of scores and find the modal class.

#### a
18 73 95 56 71 33 50 65 19 54 61 53 87 52 61 37 58 74 45 76 55 92 48 49 81 57 81 44 50 36 71 55 30 20 11 77 91

(use groups of 10–19, 20–29 and so on)

#### b
102 116 100 136 154 176 112 141 134 175 167 108 115 176 115 124 161 160 155 162 188 139 125 154 165

(use groups of 100–119, 120–139, 140–159 and so on)

#### c
103 115 107 132 123 122 109 106 123 131 109 113 116 107 118 130 103 116 130 128 122 104 101 122 103 134 122 108 111

(use groups of 100–104, 105–109, 110–114 and so on)

### Prep 3

**WORKED EXAMPLE 5**

This graph shows the results of a poll before an election.

#### a
How many people favour the Greens?

#### b
How many people favour the Liberal Party?

#### c
Which is the most popular party in the poll?

#### d
How many people favour parties other than the Greens?

#### e
How many more people favour the Labor Party than the Liberal Party?

#### f
How many people were polled altogether?

#### g
What percentage of those polled favoured the Liberal Party?
Table and charts

Question 1

In the frequency table on the right, the missing information is

A  a frequency of 10 and the class interval 15–19
B  a frequency of 3 and the class interval 15–19
C  a frequency of 3 and the class interval 10–19
D  a frequency of 5 and the class interval 15–19
E  a frequency of 47 and the class interval 15–19

Use the following information to answer Questions 2–4.

A number of teenagers were surveyed on how they rated a particular movie, and the results were shown in the following bar chart.

Question 2

The most common star rating for the movie is

A  1 star  B  2 star  C  3 star  D  4 star  E  5 star

Question 3

The number of teenagers who rated the movie is

A  75  B  100  C  150  D  175  E  200

Question 4

The percentage of teenagers who rated the movie as 2 star is closest to

A  9%  B  13%  C  25%  D  33%  E  50%
Use the following information to answer Questions 5–7.

The bar chart shows the results of a quality control survey of parts manufactured at a factory.

### Question 5

The type of data collected is best described as

A categorical and nominal data.  
B numerical and ordinal data.  
C categorical and ordinal data.  
D numerical and discrete data.  
E numerical and continuous data.

### Question 6

The total number of parts surveyed was

A 20  
B 40  
C 120  
D 125  
E 132

### Question 7

The percentage of parts rated as fair is

A 4%  
B 10%  
C 16%  
D 28%  
E 32%

Use the following information to answer Questions 8 & 9.

The following bar chart shows the distribution of wind directions recorded at a weather station at 9.00 a.m. on each of 214 days in 2011.

### Question 8

According to the bar chart, the most frequently observed wind direction was

A south-east  
B south  
C south-west  
D west  
E north-west

[VCAA 2012 1CQ1]
**Question 9**

According to the bar chart, the percentage of the 214 days on which the wind direction was observed to be east or south-east is closest to

A 10%  B 16%  C 25%  D 33%  E 35%

[VCAA 2012 1CQ2]

**Question 10**

In a small survey, twenty-five Year 8 girls were asked what they did (walked, sat, stood, ran) for most of the time during a typical school lunch time.

Their responses are recorded below.

sat stood sat walked walked sat walked ran sat walked walked walked ran walked ran walked ran walked ran ran ran walked

Use the data to

a copy and complete the following frequency table  

<table>
<thead>
<tr>
<th>Activity</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walked</td>
<td></td>
</tr>
<tr>
<td>Sat or stood</td>
<td></td>
</tr>
<tr>
<td>Ran</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>25</strong></td>
</tr>
</tbody>
</table>

b determine the percentage of Year 8 girls who ran for most of the time during a typical school lunch time.

1 mark

[VCAA 2008 2CQ1]

**Question 11**

A development index is used as a measure of the standard of living in a country.

The bar chart displays the development index for 153 countries in four categories: low, medium, high and very high.

![Development Index Bar Chart](https://example.com/bar-chart.png)

n = 153

a How many of these countries have a very high development index?  

1 mark

b What percentage of the 153 countries has either a low or medium development index?

Write your answer correct to the nearest percentage.  

1 mark

[VCAA 2013 2CQ1]
A **histogram** is a graphical way of displaying numerical data from a frequency table. It is effective when dealing with data that has been grouped into a small number (usually between 5 and 15) of intervals.

### Continuous data

Histograms for continuous numerical data have bars which

- correspond to data intervals
- sit in between the interval values on the horizontal axis.

The following is a histogram for the ages of people in a small town where the intervals are < 10 years, 10–<20 years, 20–<30 years, etc.

The interval 10–<20, for example, includes all ages starting at 10 up to but not including 20.

### Grouped discrete data

Histograms for discrete numerical data that has been grouped into intervals work much the same way as for continuous numerical data. For example, the previous histogram could be showing the number of occasions families have eaten takeaway in the last year, where the intervals are 0–9 occasions, 10–19 occasions, etc.

### Ungrouped discrete data

Histograms for discrete numerical data that *hasn't* been grouped into intervals have bars starting and ending halfway between scale marks on the horizontal axis.

The following is a histogram of the number of people living in each house in a suburban block.
**Exam hack**

Although at first glance a histogram looks similar to a bar chart, there are a number of differences.

- Histograms are used for displaying numerical data, while bar charts display categorical data.
- Histograms don’t have any spaces between the columns, while bar charts do.
- Histograms always have the frequency on the vertical axis, while bar charts can have the frequency on either axis.

**Symmetric and skewed distributions**

The shape of a histogram tells us about the underlying frequency distribution. These two histograms show symmetric distributions.

This is a common single-peaked approximately symmetric distribution.

This is a double-peaked approximately symmetric distribution (also called bi-modal).

The following two histograms show skewed distributions.

This histogram is **positively skewed**.

This histogram is **negatively skewed**.

**Exam hack**

To help you remember which skew is positive and which is negative, identify where the ‘tail’ of the histogram is. Think of the positive and negative directions of a number line. If the tail is in the positive direction, the distribution is positively skewed. If the tail is in the negative direction, the distribution is negatively skewed.
Outliers

An outlier is an extreme high or low value in the data. Outliers can indicate an error made in dealing with the data and can sometimes contaminate calculations and conclusions drawn from data sets. However, sometimes they occur without an error being involved. Histograms often make it easier to identify possible outliers.

Medians

An estimate of the median, or middle, of a distribution can often be found by looking at the histogram. The median occurs at the vertical line that splits the histogram in half with equal areas on either side.

Range

The range can be read from a histogram only when we have discrete values increasing by ones. For the following histogram,

Range = largest value – smallest value

= 7 – 1

= 6
Using CAS  Constructing histograms for ungrouped data

The following table shows the number of hours of sleep had by a group of 16-year-olds.

<table>
<thead>
<tr>
<th>Hours</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>8</td>
<td>12</td>
<td>15</td>
<td>16</td>
</tr>
</tbody>
</table>

Display this information on a histogram. Describe the graph.

1. Construct the axes using an appropriate scale. Look at the highest frequency to decide this. Add titles and labels in the centre of each column. Draw each column to match the frequency, remembering that there are no gaps.

2. Comment on the most common score, the general pattern and the presence of any outliers.

9 hours sleep was the most common. Most 16-year-olds have between 6 to 9 hours sleep, with an increasing number having more hours of sleep, up to 9 hours. There is one outlier present in the data (1 person with 2 hours sleep), indicated by the large gap between the first and second columns.

TI-Nspire CAS

STEP 1
Use a Lists & Spreadsheet page.
Name column A hours and column B freq, then enter the data.

STEP 2
Press [menu] 3: Data, 8: Summary Plot

Note that the table is drawn in two rows rather than two columns. Always remember that frequency is assigned to the vertical axis of a histogram regardless of how the table is presented.

This break in the axis indicates that not all the scale marks from zero have been included.
STEP 3
Fill in the summary plot details as per the third screen and click OK.

CLASSPAD

STEP 1
Tap \( m \), then the Statistics application.
Rename list1 as Hours and list2 as Freq.
Enter the data.

STEP 2
Tap SetGraph.
Make sure StatGraph1 only is ticked.
If necessary, tap Stat Window Auto (under the \( \square \) menu) to make sure it is on.
When everything is correct, tap SetGraph, then Setting…
Set the screen as shown and tap Set.

STEP 3
Tap the top left graph icon \( \square \).
Change the HStep to 1.
Tap OK.

STEP 4
The histogram is displayed.
Using CAS Constructing histograms for grouped data

The following table shows the exam scores, out of 100, achieved by a class of Year 11 students on their end-of-year mathematics examination.

<table>
<thead>
<tr>
<th>Exam score</th>
<th>Frequency (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>40–&lt;50</td>
<td>5</td>
</tr>
<tr>
<td>50–&lt;60</td>
<td>6</td>
</tr>
<tr>
<td>60–&lt;70</td>
<td>9</td>
</tr>
<tr>
<td>70–&lt;80</td>
<td>12</td>
</tr>
<tr>
<td>80–&lt;90</td>
<td>3</td>
</tr>
<tr>
<td>90–&lt;100</td>
<td>2</td>
</tr>
</tbody>
</table>

Display this information on a histogram.

