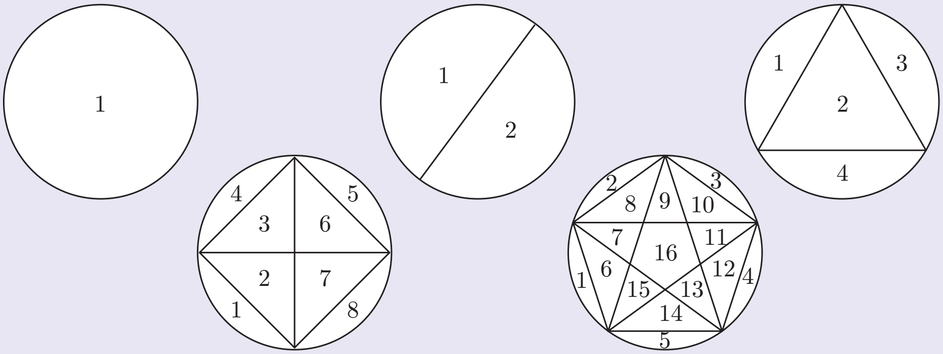
Suppose we place n points around a circle such that when we connect each point with every other point, no three lines intersect at the same point. We then count the number of regions that the circle is divided into.

The first five cases are shown below:



From these cases we *conjecture* that for n points, the circle is divided into 2^{n-1} regions.

Draw the case $n = 6$ and see if the conjecture is true!

3 Is it reasonable for a mathematician to assume a conjecture is true until it has been formally proven?

F COMPOUND INTEREST

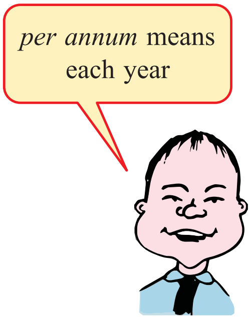
When money is deposited in a bank, it will usually earn **compound interest**.

After a certain amount of time called the **period**, the bank adds money to the account which is a percentage of the money already in there. The amount added is called the **interest**.

It is called *compound* interest because the interest generated in one period will itself earn more interest in the next period.

COMPOUND INTEREST

Suppose you invest \$1000 in the bank. You leave the money in the bank for 3 years, and are paid an interest rate of 10% per annum (p.a). The interest is added to your investment each year, so the total value *increases*.



The percentage increase each year is 10%, so at the end of the year you will have $100\% + 10\% = 110\%$ of the value at its start. This corresponds to a *multiplier* of 1.1.

After one year your investment is worth $\$1000 \times 1.1 = \$1100$.

After two years it is worth

$$\begin{aligned} &\$1100 \times 1.1 \\ &= \$1000 \times 1.1 \times 1.1 \\ &= \$1000 \times (1.1)^2 = \$1210 \end{aligned}$$

After three years it is worth

$$\begin{aligned} &\$1210 \times 1.1 \\ &= \$1000 \times (1.1)^2 \times 1.1 \\ &= \$1000 \times (1.1)^3 = \$1331 \end{aligned}$$

We have a geometric sequence with first term 1000 and common ratio 1.1.

If the money is left in your account for n years it will amount to $\$1000 \times (1.1)^n$.

THE COMPOUND INTEREST FORMULA

For interest compounding annually, $\boldsymbol{FV = PV \times \left(1 + \frac{r}{100}\right)^n}$

where: FV is the **future value** or final balance
PV is the **present value** or amount originally invested
r is the **interest rate per year**
n is the **number of years**

Example 18 🔊 **Self Tutor**

\$5000 is invested for 4 years at 7% p.a. compound interest, compounded annually. What will it amount to at the end of this period? Give your answer to the nearest cent.

$PV = 5000$
$r = 7$
$n = 4$

$$\begin{aligned} FV &= PV \times \left(1 + \frac{r}{100}\right)^n \\ &= 5000 \times \left(1 + \frac{7}{100}\right)^4 \\ &\approx 6553.98 \end{aligned}$$

The investment amounts to \$6553.98.

DIFFERENT COMPOUNDING PERIODS

Interest can be compounded more than once per year. Interest is commonly compounded:

- half-yearly (2 times per year)
- quarterly (4 times per year)
- monthly (12 times per year).

For interest compounding k times per year, $\boldsymbol{FV = PV \times \left(1 + \frac{r}{100k}\right)^{kn}}$

Example 19 🔊 **Self Tutor**

Calculate the final balance of a \$10 000 investment at 6% p.a. where interest is compounded quarterly for two years.

$PV = 10\,000$
$r = 6$
$n = 2$
$k = 4$
$\therefore \quad kn = 8$

$$\begin{aligned} FV &= PV \times \left(1 + \frac{r}{100k}\right)^{kn} \\ &= 10\,000 \times \left(1 + \frac{6}{400}\right)^8 \\ &\approx 11\,264.93 \end{aligned}$$

The final balance is \$11 264.93.

INTEREST EARNED

The **interest earned** is the difference between the original balance and the final balance.

$$\textbf{Interest} = \boldsymbol{FV - PV}$$

Example 20 Self Tutor

How much interest is earned if €8800 is placed in an account that pays $4\frac{1}{2}\%$ p.a. compounded monthly for $3\frac{1}{2}$ years?

$PV = 8800, \quad r = 4.5, \quad n = 3.5, \quad k = 12$

$\therefore \quad kn = 12 \times 3\frac{1}{2} = 42$

Now $FV = PV \times \left(1 + \frac{r}{100k}\right)^{kn}$

$= 8800 \times \left(1 + \frac{4.5}{1200}\right)^{42}$

$\approx 10\,298.08$

The interest earned $= FV - PV$

$= 10\,298.08 - 8800$

$= 1498.08$

The interest earned is €1498.08.

EXERCISE 5F.1

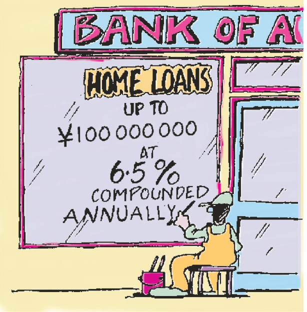
1 Find the final value of a compound interest investment of \$6000 after 3 years at 5% p.a., with interest compounded annually.

2 Luisa invests £15 000 into an account which pays 8% p.a. compounded annually. Find:

a the value of her account after 2 years
b the total interest earned after 2 years.

3 Yumi places 880 000 yen in a fixed term investment account which pays 6.5% p.a. compounded annually.

a How much will she have in her account after 6 years?
b What interest has she earned over this period?



4 Ali places £9000 in a savings account that pays 8% p.a. compounded quarterly. How much will she have in the account after 5 years?

5 How much interest would be earned on a deposit of \$2500 at 5% p.a. compounded half yearly for 4 years?

6 Jai recently inherited \$92 000. He decides to invest it for 10 years before he spends any of it. The three banks in his town offer the following terms:

Bank A: $5\frac{1}{2}\%$ p.a. compounded yearly.