Construct the axes using an appropriate scale. Look at your highest frequency to determine the scale for the vertical axis. Add a title and labels on the edges of each column. Draw each column to match its frequency, remembering that there should be no gaps between columns.

**TI-Nspire CAS**

**STEP 1**

Name column A score and column B freq, then enter the data. As class intervals are used, enter the midpoint of each interval.

Press **menu** 3: Data, 8: Summary Plot.

The midpoint of the first class interval is \( \frac{40 + 50}{2} = 45 \).
STEP 2
Fill in the summary plot details as shown on the second screen, then click OK.

The histogram that appears is not in class intervals of 10.

STEP 3

Fill in the width of 10 and Alignment 40. Click OK.

CLASSPAD

STEP 1
Tap and the Statistics application.
Rename list1 as Score and list2 as Freq.
As class intervals are used, enter the midpoints of the class intervals and the frequencies as before.

STEP 2
Use SetGraph as before, making sure StatGraph1 only is ticked and Stat Window Auto is on.
Tap Setting..., set the screen as shown and tap Set.
**STEP 3**
Tap the top left graph icon $\text{Add}$. 
Set $\text{HStart}$ to 45. The columns must be ten units wide, so set $\text{HStep}$ to 10.

Tap $\text{OK}$.

**STEP 4**
The histogram is displayed.

---

**EXAM PREP 4.3**

**Histograms**

**Prep 1**
This histogram displays the results of a Year 12 History quiz.

![Histogram](image)

---

**a** How many students did the History quiz?

**b** Describe the shape of the data.

**c** Using the histogram, copy and complete the frequency table.

**d** What percentage of students achieved a score greater than 8 for the History quiz? Give your answer to 1 decimal place.

<table>
<thead>
<tr>
<th>Score</th>
<th>Frequency $(f)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
</tr>
</tbody>
</table>
Prep 2  USING CAS: CONSTRUCTING HISTOGRAMS FOR UNGROUPED DATA

For each set of scores, construct a frequency table, display the information on a histogram, and interpret the histogram.

\[\begin{array}{cccccccccccccccc}
\text{a} & 8 & 7 & 9 & 5 & 7 & 3 & 5 & 6 & 1 & 5 & 6 & 5 & 8 \\
   & 5 & 6 & 3 & 5 & 7 & 4 & 7 & 5 & 9 & 4 & 4 & 8 \\
\text{b} & 12 & 16 & 10 & 16 & 14 & 16 & 12 & 11 & 14 & 15 & 17 & 18 & 15 \\
   & 16 & 15 & 14 & 11 & 10 & 15 & 12 & 18 & 19 & 15 & 14 & 15 \\
\text{c} & 103 & 105 & 104 & 102 & 103 & 101 & 100 & 105 & 104 & 103 & 103 & 103 \\
   & 104 & 101 & 107 & 102 & 101 & 104 & 106 & 104 & 102 & 102 & 101 \\
\end{array}\]

Prep 3  USING CAS: CONSTRUCTING HISTOGRAMS FOR GROUPED DATA

Bungee jumpers must be weighed before they 'take the plunge'. The weights in kg of 40 jumpers were recorded as follows.

\[\begin{array}{cccccccccccccccc}
41 & 58 & 63 & 37 & 49 & 58 & 71 & 33 & 85 & 58 \\
60 & 73 & 81 & 46 & 55 & 38 & 80 & 48 & 50 & 62 \\
61 & 59 & 63 & 44 & 77 & 62 & 58 & 73 & 62 & 75 \\
52 & 60 & 69 & 61 & 55 & 47 & 76 & 42 & 66 & 70 \\
\end{array}\]

\[\begin{array}{cccccccccccccccc}
a & \text{Use class intervals of } 30–<40, 40–<50 \text{ and so on to construct a frequency table for the data.} \\
b & \text{Construct a histogram to display this data.} \\
c & \text{From 200 bungee jumpers, how many would you expect to weigh 70 kg or more?} \\
   & \text{Justify your answer.} \\
\end{array}\]

Prep 4

Find the median of the following histogram.
Question 1

What is the most appropriate description for the shape of the following histogram?

A  symmetrical
B  bimodal
C  positively skewed with an outlier
D  negatively skewed with an outlier
E  none of the above

Question 2

1526 students sat for an examination. The histogram shows the distribution of marks.

The median examination mark of these students is closest to

A  50  B  55  C  60  D  65  E  70

[VCAA 2002 1CQ5]
Use the following information to answer Questions 3 & 4.
The percentage histogram shows the distribution of the fertility rates (in average births per woman) for 173 countries in 1975.

**Question 3**
In 1975, for these 173 countries, fertility rates were most frequently
- A less than 2.5
- B between 1.5 and 2.5
- C between 2.5 and 4.5
- D between 6.5 and 7.5
- E greater than 7.5

[VCAA 2009 1CQ5]

**Question 4**
In 1975, the percentage of these 173 countries with fertility rates of 4.5 or greater was closest to
- A 12%
- B 35%
- C 47%
- D 53%
- E 65%

[VCAA 2009 1CQ4]

Use the following information to answer Questions 5–7.
The histogram displays the distribution of the percentage of Internet users in 160 countries in 2007.

**Question 5**
The shape of the histogram is best described as
- A approximately symmetric.
- B bell shaped.
- C positively skewed.
- D negatively skewed.
- E bimodal.

[VCAA 2011 1CQ1]

**Question 6**
The number of countries in which less than 10% of people are Internet users is closest to
- A 10
- B 16
- C 22
- D 32
- E 54

[VCAA 2011 1CQ2]

**Question 7**
From the histogram, the median percentage of Internet users is closest to
- A 10%
- B 15%
- C 20%
- D 30%
- E 40%

[VCAA 2011 1CQ3]
Use the following information to answer Questions 8 & 9.

The distribution of test marks obtained by a large group of students is displayed in the percentage frequency histogram.

**Question 8**

The pass mark on the test was 30 marks.

The percentage of students who passed the test is

A 7%  
B 22%  
C 50%  
D 78%  
E 87%

[VCAA 2006 1CQ5]

**Question 9**

The median mark lies between

A 35 and 40  
B 40 and 45  
C 45 and 50  
D 50 and 55  
E 55 and 60

[VCAA 2006 1CQ6]

**Question 10**

The histogram above is best described as

A negatively skewed.  
B positively skewed.  
C symmetric.  
D negatively skewed with outliers.  
E positively skewed with outliers.

[VCAA 2005 1CQ3]
Question 11

The histogram shows the distribution of mean yearly rainfall (in mm) for Australia over 103 years.

Data source: ABS 2007

a Describe the shape of the histogram. 1 mark

b Use the histogram to determine
   i the number of years in which the mean yearly rainfall was 500 mm or more 1 mark
   ii the percentage of years in which the mean yearly rainfall was between 500 mm and 600 mm. Write your answer correct to one decimal place. 1 mark

[VCAA 2007 2CQ1]

Question 12

The age, in years, of employees at Burger Heaven were recorded as follows.

18 19 18 17 20 20 24 15 24 19
15 40 21 17 20 22 23 21 24 23
34 19 45 20 15 21 24 27 19 33
34 24 16 18 30 21 26 31 16 25
49 21 21 35 16 22 15 25 44 23.

a Organise the data into a frequency table using class intervals of 15–<20, 20–<25, and so on.
b How many people work at Burger Heaven?
c Construct a histogram to display this data.
d What percentage of employees are aged 40 or over?
e Describe what the graph shows.
**Quartiles**

Quartiles are the three points that divide a set of data into quarters.

- The first or **lower quartile** has 25% of the data below it.
- The second quartile (which is the same as the median) has 50% of the data below it.
- The third or **upper quartile** has 75% of the data below it.

**Worked example 6**

Find the three quartiles for the following data and show how they divide the data into quarters.

34  25  23  22  24  34  21  48  6  30  21  29

**Working**

1. Order the data from smallest to largest.
   
   $6, 21, 21, 22, 23, 24, 25, 29, 30, 34, 34, 48$

2. Find the median.
   
   There is an even number of data values, so there are two middle points (24 and 25). Add them and divide by 2 to find the median.

   $\text{Median} = \frac{24 + 25}{2} = 24.5$

3. Find the median of the lower half of the data. This is the **lower quartile**.
   
   The lower half of the data is 6, 21, 21, 22, 23, 24.
   
   $\text{Lower quartile} = \frac{21 + 22}{2} = 21.5$.

4. Find the median of the upper half of the data. This is the **upper quartile**.
   
   The upper half of the data is 25, 29, 30, 34, 34, 48.
   
   $\text{Upper quartile} = \frac{30 + 34}{2} = 32$.

5. Show how the three quartiles divide the data into quarters.

   $\text{6, 21, 21, 22, 23, 24, 25, 29, 30, 34, 34, 48}$

   $\text{Lower quartile} = 21.5 \quad \text{Upper quartile} = 32$

   $\text{Median} = 24.5$
We use the following symbols when dealing with quartiles:

- $Q_1$ is the lower quartile (the median of the lower half of the data)
- $Q_2$ is the median
- $Q_3$ is the upper quartile (the median of the upper half of the data)

**Exam hack**

If there is an odd number of data values (which means that the median will be one of the actual data values), then you can’t split the data into equal lower and upper halves. When this happens, leave the median value out when you do the quartile calculations.

**The five-number summary**

The **five-number summary** provides a good overview of a distribution. It consists of

1. The minimum data value
2. $Q_1$
3. The median
4. $Q_3$
5. The maximum data value

So, for Worked example 6, the five-number summary is:

Minimum = 6, $Q_1 = 21.5$, Median = 24.5, $Q_3 = 32$, Maximum = 48
Using CAS  Five-number summary

A CAS/calculator will calculate the five-number summary.

**TI-NSPRIE CAS**

**STEP 1**
Using a New Document, enter the data into the Lists & Spreadsheet page.
Name column A ‘data’, then add the data into this list.

**STEP 2**
In the pop up screen that appears, set the number of lists to 1 and press [enter].

**STEP 3**
Select ‘data’ for the X1 List, press ▼ then ▼ until ‘data’ is selected, then press [enter]. Leave the frequency list set as 1, then press [tab] down to OK and press [enter].

**STEP 4**
Scroll to find the values of MinX, Q1X, MedianX, Q3X and MaxX.
IQR and outliers

The interquartile range (or IQR) is the measure of the spread of the middle 50% of the data values.

\[ \text{IQR} = Q_3 - Q_1 \]

The IQR is often a better measure of spread than the range because, by looking at only the middle 50% of data, we avoid taking outliers into account.

The IQR is also used to define possible outliers, so we don’t have to rely on imprecise methods (such as simply saying something ‘looks like an outlier’).

A data value is a possible outlier if it is

- less than the lower fence \( Q_1 - 1.5 \times \text{IQR} \) or
- greater than the upper fence \( Q_3 + 1.5 \times \text{IQR} \)
For the data set 6, 21, 21, 22, 24, 25, 29, 30, 34, 34, 48, confirm whether 6 and 48 are possible outliers.

**Working**

1. Find $Q_1$ and $Q_3$ using a CAS/calculator.
   - $Q_1 = 21.5$ and $Q_3 = 32$

2. Calculate the IQR.
   - $IQR = 32 - 21.5 = 10.5$

3. Calculate $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$.
   - $Q_1 - 1.5 \times IQR = 21.5 - 1.5 \times 10.5 = 5.75$
   - $Q_3 + 1.5 \times IQR = 32 + 1.5 \times 10.5 = 47.75$

4. Check the potential outliers to see if they are
   - less than $Q_1 - 1.5 \times IQR$ or
   - greater than $Q_3 + 1.5 \times IQR$
   - 6 isn't less than 5.75, so it's not an outlier.
   - 48 is greater than 47.75, so it is a possible outlier.

**Boxplots**

**Boxplots**, also known as **box-and-whisker plots**, display numerical data based on the five-number summary, IQR and outliers.

If there are no outliers, the **whiskers** show the minimum and maximum values.

![Boxplot without outliers](image)

When there are outliers, the whiskers show the lowest or highest values that are not outliers. Outliers are shown as points.

![Boxplot with outliers](image)
Boxplots provide the following information:

- **Minimum**
- **Median**
- **Q1**
- **Q3**
- **IQR**
- **1.5 × IQR**
- **Maximum**
- **Lower fence = Q1 + 1.5 × IQR**
- **Upper fence = Q3 + 1.5 × IQR**

Boxplots can also be displayed vertically.

**Using CAS**  
**Boxplots**

Use a CAS/calculator to construct a boxplot for the data set 6, 21, 21, 22, 23, 24, 25, 29, 30, 34, 34, 48.

**TI-Nspire CAS**

**STEP 1**

Using a New Document, enter the data into the Lists & Spreadsheet page.

Name column A 'data' or something more specific (e.g. 'weights', 'scores', 'ages', 'wages'), then add the data into this list.

**STEP 2**

Press ctrl doc to add a Data & Statistics page, then press [VARS] and select 'data' (or whatever you’ve called it) as the x variable.

A dot plot will appear. This is the default display.

**STEP 3**

To change it to a boxplot, press [MENU] 1: Plot Type, 2: Box Plot.

**STEP 4**

As you move the cursor over the graph, the important values will appear.
**CLASSPAD**

**STEP 1**
Using the Statistics application, enter the data. Rename list1 as 'data' or something more specific (e.g. 'weights', 'scores', 'ages', 'wages') by first tapping the list1 cell. To access the letter keyboard, press `Keyboard` and tap `abc`.

**STEP 2**
Tap `SetGraph`, making sure StatGraph1 only is ticked and Stat WindowAuto under the menu is on.
Tap `Setting`.
For **Type**, select MedBox.
For **Xlist**, select main\wages (or whatever you renamed list1 as).
Make sure that the box for **Show Outliers** is checked.

**STEP 3**
Tap `Set`, then tap `An`.

**STEP 4**
Tap on the graph screen and then tap **Analysis** followed by **Trace**. Using the right and left arrow keys enables you to read important values from the graph.
Comparing boxplots and histograms

If you know what the histogram of a distribution looks like, you can often get some idea of what the boxplot looks like.

**Symmetric distributions**
- Median approximately in the middle
- Boxes and whiskers about the same length

**Positively skewed distributions**
- Median usually to the left of centre
- Long right-hand whisker

**Negatively skewed distributions**
- Median usually to the right of centre
- Long left-hand whisker

**Distributions with outliers**
- Boxplot matches the histogram, ignoring outliers
- Outliers shown by dots
Boxplots

Prep 1  WORKED EXAMPLE 6  USING CAS: FIVE-NUMBER SUMMARY

For the following percentage test scores:
73  65  54  90  74  51  61  88  47  92  71  66
a  find the three quartiles by hand and show how they divide the data into quarters
b  verify your answers by using a CAS/calculator and finding the five-number summary.

Prep 2

For the following boxplots, state
i  the five-number summary
ii  the values between which the middle 50% of the data lies.

a

b

Test scores

Test scores

c

Test scores

Prep 3  WORKED EXAMPLE 7  USING CAS: BOXPLOTS

For each of the following data sets, use a CAS/calculator to construct a boxplot and do a calculation to confirm possible outliers.

a  45  50  44  54  37  44  15  50  41  52  38  26  37  42  48  39  46  49  43  64
b  104  88  110  99  40  156  96  97  86  105  79  77  89  93  95  110  99  115
c  15  22  25  25  25  27  17  27  23  20  24  26  29  27  21  25
Prep 4

This boxplot represents the amount of pocket money, in dollars, earned by a sample of 48 children.

![Boxplot of pocket money](image)

**a** What percentage of children earned less than $15?

**b** How many children earned $22 or more of pocket money?

---

### EXAM PRACTICE 4.4

**Boxplots**

*Use the following information to answer Questions 1 & 2.*

**Question 1**

The percentage of data below 10 is

- **A** 25%
- **B** 50%
- **C** 75%
- **D** 100%
- **E** None of the above

**Question 2**

The value of $Q_1$ is

- **A** 5
- **B** 8
- **C** 10
- **D** 14
- **E** 16

*Use the following information to answer Questions 3–6.*

This boxplot represents the annual wages (× $1000) of the administration staff at a TAFE college.

**Question 3**

The median annual wage is

- **A** $18
- **B** $41
- **C** $18 000
- **D** $41 000
- **E** $52 000

**Question 4**

The range of annual wages is

- **A** $58
- **B** $41
- **C** $58 000
- **D** $41 000
- **E** $52 000
Chapter 4: Data distributions 173

**Question 5**
Between what two amounts are the middle 50% of staff annual wages?

A  $28 000 and $51 000  
B  $28 000 and $41 000  
C  $41 000 and $51 000  
D  $29 000 and $50 500  
E  $17 000 and $59 000

**Question 6**
What percentage of the staff earn less than $28 000 per year?

A  50%  
B  75%  
C  20%  
D  0%  
E  25%

**Question 7**
The longest time, in seconds, spent moving along this aisle is closest to

A  40  
B  60  
C  190  
D  450  
E  500

**Question 8**
The shape of the distribution is best described as

A  symmetric.  
B  negatively skewed.  
C  negatively skewed with outliers.  
D  positively skewed.  
E  positively skewed with outliers.

**Question 9**
The number of customers who spent more than 90 seconds moving along this aisle is closest to

A  7  
B  20  
C  26  
D  75  
E  79
The following data was recorded from measurements made on 12 men.

For these men, the median age ($M$) and the interquartile range (IQR), in years, are respectively

A $M = 37$ and IQR $= 10$
B $M = 37.4$ and IQR $= 7.2$
C $M = 37.4$ and IQR $= 6.9$
D $M = 38$ and IQR $= 10$
E $M = 38$ and IQR $= 25$

Use the following information to answer Questions 11 & 12.

To test the temperature control on an oven, the control is set to 180°C and the oven is heated for 15 minutes. The temperature of the oven is then measured. Three hundred ovens were tested in this way. Their temperatures were recorded and are displayed using both a histogram and a boxplot.

### Question 11

The interquartile range for temperature is closest to

A 1.3°C  
B 1.5°C  
C 2.0°C  
D 2.7°C  
E 4.0°C

### Question 12

A total of 300 ovens were tested and their temperatures were recorded.