Bank B: $5\frac{1}{4}\%$ p.a. compounded quarterly.

Bank C: 5% p.a. compounded monthly.

Which bank offers Jai the greatest interest on his inheritance?

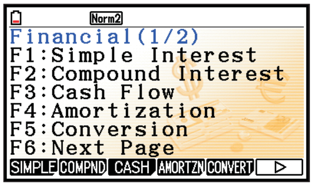
USING A GRAPHICS CALCULATOR FOR COMPOUND INTEREST PROBLEMS

Most graphics calculator have an in-built **finance program** that can be used to investigate financial scenarios. This is called a **TVM Solver**, where **TVM** stands for **time value of money**.

To access the TVM Solver:

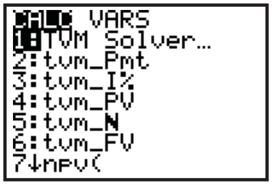
Casio fx-CG20

Select **Financial** from the Main Menu, then press F2 : **Compound Interest**.



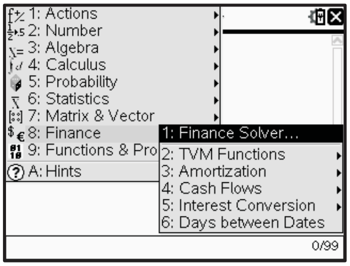
TI-84 Plus

Press APPS , then select **1 : Finance...** and **1 : TVM Solver...** .



TI-$n$spire

From the Calculator application, press menu , then select **8 : Finance > 1 : Finance Solver...** .



The TVM Solver can be used to find any variable if all the other variables are given. For the **TI-84 Plus**, the abbreviations used are:

- N represents the **number of time periods**
- $I\%$ represents the **interest rate per year**
- PV represents the **present value** of the investment
- PMT represents the **payment each time period**
- FV represents the **future value** of the investment
- P/Y is the **number of payments per year**
- C/Y is the **number of compounding periods per year**
- PMT : END BEGIN lets you choose between the payments at the end of a time period or at the beginning of a time period. Most interest payments are made at the end of the time periods.

The abbreviations used by the other calculator models are similar, and can be found in the **graphics calculator instructions** on the CD.

GRAPHICS CALCULATOR INSTRUCTIONS

Example 21

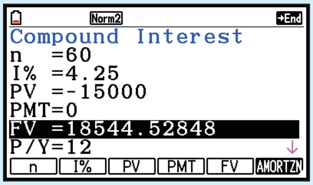
Self Tutor

Holly invests 15 000 UK pounds in an account that pays 4.25% p.a. compounded monthly. How much is her investment worth after 5 years?

To answer this using the TVM function on the calculator, first set up the TVM screen. The initial investment is considered as an outgoing and is entered as a negative value.

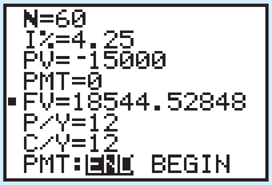
There are $5 \times 12 = 60$ month periods.

Casio fx-CG20



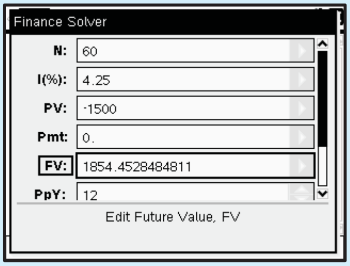
```
Compound Interest
n   =60
I%  =4.25
PV  =-15000
PMT=0
FV  =18544.52848
P/Y=12
```

TI-84 Plus



```
N=60
I%=4.25
PV=-15000
PMT=0
FV=18544.52848
P/Y=12
C/Y=12
PMT:END BEGIN
```

TI-*n*spire



```
Finance Solver
N:     60
I(%):  4.25
PV:    -1500
Pmt:   0.
FV:    1854.4528484811
PpY:   12
Edit Future Value, FV
```

Holly's investment is worth 18 544.53 UK pounds after 5 years.

In IB examinations, a correct list of entries for the TVM Solver will be awarded the method mark.

For the previous example you would write:

$$\begin{aligned} N &= 60 \\ I &= 4.25 \\ PV &= -15\,000 \\ C/Y &= 12 \\ \Rightarrow \quad FV &= 18\,544.53 \end{aligned}$$

So, Holly's investment is worth £18 544.53 .

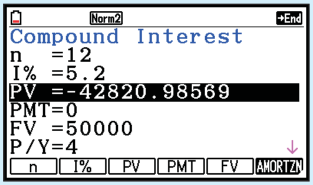
Example 22

Self Tutor

How much does Halena need to deposit into an account to collect $50 000 at the end of 3 years if the account is paying 5.2% p.a. compounded quarterly?

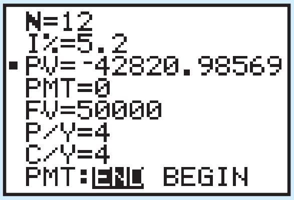
Set up the TVM screen as shown.
There are $3 \times 4 = 12$ quarter periods.

Casio fx-CG20



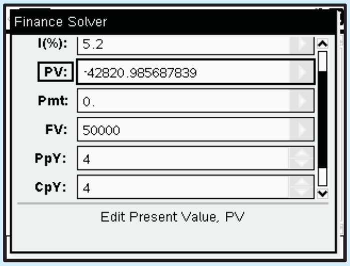
```
Compound Interest
n   =12
I%  =5.2
PV  =-42820.98569
PMT=0
FV  =50000
P/Y=4
```

TI-84 Plus



```
N=12
I%=5.2
PV=-42820.98569
PMT=0
FV=50000
P/Y=4
C/Y=4
PMT:END BEGIN
```

TI-*n*spire



```
Finance Solver
I(%):  5.2
PV:    -42820.985687839
Pmt:   0.
FV:    50000
PpY:   4
CpY:   4
Edit Present Value, PV
```

Thus, $42 821 needs to be deposited.

EXERCISE 5F.2

1 Use technology to check your answer to **Exercise 5F.1**, question **4**.

2 If I deposit £6000 in a bank account that pays 5% p.a. compounded monthly, how much will I have in my account after 2 years?

3 When my child was born I deposited $2000 in a bank account paying 4% p.a. compounded half-yearly. How much will my child receive on her 18th birthday?

4 Calculate the compound interest earned on an investment of €13 000 for 4 years if the interest rate is 7% p.a. compounded quarterly.

5 Calculate the amount you would need to invest now in order to accumulate 250 000 yen in 5 years' time, if the interest rate is 4.5% p.a. compounded monthly.

6 You would like to buy a car costing $23 000 in two years' time. Your bank account pays 5% p.a. compounded half-yearly. How much do you need to deposit now in order to be able to buy your car in two years?