The number of these temperatures that lie between 179°C and 181°C is closest to

A 40  
B 50  
C 70  
D 110  
E 150
Question 13

The percentage histogram shows the distribution of the fertility rates (in average births per woman) for 173 countries in 1975.

Which one of the boxplots below could best be used to represent the same fertility rate data as displayed in the percentage histogram?

A  B  C

D  E

[VCAA 2009 1CQ6]

Question 14

This boxplot shows the number of cigarettes smoked per day by a sample of 60 smokers who were trying to quit.

a  What is the median number of cigarettes smoked per day?

b  What is the interquartile range?

c  What is the lower extreme?

d  What percentage of people smoked between 20 and 25 cigarettes per day?

e  Estimate how many people smoked fewer than 20 cigarettes per day.

Question 15

Provide evidence why the boxplot shown below is incorrect.
Dot plots

Dot plots are the simplest way to display numerical data. They are best used for a maximum of 50 data values and when the data values are not too spread out.

Looking at the dot plot of the ages of students at a school function, we can easily read the following from the plot:

The minimum data value = 12
The maximum data value = 17
The range = largest value – smallest value = 17 – 12 = 5
The mode = 16 years.
To get more information from a dot plot, it’s necessary to count dots.
To calculate the median, you need to

1 count the total number of dots (25 in this case)
2 then count from the left to find where the middle dot occurs (in this case, it’s the 13th dot because it has 12 dots before it and 12 dots after it)
3 write down the value where this occurs. (In this example, the median is 15 years.)

The lower quartile is the median of the lower 12 values, so we need to find the 6th and 7th value, then add them together and divide by 2.

\[ Q_1 = \frac{13 + 14}{2} = 13.5 \]

The upper quartile is the median of the upper 12 values, so we need to find the 6th and 7th value from the right end, then add them together and divide by 2.

\[ Q_3 = \frac{16 + 16}{2} = 16 \]

The interquartile range, IQR = \( Q_3 - Q_1 = 16 - 13.5 = 2.5 \).

Exam hack

To decide whether a dot plot is symmetric, positively skewed or negatively skewed, look at the shape the dots are making and try to picture it as a histogram.
**Stem plots**

Stem plots, also known as stem-and-leaf plots, are an alternative to histograms whose main advantage is that the actual data values appear. A stem plot is best used with up to a maximum of 50 data values, otherwise it becomes unwieldy. We would normally use ordered stem plots where the leaves are ordered from smallest to largest. A key is always required.

This stem plot represents test scores out of 60 for a class of 33 students.

Because you can see all the data values, it is relatively straightforward to find the five-number summary and other information directly from the plot.

Looking at the example:

- the minimum data value = 1
- the maximum data value = 57
- there are 33 ordered values, so the median will be the 17th value (i.e. it will have 16 values on either side). Counting up to the 17th value from the start, the median = 35.
- the lower quartile is the median of the lower 16 values, so we need to find the 8th and 9th value, then add them together and divide by 2.
  \[ Q_1 = \frac{23 + 23}{2} = 23 \]
- the upper quartile is the median of the upper 16 values, so we need to find the 8th and 9th value from the bottom right corner of the plot, then add them together and divide by 2.
  \[ Q_3 = \frac{48 + 49}{2} = 48.5 \]
- the interquartile range, IQR = \( Q_3 - Q_1 = 48.5 - 23 = 25.5 \)
- the range = largest value – smallest value = 57 – 1 = 56
- there are two modes, 21 and 49, so we say this data set is bimodal
- the 1 score may be an outlier. Check using the lower fence
  \[ Q_1 - 1.5 \times IQR = 23 - 1.5 \times 25.5 = -15.25 \]
  Since 1 isn’t less than –15.25, it’s not an outlier. (Note that negative test scores would be impossible anyway.)

**Exam hack**

To decide whether a stem plot is symmetric, positively skewed or negatively skewed, rotate the page so that the stem forms the horizontal axis, look at the shape the data values are making, and try to picture it as a histogram.
Dot plots and stem plots

**Prep 1**

This dot plot shows the minimum daily temperatures (°C) in Hobart over a 3-week period.

\[ \begin{align*}
\text{a} & \quad \text{What is the mode?} \\
\text{b} & \quad \text{What position is the median in?} \\
\text{c} & \quad \text{What is the median?}
\end{align*} \]

**Prep 2**

This unordered stem-and-leaf plot represents the number of points scored per match by the Blues in a disappointing football season.

\[ \begin{align*}
\text{a} & \quad \text{Display the data using an ordered stem-and-leaf plot.} \\
\text{b} & \quad \text{How many matches were played in the season?} \\
\text{c} & \quad \text{What was the Blues' highest score for a match?} \\
\text{d} & \quad \text{For what percentage of matches did the Blues score below 56 points?}
\end{align*} \]

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>5 3 9</td>
</tr>
<tr>
<td>5</td>
<td>7 2 0 8</td>
</tr>
<tr>
<td>6</td>
<td>4 7 8 5 1 2</td>
</tr>
<tr>
<td>7</td>
<td>2 9 3 0</td>
</tr>
<tr>
<td>8</td>
<td>9 4 2</td>
</tr>
</tbody>
</table>

Key: 4|5 = 45
Dot plots and stem plots

Use the following information to answer Questions 1–3.

The stem plot on the right shows the times (in seconds) of skiers who finished a slalom ski race.

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>1 5 7 9</td>
</tr>
<tr>
<td>10</td>
<td>2 4 5 6 6 8</td>
</tr>
<tr>
<td>11</td>
<td>0 2 2 3 4 4 5</td>
</tr>
<tr>
<td>12</td>
<td>1 2 3 3 3 7 9</td>
</tr>
<tr>
<td>13</td>
<td>2 3 4 5 7 7</td>
</tr>
<tr>
<td>14</td>
<td>3 6 9</td>
</tr>
<tr>
<td>15</td>
<td>0 1 2</td>
</tr>
</tbody>
</table>

Key: 9|1 means 91 seconds

**Question 1**

The number of skiers who finished the race is

A 7  B 15  C 23  D 36  E 152

**Question 2**

The winning time is

A 91 seconds  B 110 seconds  C 132 seconds  D 143 seconds  E 152 seconds

**Question 3**

If skiers with times under 110 seconds were of Olympic standard, the percentage of skiers of this standard is closest to

A 12%  B 22%  C 25%  D 27%  E 28%

Use the following information to answer Questions 4 & 5.

The dot plot shows the distribution of the number of bedrooms in each of 21 apartments advertised for sale in a new high-rise apartment block.

**Question 4**

The mode of this distribution is

A 1  B 2  C 3  D 7  E 8

[VCAA 2007 1CQ1]

**Question 5**

The median of this distribution is

A 1  B 2  C 3  D 4  E 5

[VCAA 2007 1CQ2]
Use the following information to answer Questions 6 & 7.

The ordered stem plot on the right shows the percentage of homes connected to broadband Internet for 24 countries in 2007.

Question 6

The number of these countries with more than 22% of homes connected to broadband Internet in 2007 is

A 4  B 5  C 19  D 20  E 22

[VCAA 2013 1CQ1]

Question 7

Which one of the following statements relating to the data in the ordered stem plot is not true?

A The minimum is 16%.
B The median is 30%.
C The first quartile is 23.5%.
D The third quartile is 32%.
E The maximum is 38%.

[VCAA 2013 1CQ2]

Use the following information to answer Questions 8 & 9.

The marks obtained by students who sat for a test are displayed as an ordered stem plot as shown.

Question 8

The number of students who sat the test is

A 25  B 26  C 27  D 32  E 50

[VCAA 2004 1CQ1]

Question 9

The interquartile range of these test marks is closest to

A 9  B 13  C 30  D 36  E 41

[VCAA 2004 1CQ2]
**Question 10**

Which of these displays could be described as positively skewed with a possible outlier?

A

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>1 5 6 8 9 9</td>
</tr>
<tr>
<td>8</td>
<td>0 0 2 3 4 6 8 9</td>
</tr>
<tr>
<td>9</td>
<td>0 2 5 7 9</td>
</tr>
<tr>
<td>10</td>
<td>2 3 5 6 7 7 8</td>
</tr>
</tbody>
</table>

Key: 7|9 = 79

B

C

D

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>4 7 9 9 9</td>
</tr>
<tr>
<td>16</td>
<td>0 2 3 7 8 8</td>
</tr>
<tr>
<td>17</td>
<td>0 0 2 4 4 5 5 7 7 7 9</td>
</tr>
<tr>
<td>18</td>
<td>3 9</td>
</tr>
<tr>
<td>19</td>
<td>0 1 2</td>
</tr>
</tbody>
</table>

Key: 12|3 = 123

E

**Question 11**

The development index for each country is a whole number between 0 and 100.

The dot plot below displays the values of the development index for each of 28 countries that have a high development index.

a Using the information in the dot plot, determine the mode and the range. 1 mark

b Write down an appropriate calculation and use it to explain why the country with a development index of 70 is an outlier for this group of countries. 2 marks

[VCAA 2013 2CQ2]
The table shows the number of rainy days recorded in a high rainfall area for each month during 2008.