7 You have just won the lottery and decide to invest the money. Your accountant advises you to deposit your winnings in an account that pays 6.5% p.a. compounded annually. After four years your winnings have grown to €102 917.31. How much did you win in the lottery?

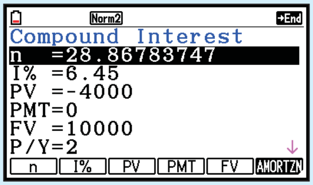
Example 23

Self Tutor

For how long must Magnus invest €4000 at 6.45% p.a. compounded half-yearly for it to amount to €10 000?

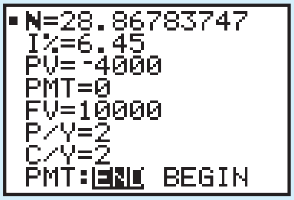
Set up the TVM screen as shown. We then need to find n, the number of periods required.

Casio fx-CG20



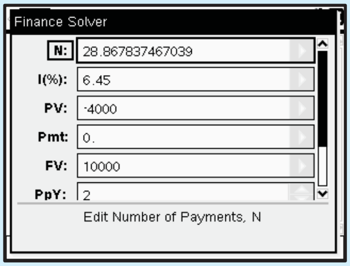
```
Norm2
Compound Interest
n   =28.86783747
I%  =6.45
PV  =-4000
PMT=0
FV  =10000
P/Y=2
```

TI-84 Plus



```
N=28.86783747
I%=6.45
PV=-4000
PMT=0
FV=10000
P/Y=2
C/Y=2
PMT:END BEGIN
```

TI-*n*spire



```
Finance Solver
N:     28.867837467039
I(%):  6.45
PV:    -4000
Pmt:   0.
FV:    10000
PpY:   2
Edit Number of Payments, N
```

$n \approx 28.9$, so 29 half-years are required, or 14.5 years.

8 Your parents give you $8000 to buy a car, but the car you want costs $9200. You deposit the $8000 into an account that pays 6% p.a. compounded monthly. How long will it be before you have enough money to buy the car you want?

9 A couple inherited €40 000 and deposited it in an account paying $4\frac{1}{2}\%$ p.a. compounded quarterly. They withdrew the money as soon as they had over €45 000. How long did they keep the money in that account?

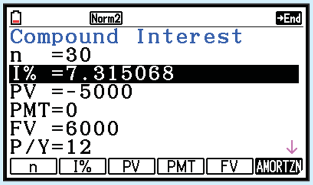
10 A business deposits £80 000 in an account that pays $5\frac{1}{4}\%$ p.a. compounded monthly. How long will it take before they double their money?

Example 24 **Self Tutor**

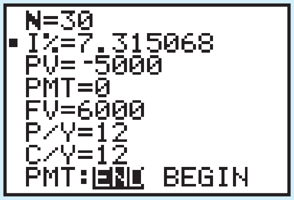
Iman deposits \$5000 in an account that compounds interest monthly. 2.5 years later the account totals \$6000. What annual rate of interest was paid?

Set up the TVM screen as shown. In this case $n = 2.5 \times 12 = 30$ months.

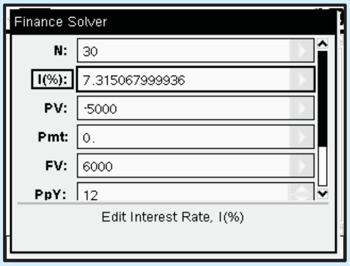
Casio fx-CG20


```
Compound Interest
n   =30
I% =7.315068
PV =-5000
PMT=0
FV =6000
P/Y=12
```

TI-84 Plus


```
N=30
▪I%=7.315068
PV=-5000
PMT=0
FV=6000
P/Y=12
C/Y=12
PMT:END BEGIN
```

TI-*n*spire


Finance Solver

N:	30
I(%):	7.315067999936
PV:	-5000
Pmt:	0.
FV:	6000
PpY:	12

Edit Interest Rate, I(%)

An annual interest rate of 7.32% p.a. is required.

11 An investor purchases rare medals for \$10 000 and hopes to sell them 3 years later for \$15 000. What must the annual increase in the value of the medals be over this period, in order for the investor's target to be reached?

12 I deposited €5000 into an account that compounds interest monthly, and $3\frac{1}{2}$ years later the account totals €6165. What annual rate of interest did the account pay?

13 A young couple invests their savings of 900 000 yen in an account where the interest is compounded annually. Three years later the account balance is 1 049 322 yen. What interest rate has been paid?

G DEPRECIATION

Assets such as computers, cars, and furniture lose value as time passes. This is due to wear and tear, technology becoming old, fashions changing, and other reasons. We say that they **depreciate** over time.

Depreciation is the loss in value of an item over time.

Suppose a truck is bought for \$36 000, and depreciates at 25% each year.

Each year, the truck is worth $100\% - 25\% = 75\%$ of its previous value.

We therefore have a geometric sequence with initial value \$36 000 and common ratio 0.75 .

After 1 year, the value is $\$36\,000 \times 0.75 = \$27\,000$

After 2 years, the value is $\$36\,000 \times 0.75^2 = \$20\,250$

After n years, the value is $\$36\,000 \times 0.75^n$.

When calculating depreciation, the **annual multiplier** is $\left(1 + \frac{r}{100}\right)$, where r is the *negative* annual depreciation rate as a percentage.

The **depreciation formula** is $\quad FV = PV \times \left(1 + \frac{r}{100}\right)^n$

where FV is the **future value** after n time periods
PV is the **original purchase value**
r is the **depreciation rate per period** and r is **negative**
n is the **number of periods**.

Example 25 **Self Tutor**

An industrial dishwasher was purchased for £2400 and depreciated at 15% each year.

a Find its value after six years. **b** By how much did it depreciate?

a $PV = 2400$, $r = -15$, $n = 6$

Now $FV = PV \times \left(1 + \frac{r}{100}\right)^n$

$$= 2400 \times (1 - 0.15)^6$$
$$= 2400 \times (0.85)^6$$
$$\approx 905.16$$

So, after 6 years the value is £905.16.

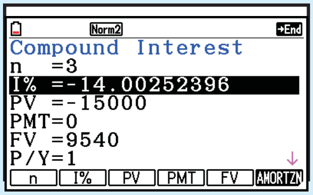
b Depreciation = £2400 − £905.16 = £1494.84

Example 26 **Self Tutor**

A vending machine bought for \$15 000 is sold 3 years later for \$9540. Calculate its annual rate of depreciation.

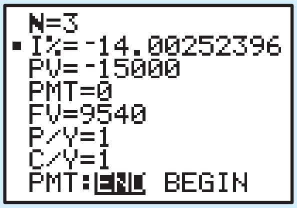
Set up the TVM screen with $N = 3$, $PV = -15\,000$, $PMT = 0$, $FV = 9540$, $P/Y = 1$, $C/Y = 1$.