<table>
<thead>
<tr>
<th>Month</th>
<th>Number of rainy days</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>12</td>
</tr>
<tr>
<td>February</td>
<td>8</td>
</tr>
<tr>
<td>March</td>
<td>12</td>
</tr>
<tr>
<td>April</td>
<td>14</td>
</tr>
<tr>
<td>May</td>
<td>18</td>
</tr>
<tr>
<td>June</td>
<td>18</td>
</tr>
<tr>
<td>July</td>
<td>20</td>
</tr>
<tr>
<td>August</td>
<td>19</td>
</tr>
<tr>
<td>September</td>
<td>17</td>
</tr>
<tr>
<td>October</td>
<td>16</td>
</tr>
<tr>
<td>November</td>
<td>15</td>
</tr>
<tr>
<td>December</td>
<td>13</td>
</tr>
</tbody>
</table>

The dot plot below displays the distribution of the number of rainy days for the 12 months of 2008.

a Copy the dot plot and circle the dot that represents the number of rainy days in April 2008.

b For the year 2008, determine
   i the median number of rainy days per month
   ii the percentage of months that have more than 10 rainy days.

   Write your answer correct to the nearest per cent.

[VCAA 2009 2CQ1]
**Question 13**

The stem plot in Figure 1 shows the distribution of the average age, in years, at which women first marry in 17 countries.

a For these countries, determine
   i the lowest average age of women
      at first marriage  1 mark
   ii the median average age of women
      at first marriage.  1 mark

The stem plot in Figure 2 shows the distribution of the average age, in years, at which men first marry in 17 countries.

b For these countries, determine the interquartile range (IQR) for the average age of men
   at first marriage.  1 mark

c If the data values displayed in Figure 2 were used to construct a boxplot with outliers, then the country for which the average age of men at first marriage is 26.0 years would be shown as an outlier. Explain why this is so. Show an appropriate calculation to support your explanation.  2 marks

**Figure 1: Average age, in years, of women at first marriage**

<table>
<thead>
<tr>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
</tr>
<tr>
<td>25</td>
</tr>
<tr>
<td>26</td>
</tr>
<tr>
<td>27</td>
</tr>
<tr>
<td>28</td>
</tr>
<tr>
<td>29</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>31</td>
</tr>
</tbody>
</table>

Key: 27|3 represents 27.3 years

**Figure 2: Average age, in years, of men at first marriage**

<table>
<thead>
<tr>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
</tr>
<tr>
<td>26</td>
</tr>
<tr>
<td>27</td>
</tr>
<tr>
<td>28</td>
</tr>
<tr>
<td>29</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>31</td>
</tr>
<tr>
<td>32</td>
</tr>
<tr>
<td>33</td>
</tr>
</tbody>
</table>

Key: 32|5 represents 32.5 years

[**VCAA 2011 2CQ1**]

**Question 14**

This dot plot shows the number of children in each family living on Willard Crescent.

a How many families live on Willard Crescent?

b What is the mode?

c What is the median?

d State any outliers.

e If the outlier is removed from the data set, explain why there is no effect on
   i the mode
   ii the median
4.6 Back-to-back stem plots and parallel boxplots

There are times when we want to investigate the association between a numerical variable and a categorical variable. To do this we can use a back-to-back stem plot or parallel boxplot.

**Back-to-back stem plots**

Sometimes we want compare the distribution of a numerical variable for two groups. For example, we may want to compare the test scores for two classes or heights of two football teams. One way of displaying this sort of data is through back-to-back stem plots.

A back-to-back stem plot has two sets of leaves, one on the left of the stem and one on the right. This allows us to display the data for the two groups being compared, as in the example below.

Test results for two classes

<table>
<thead>
<tr>
<th>Class A</th>
<th>Stem</th>
<th>Class B</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 3 2</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>8 5 4 3 1</td>
<td>6</td>
<td>0 0 2 4</td>
</tr>
<tr>
<td>9 6 4 4 3 1 0</td>
<td>7</td>
<td>3 5 6 7 8 8 8 9</td>
</tr>
<tr>
<td>8 8 7 5 3 2 0</td>
<td>8</td>
<td>0 1 4 6 9</td>
</tr>
<tr>
<td>1 0</td>
<td>9</td>
<td>2 3 5 6 7 7 8</td>
</tr>
</tbody>
</table>

Key: 5|6 = 56

Exam hack

To comment on the shape of the data on the left side of a back-to-back stem plot, it may help to picture what it would look like if it was on the right side of a back-to-back stem plot.
Liz and George deliver pamphlets to letterboxes. The number of pamphlets delivered per hour over 12 hours is shown below.

**Liz:** 32 24 27 35 28 26 32 28 31 35 32 25

**George:** 21 18 24 31 25 38 32 15 45 29 31 38

**a** Display the data with a back-to-back stem plot.

**b** Comment on the shape of the data for each person.

**c** Calculate the median, range and IQR for the number of pamphlets delivered by each person.

**d** Who would you say is the better delivery person and why?

**Working**

**a** As the data range is reasonably small, split the stems so that the shape of the data will be easier to interpret. Remember to order the values of the leaves.

**George** | **Stem** | **Liz**
---|---|---
8 5 | 1 | 4
4 1 | 2 | 4
9 5 | 2 | 5 6 7 8 8
2 1 1 | 3 | 1 2 2 2
8 8 | 3 | 5 5
4 |
5 |

**b** Comment on symmetry and skewness for each distribution.

Liz’s data is approximately symmetrical, whereas George’s looks more negatively skewed.

**c** Calculate the median, range and IQR from the back-to-back stem plot.

Liz: Median = 29.5, range = 11, IQR = 32 – 26.5 = 5.5

George: Median = 30, range = 30, IQR = 35 – 22.5 = 12.5

**d** Use the results to decide who is the better delivery person.

The medians are about the same, but George’s range and IQR are considerably higher than Liz’s. This means Liz’s deliveries have less variability and are more consistent than George’s, which indicates that Liz is the better delivery person.
Parallel boxplots

Parallel boxplots can be used to compare the distribution of a numerical variable across two or more groups. While back-to-back stem plots are limited to two groups and small sets of data, parallel boxplots aren’t restricted in this way. It is also easier to compare medians, quartiles and ranges visually from parallel boxplots than from back-to-back stem plots.

Here’s an example based on the test results for four Year 11 classes. It’s relatively easy to find which class has the highest median, lowest $Q_1$ or highest maximum value etc.

![Parallel boxplots example](image)

Using CAS Constructing parallel boxplots

The test results of two Year 11 classes in General Mathematics are shown below.

**Class A:**
58  46  53  52  67  36  61  49  47  
59  66  53  94  69  46  44  57  

**Class B:**
60  50  70  69  86  43  60  60  44  56  
49  50  56  56  42  65  47  67  25  46  

Construct parallel boxplots for the data.

**TI-Nspire CAS**

**STEP 1**
Open a New Document with a Lists & Spreadsheet page.

Type each set of data into a column.

**STEP 2**
Add a Data & Statistics page.

Click in the ‘Click to add variable’ space at the bottom of the page.

Select classa and a dot plot will appear.
### STEP 3
Press \( \text{menu} \)
1. Plot Type
2. Box Plot

### STEP 4
Press
2. Plot Properties
5. Add X Variable
Select classb.

### STEP 5
Use the touch pad to move around the screen and the five-number summary values will appear.

---

**CLASSPAD**

### STEP 1
Use the \( \text{Statistics} \) application.
Type each set of data into a column.

### STEP 2
Tap SetGraph.
Make sure that StatGraph1 and StatGraph2 are checked.
Tap Setting.

### STEP 3
Complete the window as shown, then tap the tab numbered 2.
STEP 4
Complete the window as shown, then tap [Set].

STEP 5
Tap †, then View Window and set xmin and max below it and ymin and max below that as shown on the screen on the right. Tap OK.

STEP 6
Tap [DEL], then Analysis and Trace. Use the arrow keys to move from one plot to the other and read the five-number summary from each boxplot, which appears at the bottom of the screen.
The parallel boxplots below represent the average temperature in June for Victoria, New South Wales and Queensland over a number of years.

a Which state has the highest median average June temperature?
b Which state has the largest range of average June temperatures?
c Which state’s data is best described as positively skewed?
d Which state had the lowest average June temperature?
e Write a brief summary comparing the average June temperature for each state.
f Do you think there is an association between the states and their average June temperatures? Provide some evidence for your view.

**Working**

a Look for the state whose median line is furthest along the scale. Queensland  

b Look for the longest boxplot, including whiskers. New South Wales  

c Look for the boxplot with its median to the left of the box and a right whisker being longer than the left one. Queensland  

d Look for the lowest left endpoint. New South Wales  

e Compare the states, noting similarities and differences. Victoria and New South Wales have similar average June temperatures, whereas Queensland has higher average June temperatures than the other two states.  

f Are there differences? There may be an association between the state and average June temperature as Queensland’s data is significantly different from the other two states. The middle 50% of the Queensland data has no overlap at all with the middle 50% of the NSW or Victorian data.
Which display do you use?