Casio fx-CG20



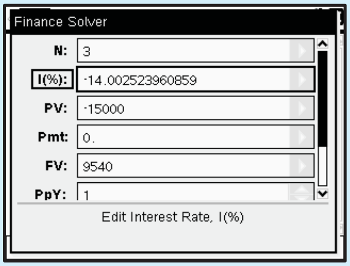
```
Norm2                    →End
Compound Interest
n   =3
I%  =-14.00252396
PV  =-15000
PMT=0
FV  =9540
P/Y=1
 n | I% | PV | PMT | FV | AMORTZN
```

TI-84 Plus



```
 N=3
▪I%=-14.00252396
 PV=-15000
 PMT=0
 FV=9540
 P/Y=1
 C/Y=1
 PMT:END BEGIN
```

TI-*n*spire



```
Finance Solver
   N: 3
I(%): -14.002523960859
  PV: -15000
 Pmt: 0.
  FV: 9540
 PpY: 1
Edit Interest Rate, I(%)
```

The annual depreciation rate is 14.0%.

EXERCISE 5G

1 A lathe, purchased by a workshop for €2500, depreciates by 20% each year. Find the value of the lathe after 3 years.

2 A tractor was purchased for €110 000, and depreciates at 25% p.a. for 5 years.

a Find its value at the end of this period. **b** By how much did it depreciate?

3 **a** I buy a laptop for ¥87 500 and keep it for 3 years. During this time it depreciates at an annual rate of 30%. What will its value be after 3 years?

b By how much has the laptop depreciated?

4 A printing press costing £250 000 was sold 4 years later for £80 000. At what yearly rate did it depreciate in value?

5 A 4-wheel-drive vehicle was purchased for \$45 000 and sold for \$28 500 after 2 years and 3 months. Find its annual rate of depreciation.

REVIEW SET 5A

1 Identify the following sequences as arithmetic, geometric, or neither:

a $7, -1, -9, -17, ...$ **b** $9, 9, 9, 9, ...$ **c** $4, -2, 1, -\frac{1}{2}, ...$

d $1, 1, 2, 3, 5, 8, ...$ **e** the set of all multiples of 4 in ascending order.

2 Find k if $3k$, $k-2$, and $k+7$ are consecutive terms of an arithmetic sequence.

3 Show that $28, 23, 18, 13, ...$ is an arithmetic sequence. Hence find u_n and the sum S_n of the first n terms in simplest form.

4 Find k given that 4, k, and $k^2 - 12$ are consecutive terms of a geometric sequence.

5 Determine the general term of a geometric sequence given that its sixth term is $\frac{16}{3}$ and its tenth term is $\frac{256}{3}$.

6 Insert six numbers between 23 and 9 so that all eight numbers are in arithmetic sequence.

7 Find the 8th term of each of the following sequences:

a $5, 1, \frac{1}{5}, ...$ **b** $-11, -8\frac{1}{2}, -6, ...$ **c** $a, a-d, a-2d, ...$

8 At the start of the dry season, Yafiah's 3000 L water tank is full. She uses 183 L of water each week to water her garden.

a Find the amount of water left in the tank after 1, 2, 3, and 4 weeks.

b Explain why the amount of water left in the tank after n weeks forms an arithmetic sequence.

c When does Yafiah's tank run out of water?

9 Find the sum of:

a $14 + 11 + 8 + ... + (-55)$ **b** $3 + 15 + 75 + ...$ to 10 terms

10 Consider the arithmetic sequence $12, 19, 26, 33, ...$

a Find the 8th term of the sequence.

b Find the sum of the first 10 terms of the sequence.

c The sum of the first n terms is 915. Find the value of n.

11 Val receives a $285 000 superannuation payment when she retires. She finds the following investment rates are offered:

Bank A : 6% p.a. compounded quarterly *Bank B* : $5\frac{3}{4}$% p.a. compounded monthly.

Compare the interest that would be received from these banks over a ten year period. In which bank should Val deposit her superannuation?

12 Sven sells his stamp collection and deposits the proceeds of $8700 in a term deposit account for nine months. The account pays $9\frac{3}{4}$% p.a. compounded monthly. How much interest will he earn over this period?

13 **a** Find the future value of a truck which is purchased for $135 000 and depreciates at 15% p.a. for 5 years.

b By how much does it depreciate?

14 Ena currently has £7800, and wants to buy a car valued at £9000. She puts her money in an account paying 4.8% p.a. compounded quarterly. When will she be able to buy the car?

REVIEW SET 5B

1 A sequence is defined by $u_n = 6(\frac{1}{2})^{n-1}$.

a Prove that the sequence is geometric. **b** Find u_1 and r.

c Find the 16th term to 3 significant figures.

2 Consider the sequence $24, 23\frac{1}{4}, 22\frac{1}{2}, ..., -36$. Find:

a the number of terms in the sequence. **b** the value of u_{35} for the sequence.

c the sum of the terms in the sequence.

3 Find the sum of:

a $3 + 9 + 15 + 21 + ...$ to 23 terms **b** $24 + 12 + 6 + 3 + ...$ to 12 terms.

4 List the first five terms of the sequence:

a $\{(\frac{1}{3})^n\}$ **b** $\{12 + 5n\}$ **c** $\left\{\dfrac{4}{n+2}\right\}$

5 **a** What will an investment of €6000 at 7% p.a. compound interest amount to after 5 years?

b What part of this is interest?

6 Find the first term of the sequence $24, 8, \frac{8}{3}, \frac{8}{9}, ...$ which is less than 0.001.

7 A geometric sequence has $u_6 = 24$ and $u_{11} = 768$.

a Determine the general term of the sequence. **b** Hence find u_{17}.

c Find the sum of the first 15 terms.

8 The nth term of a sequence is given by the formula $u_n = 4n - 7$.

a Find the value of u_{10}.

b Write down an expression for $u_{n+1} - u_n$ and simplify it.

c Hence explain why the sequence is arithmetic.

d Evaluate $u_{15} + u_{16} + u_{17} + ... + u_{30}$.

9 **a** Determine the number of terms in the sequence $128, 64, 32, 16, ..., \frac{1}{512}$.

b Find the sum of these terms.

10 For the geometric sequence $180, 60, 20, ...,$ find:

a the common ratio for this sequence. **b** the 6th term of the sequence.

c the least number of terms required for the sum of the terms to exceed 269.9.

11 Before leaving overseas on a three year trip to India, I leave a sum of money in an account that pays 6% p.a. compounded half-yearly. When I return from the trip, there is €5970.26 in my account. How much interest has been added since I have been away?

12 Megan deposits £3700 in an account paying interest compounded monthly for two years. If she ends up with £4072, what rate of interest did Megan receive?

13 Kania purchases office equipment valued at $17 500.

a At the end of the first year, the value of the equipment is $15 312.50. Find the rate of depreciation.

b If the value of the equipment continued to depreciate at the same rate, what would it be worth after $3\frac{1}{2}$ years?

REVIEW SET 5C

1 A sequence is defined by $u_n = 68 - 5n$.

a Prove that the sequence is arithmetic. **b** Find u_1 and d.

c Find the 37th term of the sequence.

d State the first term of the sequence which is less than -200.

2 **a** Show that the sequence $3, 12, 48, 192, ...$ is geometric.

b Find u_n and hence find u_9.

3 Find the general term of the arithmetic sequence with $u_7 = 31$ and $u_{15} = -17$.
Hence, find the value of u_{34}.

4 Consider the sequence $24, a, 6, ...$
Find the value of a if the sequence is: **a** arithmetic **b** geometric.

5 Find the 10th term of the sequence:

a $32, 25, 18, 11, ...$ **b** $\frac{1}{81}, \frac{1}{27}, \frac{1}{9}, \frac{1}{3}, ...$

6 There were originally 3000 koalas on Koala Island. Since then, the population of koalas on the island has increased by 5% each year.

a How many koalas were on the island after 3 years?

b How long will it take for the population to exceed 5000?

7 Find the formula for u_n, the general term of:

a $86, 83, 80, 77, ...$ **b** $\frac{3}{4}, 1, \frac{7}{6}, \frac{9}{7}, ...$ **c** $100, 90, 81, 72.9, ...$

Hint: One of these sequences is neither arithmetic nor geometric.

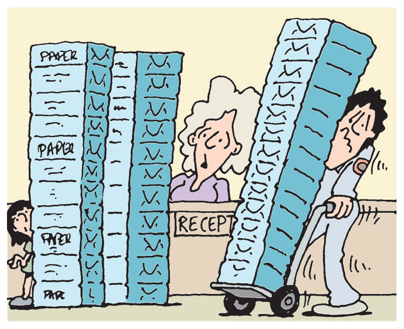
8 Find the first term of the sequence $5, 10, 20, 40, ...$ which exceeds 10 000.

9 $-1,\ k,\ k^2 - 7$ are consecutive terms of an arithmetic sequence. Find k.

10 Each year, a school manages to use only 90% as much paper as the previous year. In the year 2000, they used 700 000 sheets of paper.

a Find how much paper the school used in the years 2001 and 2002.

b How much paper did the school use in total in the decade from 2000 to 2009?



11 Find the final value of a compound interest investment of €8000 after 7 years at 3% p.a. with interest compounded annually.

12 Ned would like to have £15 000 in 3 years' time to install a swimming pool. His bank pays 4.5% p.a. interest, compounded half-yearly. How much does Ned need to deposit now?

13 A motorbike, purchased for £2300, was sold for £1300 after 4 years. Calculate the average annual rate of depreciation.

Chapter 6

Descriptive statistics

Syllabus reference: 2.1, 2.2, 2.3, 2.4, 2.5, 2.6

Contents:

OPENING PROBLEM

A farmer is investigating the effect of a new organic fertiliser on his crops of peas. He has divided a small garden into two equal plots and planted many peas in each. Both plots have been treated the same except that fertiliser has been used on one but not the other.

A random sample of 150 pods is harvested from each plot at the same time, and the number of peas in each pod is counted. The results are:

Without fertiliser

4 6 5 6 5 6 4 6 4 9 5 3 6 8 5 4 6 8 6 5 6 7 4 6 5 2 8 6 5 6 5 5 5 4 4 4 6 7 5 6
7 5 5 6 4 8 5 3 7 5 3 6 4 7 5 6 5 7 5 7 6 7 5 4 7 5 5 5 6 6 5 6 7 5 8 6 8 6 7 6
6 3 7 6 8 3 3 4 4 7 6 5 6 4 5 7 3 7 7 6 7 7 4 6 6 5 6 7 6 3 4 6 6 3 7 6 7 6 8 6
6 6 6 4 7 6 6 5 3 8 6 7 6 8 6 7 6 6 6 8 4 4 8 6 6 2 6 5 7 3

With fertiliser

6 7 7 4 9 5 5 5 8 9 8 9 7 7 5 8 7 6 6 7 9 7 7 7 8 9 3 7 4 8 5 10 8 6 7 6 7 5 6 8
7 9 4 4 9 6 8 5 8 7 7 4 7 8 10 6 10 7 7 7 9 7 7 8 6 8 6 8 7 4 8 6 8 7 3 8 7 6 9 7
6 9 7 6 8 3 9 5 7 6 8 7 9 7 8 4 8 7 7 7 6 6 8 6 3 8 5 8 7 6 7 4 9 6 6 6 8 4 7 8
9 7 7 4 7 5 7 4 7 6 4 6 7 7 6 7 8 7 6 6 7 8 6 7 10 5 13 4 7 11

Things to think about:

- Can you state clearly the problem that the farmer wants to solve?
- How has the farmer tried to make a fair comparison?
- How could the farmer make sure that his selection was at random?
- What is the best way of organising this data?
- What are suitable methods of displaying the data?
- Are there any abnormally high or low results and how should they be treated?
- How can we best describe the most typical pod size?
- How can we best describe the spread of possible pod sizes?
- Can the farmer make a reasonable conclusion from his investigation?

Statistics is the study of data collection and analysis. In a statistical investigation we collect information about a group of individuals, then analyse this information to draw conclusions about those individuals.

Statistics are used every day in many professions including:

- medical research to measure the effectiveness of different treatment options for a particular medical condition
- psychology for personality testing
- manufacturing to aid in quality control
- politics to determine the popularity of a political party
- sport to monitor team or player performances
- marketing to assess consumer preferences and opinions.

You should already be familiar with these words which are commonly used in statistics:

- **Population** A defined collection of individuals or objects about which we want to draw conclusions.
- **Census** The collection of information from the **whole population**.
- **Sample** A subset of the population which we want to collect information from. It is important to choose a sample at **random** to avoid **bias** in the results.
- **Survey** The collection of information from a **sample**.
- **Data** (singular **datum**) Information about individuals in a population.
- **Parameter** A numerical quantity measuring some aspect of a population.
- **Statistic** A quantity calculated from data gathered from a sample. It is usually used to estimate a population parameter.

A TYPES OF DATA

When we collect data, we measure or observe a particular feature or **variable** associated with the population. The variables we observe are described as either categorical or numerical.

CATEGORICAL VARIABLES

A **categorical variable** describes a particular quality or characteristic.

The data is divided into **categories**, and the information collected is called **categorical data**.

Some examples of categorical data are:

- *computer operating system*: the categories could be Windows, Macintosh, or Linux.
- *gender*: the categories are male and female.

QUANTITATIVE OR NUMERICAL VARIABLES

A **quantitative variable** has a numerical value. The information collected is called **numerical data**.