Often there is more than one suitable display, but here are some guidelines on which statistical display is the best one to use:

<table>
<thead>
<tr>
<th>Display</th>
<th>Type of data</th>
<th>Guidelines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bar chart</td>
<td>Categorical data</td>
<td>Categories can be represented on the horizontal or vertical axis.</td>
</tr>
<tr>
<td>Histogram</td>
<td>Numerical data</td>
<td>Best if data has been grouped into between 5 and 15 intervals. Can be used for a large number of data values.</td>
</tr>
<tr>
<td>Dot plot</td>
<td>Numerical data</td>
<td>Best used with a maximum of 50 data values and when the data values are not too spread out.</td>
</tr>
<tr>
<td>Boxplot</td>
<td>Numerical data</td>
<td>Best if you want to read the five-number summary easily.</td>
</tr>
<tr>
<td>Stem plot</td>
<td>Numerical data</td>
<td>Best used with a maximum of 50 data values.</td>
</tr>
<tr>
<td>Back-to-back stem plots</td>
<td>Categorical and numerical data</td>
<td>Used to compare the distribution of a single numerical variable across two categorical groups, where the two data sets are small.</td>
</tr>
<tr>
<td>Parallel boxplots</td>
<td>Categorical and numerical data</td>
<td>Used to compare the distribution of a single numerical variable across two or more categorical groups, where the data sets may be large and you want to compare the five-number summaries.</td>
</tr>
</tbody>
</table>

**EXAM PREP 4.6**

**Back-to-back stem plots and parallel boxplots**

**Prep 1 WORKED EXAMPLE 8**

Two radar cameras that were positioned on different roads recorded car speeds (in km/h) as follows.

**Camera 1:**
78 63 75 69 71 83 80 67 74 72 73 74 90 83 65 73 69 89 76 102 83 78 69 71

**Camera 2:**
112 139 120 116 116 136 140 123 135 131 120 117 138 131 127 119 125 130 130 134 123 148 169 130

**a** Display the data with a back-to-back stem plot.
**b** Comment on the shape of the data for each camera.
**c** Calculate the median, range and IQR for the car speeds recorded by each of the two radar cameras.
**d** Is there an association between the position of the cameras and the speed recorded? Justify your answer.
4.6

Prep 2

USING CAS: CONSTRUCTING PARALLEL BOXPLOTS

Construct parallel boxplots for the data in Prep 1.

Prep 3

WORKED EXAMPLE 9

Yashneel decided to investigate his theory that ‘People tend to underestimate the length of a piece of string’. He asked students to estimate the lengths of several pieces of string, measured the actual lengths and displayed the results in parallel boxplots.

![Parallel Boxplot Example]

- Write down the median of:
  - the estimated lengths
  - the actual lengths.
- Compare the range of each data set.
- Compare the interquartile range of each data set.
- Classify the shape of each data set.
- Do you think that the data supports Yashneel’s theory? Justify your answer.

EXAM PRACTICE 4.6

Back-to-back stem plots and parallel boxplots

Question 1

Kaylee measured the heights of 50 Year 11 students for her statistics assignment. The most appropriate graphical display for the data is

A  a bar chart  
B  a dot plot  
C  parallel boxplots  
D  a histogram  
E  a back-to-back stem plot

Use the following information to answer Questions 2–4.

The back-to-back ordered stem plot below shows the female and male smoking rates, expressed as a percentage, in 18 countries.

<table>
<thead>
<tr>
<th>Smoking rates (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Female</strong></td>
</tr>
<tr>
<td>9 9 9 7 7 6 5</td>
</tr>
<tr>
<td>8 6 5 5 5 5 5 3 2 1 0</td>
</tr>
<tr>
<td><strong>Male</strong></td>
</tr>
<tr>
<td>2 4 4 4 5 6 7 7 7</td>
</tr>
<tr>
<td>3 0 0 1 1 6 9</td>
</tr>
<tr>
<td>4 7</td>
</tr>
</tbody>
</table>
Question 2
For these 18 countries, the lowest female smoking rate is

A 5%  B 7%  C 9%  D 15%  E 19%

[VCAA 2009 1CQ1]

Question 3
For these 18 countries, the interquartile range (IQR) of the female smoking rates is

A 4  B 6  C 19  D 22  E 23

[VCAA 2009 1CQ2]

Question 4
For these 18 countries, the smoking rates for females are generally

A lower and less variable than the smoking rates for males.
B lower and more variable than the smoking rates for males.
C higher and less variable than the smoking rates for males.
D higher and more variable than the smoking rates for males.
E about the same as the smoking rates for males.

[VCAA 2009 1CQ3]

Use the following information to answer Questions 5–7.
The back-to-back ordered stem plot shows the distribution of maximum temperatures (in °Celsius) of two towns, Beachside and Flattown, over 21 days in January.

<table>
<thead>
<tr>
<th>Beachside</th>
<th></th>
<th>Flattown</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>9 8 7 5</td>
<td>1 8 9</td>
<td></td>
</tr>
<tr>
<td>4 3 2 2 1</td>
<td>1 1 0 0 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 9 8 7 6</td>
<td>2 8 9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 2 3 3 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 3 5 5 6</td>
<td>7 7 7 8 8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 0 0 1 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 5 6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Question 5
The variables temperature (°Celsius) and town (Beachside or Flattown) are

A both categorical variables.
B both numerical variables.
C categorical and numerical variables respectively.
D numerical and categorical variables respectively.
E neither categorical nor numerical variables.

[VCAA 2006 1CQ1]
Question 6

For Beachside, the range of maximum temperatures is

A 3°C  B 23°C  C 32°C  D 33°C  E 38°C

[VCAA 2006 1CQ2]

Question 7

The distribution of maximum temperatures for Flattown is best described as


[VCAA 2006 1CQ3]

Use the following information to answer Questions 8 & 9.

Samples of jellyfish were selected from two different locations, A and B. The diameter (in mm) of each jellyfish was recorded and the resulting data is summarised in the boxplots shown.

Question 8

The percentage of jellyfish taken from location A with a diameter greater than 14 mm is closest to

A 2%  B 5%  C 25%  D 50%  E 75%

[VCAA 2007 1CQ5]

Question 9

From the boxplots, it can be concluded that the diameters of the jellyfish taken from location A are generally

A similar to the diameters of the jellyfish taken from location B.  B less than the diameters of the jellyfish taken from location B and less variable.  C less than the diameters of the jellyfish taken from location B and more variable.  D greater than the diameters of the jellyfish taken from location B and less variable.  E greater than the diameters of the jellyfish taken from location B and more variable.

[VCAA 2007 1CQ6]
Question 10

As part of an experiment, three samples of pine trees were planted. Each sample contained 50 trees.

One sample was grown under hot conditions, one sample was grown under mild conditions and one sample was grown under cool conditions.

The parallel boxplots show the rate of growth (in centimetres per year) of these three samples.

From the parallel boxplots it can be concluded that, as conditions change from hot to mild to cool, the rate of growth for these trees

A decreases on average and becomes less variable.
B decreases on average and becomes more variable.
C does not change on average but becomes more variable.
D increases on average and becomes less variable.
E increases on average and becomes more variable.

[VCAA 2005 1CQ7]

Question 11

The boxplots display the distribution of average pay rates, in dollars per hour, earned by workers in 35 countries for the years 1980, 1990 and 2000.

Based on the information contained in the boxplots, which one of the following statements is not true?

A In 1980, over 50% of the countries had an average pay rate of less than $8.00 per hour.
B In 1990, over 75% of the countries had an average pay rate of greater than $5.00 per hour.
C In 1990, the average pay rate in the top 50% of the countries was higher than the average pay rate for any of the countries in 1980.
D In 1990, over 50% of the countries had an average pay rate less than the median average pay rate in 2000.
E In 2000, over 75% of the countries had an average pay rate greater than the median average pay rate in 1980.

[VCAA 2011 1CQ5]
A weather station records the wind speed and the wind direction each day at 9.00 a.m. The wind speed is recorded correct to the nearest whole number.

The parallel boxplots have been constructed from data that was collected on the 214 days from June to December in 2011.

a  Complete the following statements.

The wind direction with the lowest recorded wind speed was _______.

The wind direction with the largest range of recorded wind speeds was _______.  1 mark

b  The wind blew from the south on eight days.

Reading from the parallel boxplots we know that, for these eight wind speeds, the

First quartile $Q_1 = 2$ km/h

Median $M = 3.5$ km/h

Third quartile $Q_3 = 4$ km/h

Given that the eight wind speeds were recorded to the nearest whole number, write down the eight wind speeds.  1 mark

[VCAA 2012 2CQ3]
The arm spans (in cm) of Year 6, 8 and 10 girls were recorded in a survey. The results are summarised in the three parallel box plots displayed below.

![Box plots showing arm spans for Year 6, 8, and 10 girls.](image)

**a** Complete the following sentence.

The middle 50% of Year 6 students have an arm span between ___ and ___ cm. 1 mark

**b** The three parallel boxplots suggest that arm span and Year level are associated. Explain why. 1 mark

**c** The arm span of 110 cm for a Year 10 girl is shown as an outlier on the boxplot. This value is an error. Her real arm span is 140 cm. If the error is corrected, would this girl's arm span still show as an outlier on the boxplot? Give reasons for your answer, showing an appropriate calculation. 2 marks

[VCAA 2008 2CQ3]
Chapter 4: Data distributions

The mean and standard deviation

4.7

The mean

The mean is what is often referred to in everyday life as the average. The symbol for the mean of a set of data is \( \bar{x} \) (called ‘x bar’).