Quantitative variables can either be **discrete** or **continuous**.

A **quantitative discrete variable** takes exact number values and is often a result of **counting**.

Some examples of quantitative discrete variables are:

- *the number of apricots on a tree*: the variable could take the values 0, 1, 2, 3, ... up to 1000 or more.
- *the number of players in a game of tennis*: the variable could take the values 2 or 4.

A **quantitative continuous variable** can take any numerical value within a certain range. It is usually a result of **measuring**.

Some examples of quantitative continuous variables are:

- *the times taken to run a* 100 *m race*: the variable would likely be between 9.8 and 25 seconds.
- *the distance of each hit in baseball*: the variable could take values from 0 m to 100 m.

Example 1 Self Tutor

Classify these variables as categorical, quantitative discrete, or quantitative continuous:

a the number of heads when 3 coins are tossed

b the brand of toothpaste used by the students in a class

c the heights of a group of 15 year old children.

a The value of the variable is obtained by counting the number of heads. The result can only be one of the values 0, 1, 2 or 3. It is a quantitative discrete variable.

b The variable describes the brands of toothpaste. It is a categorical variable.

c This is a numerical variable which can be measured. The data can take any value between certain limits, though when measured we round off the data to an accuracy determined by the measuring device. It is a quantitative continuous variable.

EXERCISE 6A

1 Classify the following variables as categorical, quantitative discrete, or quantitative continuous:

a the number of brothers a person has

b the colours of lollies in a packet

c the time children spend brushing their teeth each day

d the height of trees in a garden

e the brand of car a person drives

f the number of petrol pumps at a service station

g the most popular holiday destinations

h the scores out of 10 in a diving competition

i the amount of water a person drinks each day

j the number of hours spent per week at work

k the average temperatures of various cities

l the items students ate for breakfast before coming to school

m the number of televisions in each house.

2 For each of the variables in **1**:

- if the variable is categorical, list some possible categories for the variable
- if the variable is quantitative, give the possible values or range of values the variable may take.

B SIMPLE QUANTITATIVE DISCRETE DATA

ORGANISATION OF DATA

There are several different ways we can organise and display quantitative discrete data. One of the simplest ways to organise the data is using a **frequency table**.

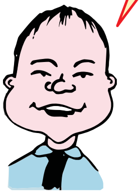
For example, consider the **Opening Problem** in which the quantitative discrete variable is *the number of peas in a pod*. For the data *without fertiliser* we count the data systematically using a **tally**.

The **frequency** of a data value is the number of times that value occurs in the data set.

The **relative frequency** of a data value is the frequency divided by the total number of recorded values. It indicates the proportion of results which take that value.

Number of peas in a pod	*Tally*	*Frequency*	*Relative frequency*
1		0	0
2	||	2	0.013
3	𝍸 𝍸 |	11	0.073
4	𝍸 𝍸 𝍸 ||||	19	0.127
5	𝍸 𝍸 𝍸 𝍸 𝍸 ||||	29	0.193
6	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 |	51	0.34
7	𝍸 𝍸 𝍸 𝍸 𝍸	25	0.167
8	𝍸 𝍸 ||	12	0.08
9	|	1	0.007
	Total	150	

A tally column is not essential for a frequency table, but is useful in the counting process for large data sets.



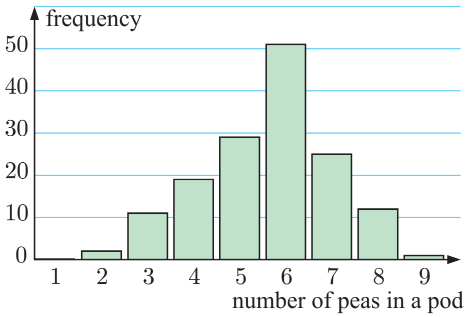
DISPLAY OF DATA

Quantitative discrete data is displayed using a **column graph**. For this graph:

- the range of data values is on the horizontal axis
- the frequency of data values is on the vertical axis
- the column widths are equal and the column height represents frequency
- there are gaps between columns to indicate the data is discrete.

A column graph for *the number of peas in a pod without fertiliser* is shown alongside.

Number of peas in a pod without fertiliser

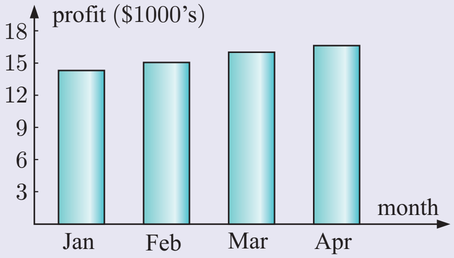


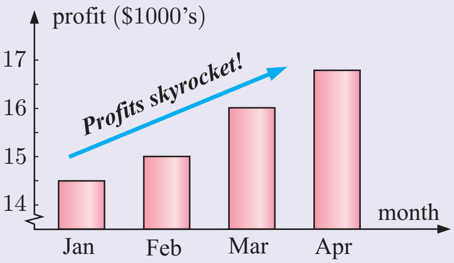
The **mode** of a data set is the most frequently occurring value. On a column graph the mode will have the highest column. In this case the mode is 6 peas in a pod.

THEORY OF KNOWLEDGE

Statistics are often used to give the reader a misleading impression of what the data actually means. In some cases this happens by accident through mistakes in the statistical process. Often, however, it is done deliberately in an attempt to persuade the reader to believe something.

Even simple things like the display of data can be done so as to create a false impression. For example, the two graphs below show the profits of a company for the first four months of the year.





Both graphs accurately display the data, but on one graph the vertical axis has a break in its scale which can give the impression that the increase in profits is much larger than it really is. The comment 'Profits skyrocket!' encourages the reader to come to that conclusion without looking at the data more carefully.

1 Given that the data is presented with mathematical accuracy in both graphs, would you say the author in the second case has lied?

When data is collected by sampling, the choice of a biased sample can be used to give misleading results. There is also the question of whether outliers should be considered as genuine data, or ignored and left out of statistical analysis.

2 In what other ways can statistics be used to deliberately mislead the target audience?

The use of statistics in science and medicine has been widely debated, as companies employ scientific 'experts' to back their claims. For example, in the multi-billion dollar tobacco industry, huge volumes of data have been collected which claim that smoking leads to cancer and other harmful effects. However, the industry has sponsored other studies which deny these claims.

There are many scientific articles and books which discuss the uses and misuses of statistics. For example:

- *Surgeons General's reports on smoking and cancer: uses and misuses of statistics and of science*, R J Hickey and I E Allen, Public Health Rep. 1983 Sep-Oct; **98**(5): 410-411.
- *Misusage of Statistics in Medical Researches*, I Ercan, B Yazici, Y Yang, G Ozkaya, S Cangur, B Ediz, I Kan, 2007, European Journal of General Medicine, **4**(3),127-133.
- *Sex, Drugs, and Body Counts: The Politics of Numbers in Global Crime and Conflict*, P Andreas and K M Greenhill, 2010, Cornell University Press.