The mean for a list of data values:

\[
\bar{x} = \frac{\text{sum of all the values}}{\text{number of values}} = \frac{\sum x}{n}
\]

\( \Sigma \) means ‘sum of’.

The mean for data in a frequency table:

\[
\bar{x} = \frac{\text{sum of (each value } \times \text{ its frequency)}}{\text{sum of frequencies}} = \frac{\sum fx}{\Sigma f}
\]

Worked example 10

The scores for the players in a nine-hole golf competition are represented in the following frequency table. Calculate the mean score, correct to one decimal place.

<table>
<thead>
<tr>
<th>Score (x)</th>
<th>Frequency (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>2</td>
</tr>
<tr>
<td>38</td>
<td>4</td>
</tr>
<tr>
<td>39</td>
<td>7</td>
</tr>
<tr>
<td>40</td>
<td>4</td>
</tr>
<tr>
<td>41</td>
<td>1</td>
</tr>
</tbody>
</table>

Working

1. Add an extra column to the table and an extra row for totals. Fill in the \( f \times x \) column and totals.

<table>
<thead>
<tr>
<th>Score (x)</th>
<th>Frequency (f)</th>
<th>( f \times x )</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>2</td>
<td>74</td>
</tr>
<tr>
<td>38</td>
<td>4</td>
<td>152</td>
</tr>
<tr>
<td>39</td>
<td>7</td>
<td>273</td>
</tr>
<tr>
<td>40</td>
<td>4</td>
<td>160</td>
</tr>
<tr>
<td>41</td>
<td>1</td>
<td>41</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>18</strong></td>
<td><strong>700</strong></td>
</tr>
</tbody>
</table>

\[
\bar{x} = \frac{\sum fx}{\Sigma f}
\]

\[
\bar{x} = \frac{700}{18} = 38.9
\]

The mean score for the nine-hole golf competition was 38.9.
The median vs the mean

Both the median and the mean are measures of the centre of a distribution. While the median is the midpoint of a distribution, the mean is the balance-point of the distribution.

When looking at the median vs the mean, be aware of the following.

- For symmetric distributions, the median = the mean.
- For distributions that are approximately symmetric, the median and the mean will be approximately equal.
- The mean is greater than the median for positively skewed distributions.
- The mean is less than the median for negatively skewed distributions.
- Outliers don’t generally affect the median, but they can significantly affect the mean.

Exam hack

<table>
<thead>
<tr>
<th>How do you choose whether to use the median or mean as the measure of the centre of a distribution?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approximately symmetric distributions with no outliers</td>
</tr>
<tr>
<td>Approximately symmetric distributions with outliers</td>
</tr>
<tr>
<td>Skewed distributions</td>
</tr>
</tbody>
</table>
The standard deviation

The standard deviation, like the range and the interquartile range, is a measure of the spread of data. Smaller values of the standard deviation indicate less spread. Larger values indicate more spread. While the interquartile range measures the spread of data around the median, the standard deviation measures the spread of data around the mean.

The formula for the sample standard deviation is

\[ s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} \]

although calculations are done quickly through a CAS/calculator.

Using CAS Finding the mean and standard deviation

To find the mean \( \bar{x} \) and the standard deviation \( s \) for a list of data values, follow the same steps as for ‘Using CAS: Five-number summary’ (Section 4.4) and locate \( \bar{x} \) and \( s_x \) from the choices at the end.

To find the mean \( \bar{x} \) and the standard deviation \( s \) (and any of the values for the five-number summary) for data in a frequency table, follow the steps for this frequency table example. Write your answer rounded to two decimal places.

<table>
<thead>
<tr>
<th>Score</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>

TI-Nspire CAS

**STEP 1**

Open a New Document with a Lists & Spreadsheet page. Name column A ‘score’ and column B ‘freq’, then enter the scores into the list called ‘score’ and the corresponding frequencies into the list called ‘freq’.

**STEP 2**


In the pop-up screen that appears, set the number of lists to 1 and press enter.
STEP 3
Select variable name ‘score’ for X1 List and select ‘freq’ for Frequency List, then press \( \text{tab} \), down to \( \text{OK} \) and press \( \text{enter} \).

STEP 4
Write the answer rounded to two decimal places.
\[
\bar{x} = 3.06 \\
s = 1.61
\]

CLASSPAD

STEP 1
Use the \( \text{Statistics} \) application. Rename list1 as ‘score’ by first tapping the list1 cell. To access the letter keyboard, press \( \text{Keyboard} \) and tap \( \text{abc} \).

Follow a similar process to rename the list2 column as ‘freq’ and enter the values.
Tap \( \text{Calc} \) then \( \text{One-Variable} \).

STEP 2
In the pop-up window that appears, set the \( \text{XList} \) to main\(\text{\textbackslash score} \) and \( \text{Freq} \) to main\(\text{\textbackslash freq} \), then tap \( \text{OK} \).

STEP 3
Write the answer rounded to two decimal places.
\[
\bar{x} = 3.06 \\
s = 1.61
\]
The mean and standard deviation

**Prep 1**  
**WORKED EXAMPLE 10**

Students were surveyed about the number of movies they had downloaded in the last six months. The results are shown in the frequency table.

<table>
<thead>
<tr>
<th>Movies (x)</th>
<th>Frequency (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
</tr>
</tbody>
</table>

**a** How many students were surveyed?

**b** Calculate the mean number of movies downloaded, correct to one decimal place.

**Prep 2**  
**USING CAS: FINDING THE MEAN AND STANDARD DEVIATION**

Find the mean and standard deviation of each of the following data sets. Answer correct to 2 decimal places.

| a | 26 27 23 24 27 25 23 |
|   | 19 25 21 22 20 27 21 |

<table>
<thead>
<tr>
<th>b</th>
<th>Score</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>c</th>
<th>Score</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>23</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>24</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>25</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>26</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>27</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>28</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>29</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>30</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>31</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>32</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
</tr>
</tbody>
</table>

**Age (years)**

**Prep 3**

Find the mean, median, mode, range, interquartile range and standard deviation of the following data. Where necessary, round answers to one decimal place.

23, 28, 29, 25, 26, 25, 29, 28, 22, 24, 21, 31, 32, 24, 27, 24, 26
The mean and standard deviation

Use the following information to answer Questions 1 & 2.

Joel is training for a triathlon. He swam the following times, in minutes, in his last ten races.
28 34 22 24 25 24 26 26 24 27

Question 1
Joel's mean swim time is
A. 24 min.  B. 25 min.  C. 25.5 min.  D. 26 min.  E. 27 min.

Question 2
If the outlier score of 34 was removed from the set of data, which of the following would happen?
A. The mean would stay the same.  B. The mean would have 34 added to it.
C. The mean would decrease.  D. The mean would increase.
E. None of the above.

Question 3
The total weight of nine oranges is 1.53 kg.
Using this information, the mean weight of an orange would be calculated to be closest to
A. 115 g  B. 138 g  C. 153 g  D. 162 g  E. 170 g

Question 4
A sample of 14 people were asked to indicate the time (in hours) they had spent watching television on the previous night. The results are displayed in the dot plot.
Correct to one decimal place, the mean and standard deviation of these times are respectively
A. $\bar{x} = 2.0, \ s = 1.5$  B. $\bar{x} = 2.1, \ s = 1.5$  C. $\bar{x} = 2.1, \ s = 1.6$
D. $\bar{x} = 2.6, \ s = 1.2$  E. $\bar{x} = 2.6, \ s = 1.3$
Use the following information to answer Questions 5 & 6.

The number of DVD players in each of 20 households is recorded in the frequency table.

<table>
<thead>
<tr>
<th>Number of DVD players</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>20</strong></td>
</tr>
</tbody>
</table>

**Question 5**

For this sample of households, the percentage of households with at least one DVD player is

A 30%  \hspace{1cm} B 45%  \hspace{1cm} C 50%  \hspace{1cm} D 70%  \hspace{1cm} E 90%  

[VCAA 2004 1CQ4]

**Question 6**

For this sample of households, the mean number of DVD players in these 20 households is

A 0.75  \hspace{1cm} B 1.00  \hspace{1cm} C 1.15  \hspace{1cm} D 1.64  \hspace{1cm} E 2.00  

[VCAA 2004 1CQ5]

**Question 7**

The boxplot shows the distribution of the time, in seconds, that 79 customers spent moving along a particular aisle in a large supermarket.

From the boxplot, it can be concluded that the median time spent moving along the supermarket aisle is

A less than the mean time.  \hspace{1cm} B equal to the mean time.  
C greater than the mean time.  \hspace{1cm} D half of the interquartile range.  
E one quarter of the range.  

[VCAA 2008 1CQ4]
Question 8

The table shows the percentage of women ministers in the parliaments of 22 countries in 2008.