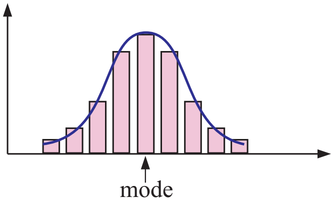
3 Can we trust statistical results published in the media and in scientific journals?

4 What role does ethics have to play in mathematics?

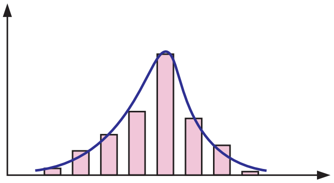
DESCRIBING THE DISTRIBUTION OF A DATA SET

Many data sets show **symmetry** or **partial symmetry** about the mode.

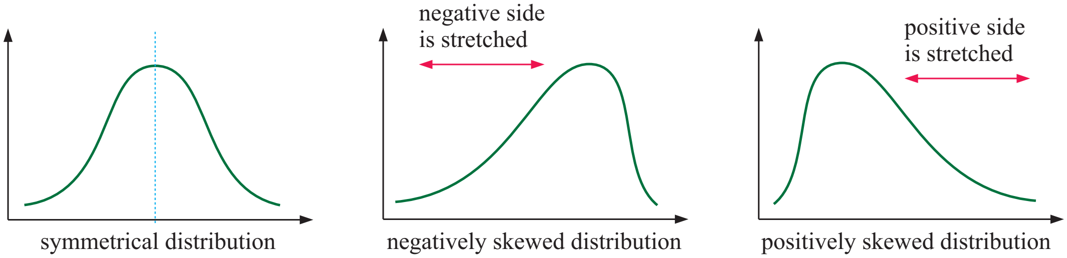
If we place a curve over the column graph we see that this curve shows symmetry. We have a **symmetrical distribution** of the data.



Comparing the *peas in a pod without fertiliser* data with the symmetrical distribution, we can see it has been 'stretched' on the left or negative side of the mode. We say the data is **negatively skewed**.



The descriptions we use are:

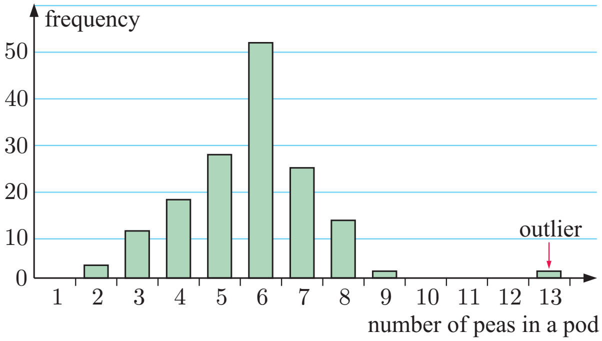


OUTLIERS

Outliers are data values that are either much larger or much smaller than the general body of data. Outliers appear separated from the body of data on a column graph.

For example, suppose the farmer in the **Opening Problem** found one pod without fertiliser that contained 13 peas. The data value 13 would be considered an outlier since it is much larger than the other data in the sample.

While knowledge of outliers is not examinable, it may be useful for statistically based projects.



Example 2 **Self Tutor**

30 children attended a library holiday programme. Their year levels at school were:

8	7	6	7	7	7	9	7	7	11	8	10	8	8	9
10	7	7	8	8	8	8	7	6	6	6	6	9	6	9

a Record this information in a frequency table. Include a column for relative frequency.

b Construct a column graph to display the data.

c What is the modal year level of the children?

d Describe the shape of the distribution. Are there any outliers?

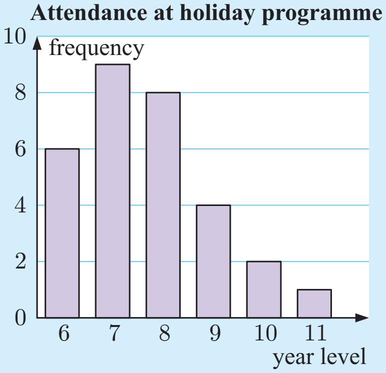
e What percentage of the children were in year 8 or below?

f What percentage of the children were above year 9?

a

Year level	*Tally*	*Frequency*	*Relative frequency*
6	𝍸 |	6	0.2
7	𝍸 ||||	9	0.3
8	𝍸 |||	8	0.267
9	||||	4	0.133
10	||	2	0.067
11	|	1	0.033
	Total	30	

b



c The modal year level is year 7.

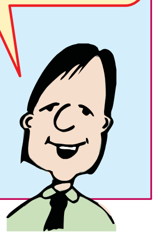
d The distribution of children's year levels is positively skewed. There are no outliers.

e $\dfrac{6+9+8}{30} \times 100\% \approx 76.7\%$ were in year 8 or below.

or the sum of the relative frequencies is $0.2 + 0.3 + 0.267 = 0.767$

$\therefore$ 76.7% were in year 8 or below.

Due to rounding, the relative frequencies will not always appear to add to *exactly* 1.



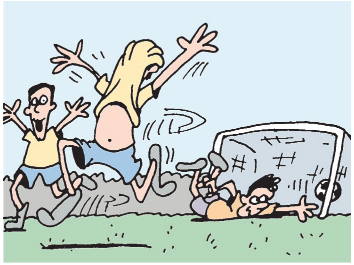
f $\dfrac{2+1}{30} \times 100\% = 10\%$ were above year 9.

or $0.067 + 0.033 = 0.1$ $\quad\therefore$ 10% were above year 9.

EXERCISE 6B

1 In the last football season, the Flames scored the following numbers of goals in each game:

2	0	1	4	0	1	2	1	1	0	3	1
3	0	1	1	6	2	1	3	1	2	0	2



a What is the variable being considered here?

b Explain why the data is discrete.

c Construct a frequency table to organise the data. Include a column for relative frequency.

d Draw a column graph to display the data.

e What is the modal score for the team?

f Describe the distribution of the data. Are there any outliers?

g In what percentage of games did the Flames fail to score?

2 Prince Edward High School prides itself on the behaviour of its students. However, from time to time they do things they should not, and as a result are placed on detention. The studious school master records the number of students on detention each week throughout the year:

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

a Construct a column graph to display the data.

b What is the modal number of students on detention in a week?

c Describe the distribution of the data, including the presence of outliers.

d In what percentage of weeks were more than 4 students on detention?

3 While watching television, Joan recorded the number of commercials in each break. She obtained these results:

5	7	6	4	6	5	6	7	5	8
7	6	9	8	7	6	6	9	6	7
6	4	7	5	8	7	6	8	7	8
5	6	9	7						

a Construct a frequency table to organise the data.