<table>
<thead>
<tr>
<th>Country</th>
<th>Percentage of women ministers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norway</td>
<td>56</td>
</tr>
<tr>
<td>Sweden</td>
<td>48</td>
</tr>
<tr>
<td>France</td>
<td>47</td>
</tr>
<tr>
<td>Spain</td>
<td>44</td>
</tr>
<tr>
<td>Switzerland</td>
<td>43</td>
</tr>
<tr>
<td>Austria</td>
<td>38</td>
</tr>
<tr>
<td>Denmark</td>
<td>37</td>
</tr>
<tr>
<td>Iceland</td>
<td>36</td>
</tr>
<tr>
<td>Germany</td>
<td>33</td>
</tr>
<tr>
<td>Netherlands</td>
<td>33</td>
</tr>
<tr>
<td>New Zealand</td>
<td>32</td>
</tr>
<tr>
<td>Australia</td>
<td>24</td>
</tr>
<tr>
<td>Italy</td>
<td>24</td>
</tr>
<tr>
<td>United States</td>
<td>24</td>
</tr>
<tr>
<td>Belgium</td>
<td>23</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>23</td>
</tr>
<tr>
<td>Ireland</td>
<td>21</td>
</tr>
<tr>
<td>Liechtenstein</td>
<td>20</td>
</tr>
<tr>
<td>Canada</td>
<td>16</td>
</tr>
<tr>
<td>Luxembourg</td>
<td>14</td>
</tr>
<tr>
<td>Japan</td>
<td>12</td>
</tr>
<tr>
<td>Singapore</td>
<td>0</td>
</tr>
</tbody>
</table>

a  What proportion of these 22 countries have a higher percentage of women ministers in their parliament than Australia?

b  Determine the median, range and interquartile range of this data.

c  The ordered stem plot displays the distribution of the percentage of women ministers in parliament for 21 of these countries.

   The value for Canada is missing.

   Copy and complete the stem plot by adding the value for Canada.

d  Both the median and the mean are appropriate measures of centre for this distribution. Explain why.

<table>
<thead>
<tr>
<th>Stem (10s)</th>
<th>Leaf (units)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>2 4</td>
</tr>
<tr>
<td>2</td>
<td>0 1 3 3 4 4 4</td>
</tr>
<tr>
<td>3</td>
<td>2 3 3 6 7 8</td>
</tr>
<tr>
<td>4</td>
<td>3 4 7 8</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

[VCAA 2010 2CQ1]

Question 9

This dot plot shows the number of vehicles driving past Westvale High School every minute for a 20-minute period.

a  Find the mean.

b  Calculate the standard deviation, correct to two decimal places.

c  How many scores were within one standard deviation of the mean?
**Types of data**

**Categorical data**
Data that can be sorted into categories or groups.
*Ask: Is the number being used like a name?*
- **YES**

**Numerical data**
Data that can be counted or measured to an increasing level of accuracy.
*Ask: Is the number being used like a name?*
- **NO**

**Nominal data**
Categorical data that doesn’t have a natural order, even when numbers are involved.
e.g. Colour of first car this morning
Postcodes of home address
Numbers on basketball player uniforms
*Ask: Does it make sense to order them?*
- **NO**

**Discrete data**
Numerical data that can be counted
e.g. The number of mobile phones in home
The number of people who like Vegemite
The amount of money spent in a shop
*Ask: Does it involve counting something?*
- **YES**

**Ordinal data**
Categorical data that has a natural order, but doesn’t involve counting anything and can’t be measured to an increasing level of accuracy.
e.g. Ratings given to the latest Star Wars movie
(1 = great, 2 = okay, 3 = awful, 4 = didn’t see it)
House numbers
Shoe sizes
Grades in a test (A, B, C, D, E)
Clothing sizes (small, medium, large, extra large)
*Ask: Does it make sense to order them?*
- **YES**

**Continuous data**
Numerical data that can be measured to an increasing level of accuracy
e.g. The age of oldest living pet
Time taken to run 100 metres
The height of players in a basketball team
*Ask: Could I measure this more accurately if I wanted to?*
- **YES**

**Tables, charts and plots**

- **Frequency tables** can be used to display both categorical and numerical data. The data values are listed in one column and the corresponding frequencies are displayed in a frequency column.

- **Bar charts** help us see patterns when dealing with categorical variables. Categories can be represented on the horizontal or vertical axis, with their corresponding frequency on the other axis.

- A **histogram** is a graphical way of displaying numerical data from a frequency table. It is effective when dealing with data that has been grouped into a small number (usually between 5 and 15) of intervals.
  - Histograms for continuous numerical data and grouped discrete numerical data have bars that sit in between the interval values on the horizontal axis.
  - Histograms for discrete numerical data that hasn’t been grouped into intervals have bars starting and ending halfway between scale marks on the horizontal axis.
• **Boxplots** display numerical data based on the five-number summary, IQR and outliers.

![Boxplot Diagram](image)

Lower fence = $Q_1 + 1.5 \times IQR$

Upper fence = $Q_3 + 1.5 \times IQR$

• **Dot plots** are the simplest way to display numerical data. They are best used for a maximum of 50 data values and when the data values are not too spread out.

• **Stem plots** are an alternative to histograms. Their main advantage is that the actual data values appear. A key is always required.

**Distribution shapes**

- Positively skewed distribution
- Negatively skewed distribution
- Symmetric distribution
- Distribution with outliers
Statistical values

- **The range** is a measure of the spread of the data. 
  \[ \text{Range} = \text{largest value} - \text{smallest value} \]

- **The median** is a measure of the centre of the data.
  - The median is the middle value when the data is ordered from smallest to largest.
  - When there are two middle values, we add them and divide by 2 to find the median.
  - The median on a histogram occurs at the vertical line that splits the histogram in half, with equal areas on either side.

- **Quartiles** are the three points that divide a set of data into quarters.
  - The first or lower quartile has 25% of the data below it.
  - The second quartile (which is the same as the median) has 50% of the data below it.
  - The third or upper quartile has 75% of the data below it.

- **The five-number summary** consists of
  1. The minimum data value
  2. \( Q_1 \) (lower quartile)
  3. The median
  4. \( Q_3 \) (upper quartile)
  5. The maximum data value

- **The interquartile range** (or IQR) is the measure of the spread of the middle 50% of the data values. \( \text{IQR} = Q_3 - Q_1 \)

- A data value is a possible **outlier** if it is either
  - less than \( Q_1 - 1.5 \times \text{IQR} \) or
  - greater than \( Q_3 + 1.5 \times \text{IQR} \)

Back-to-back stem plots

- can be used to compare the distribution of a numerical variable across two groups
- have two sets of leaves, one on the left of the stem and one on the right
- are limited to small sets of data

Parallel boxplots

- can be used to compare the distribution of a numerical variable across two or more groups
- make it easy to visually compare medians, quartiles and ranges
- can be used for large data sets
Mean

- The mean is what is often referred to in everyday life as the average.
- The mean is the balance-point of the distribution.
- The symbol for the mean of a set of data is \( \bar{x} \) (called ‘x bar’).

- The mean for a list of data values:
  \[
  \bar{x} = \frac{\text{sum of all the values}}{\text{number of values}} = \frac{\sum x}{n}
  \]

- The mean for data in a frequency table:
  \[
  \bar{x} = \frac{\text{sum of (each value } \times \text{ its frequency)}}{\text{sum of frequencies}} = \frac{\sum fx}{\sum f}
  \]

- When looking at the median vs the mean, be aware of the following.
  - For symmetric distributions, the median = the mean.
  - For distributions that are approximately symmetric, the median and the mean will be approximately equal.
  - The mean is greater than the median for positively skewed distributions.
  - The mean is less than the median for negatively skewed distributions.
  - Outliers don’t generally affect the median, but they can significantly affect the mean.

- Choosing between the median or mean as the measure of the centre of a distribution:

<table>
<thead>
<tr>
<th>Distribution Type</th>
<th>Measure of Centre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approximately symmetric distributions with no outliers</td>
<td>Mean or median</td>
</tr>
<tr>
<td>Approximately symmetric distributions with outliers</td>
<td>Median</td>
</tr>
<tr>
<td>Skewed distributions</td>
<td>Median</td>
</tr>
</tbody>
</table>

Standard deviation

- The standard deviation is a measure of the spread of data.
- While the interquartile range measures the spread of data around the median, the standard deviation measures the spread of data around the mean.

- The formula for the standard deviation is
  \[
  s = \sqrt{\frac{\sum(x - \bar{x})^2}{n-1}}, \text{ although calculations are done quickly through a CAS/calculator.}
  \]
• Introducing Nelson VCE Maths: the only VCE series where every question is exam practice, whose structure and content are based on actual VCE exam questions
• Clear and concise theory and worked examples
• VCE exam questions from 2002 to 2014 integrated into each book
• Contains Exam prep and Exam practice exercises
• Using CAS contains instructions for the TI-Nspire and ClassPad calculators, integrated into all chapters
• Exam hack provides ‘insider advice’ to help students overcome common errors, achieve learning success and gain valuable marks
• Practice exams in Units 3 and 4 cover a number of related topics in an area of study
• Solutions at the back of the book include worked solutions to selected questions, with examiners’ comments and student performance percentages from the VCAA examination reports
• Worked solutions to all questions can be downloaded by teachers*
• All teacher resource material included in the price of the book*

*Please note that complimentary access to NelsonNet is only available to teachers who use the accompanying student textbook as a core educational resource in their classroom. Contact your sales representative for information about access codes and conditions.