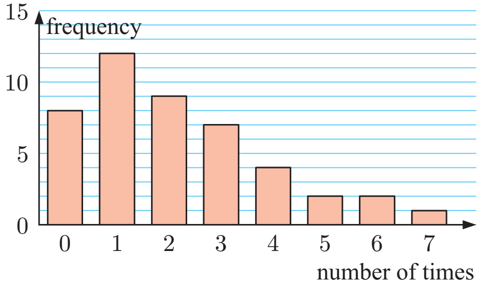
b Draw a column graph to display the data.

c Find the mode of the data.

d Describe the distribution of the data. Are there any outliers?

e What percentage of breaks contained at least 6 commercials?

4 A random sample of people were asked "How many times did you eat at a restaurant last week?" A column graph was used to display the results.



a How many people were surveyed?

b Find the mode of the data.

c How many people surveyed did not eat at a restaurant at all last week?

d What percentage of people surveyed ate at a restaurant more than three times last week?

e Describe the distribution of the data.

5 Consider *the number of peas in a pod with fertiliser* in the **Opening Problem**.

a Construct a frequency table to organise the data.

b Draw a column graph to display the data.

c Describe fully the distribution of the data.

d Is there evidence to suggest that the fertiliser increases the number of peas in each pod?

e Is it reasonable to say that using the fertiliser will increase the farmer's profits?

C GROUPED QUANTITATIVE DISCRETE DATA

A local kindergarten is concerned about the number of vehicles passing by between 8:45 am and 9:00 am. Over 30 consecutive week days they recorded data:

27, 30, 17, 13, 46, 23, 40, 28, 38, 24, 23, 22, 18, 29, 16,
35, 24, 18, 24, 44, 32, 52, 31, 39, 32, 9, 41, 38, 24, 32

In situations like this there are many different data values with very low frequencies. This makes it difficult to study the data distribution. It is more statistically meaningful to group the data into **class intervals** and then compare the frequency for each class.

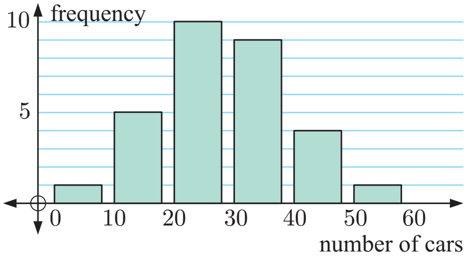
For the data given we use class intervals of width 10. The frequency table is shown opposite.

We see the **modal class**, or class with the highest frequency, is from 20 to 29 cars.

Number of cars	*Tally*	*Frequency*
0 to 9	|	1
10 to 19	卌	5
20 to 29	卌 卌	10
30 to 39	卌 ||||	9
40 to 49	||||	4
50 to 59	|	1
	Total	30

We can construct a **column graph** for grouped discrete data in the same way as before.

Vehicles passing kindergarten between 8:45 am and 9:00 am



DISCUSSION

- If we are given a set of raw data, how can we efficiently find the lowest and highest data values?
- If the data values are grouped in classes on a frequency table or column graph, do we still know what the highest and lowest values are?

EXERCISE 6C

1 Arthur catches the train to school from a busy train station. Over the course of 30 days he counts the number of people waiting at the station when the train arrives.

17	25	32	19	45	30	22	15	38	8
21	29	37	25	42	35	19	31	26	7
22	11	27	44	24	22	32	18	40	29

a Construct a tally and frequency table for this data using class intervals 0 - 9, 10 - 19, ..., 40 - 49.

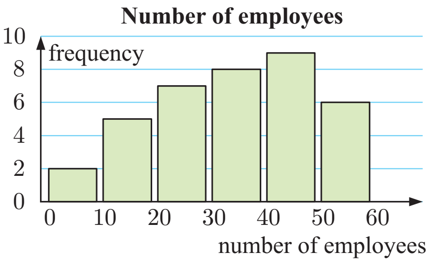
b On how many days were there less than 10 people at the station?

c On what percentage of days were there at least 30 people at the station?

d Draw a column graph to display the data.

e Find the modal class of the data.

2 A selection of businesses were asked how many employees they had. A column graph was constructed to display the results.



a How many businesses were surveyed?

b Find the modal class.

c Describe the distribution of the data.

d What percentage of businesses surveyed had less than 30 employees?

e Can you determine the highest number of employees a business had?

3 A city council does a survey of the number of houses per street in a suburb.

42	15	20	6	34	19	8	5	11	38	56	23	24	24
35	47	22	36	39	18	14	44	25	6	34	35	28	12
27	32	36	34	30	40	32	12	17	6	37	32		

a Construct a frequency table for this data using class intervals 0 - 9, 10 - 19, ..., 50 - 59.

b Hence draw a column graph to display the data.

c Write down the modal class.

d What percentage of the streets contain at least 20 houses?

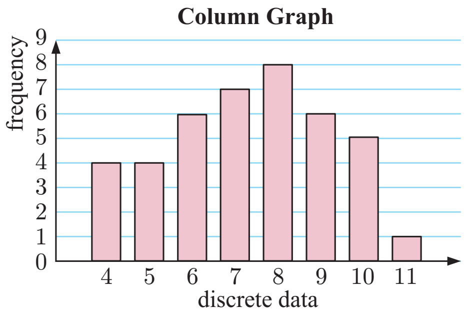
D QUANTITATIVE CONTINUOUS DATA

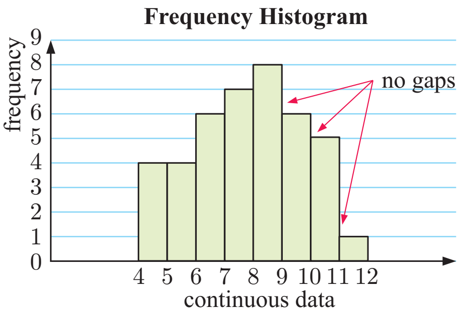
When we measure data that is **continuous**, we cannot write down an exact value. Instead we write down an approximation which is only as accurate as the measuring device.

Since no two data values will be *exactly* the same, it does not make sense to talk about the frequency of particular values. Instead we group the data into **class intervals** of **equal width**. We can then talk about the frequency of each class interval.

A special type of graph called a **frequency histogram** or just **histogram** is used to display grouped continuous data. This is similar to a column graph, but the 'columns' are joined together and the values at the edges of the column indicate the boundaries of each class interval.

The **modal class**, or class of values that appears most often, is easy to identify from a frequency histogram.





INVESTIGATION 1 CHOOSING CLASS INTERVALS

When dividing data values into intervals, the choice of how many intervals to use, and hence the width of each class, is important.

DEMO

What to do:

1. Click on the icon to experiment with various data sets and the number of classes. How does the number of classes alter the way we can interpret the data?
2. Write a brief account of your findings.

As a rule of thumb we use approximately $\sqrt{n}$ classes for a data set of n individuals. For very large sets of data we use more classes rather than less.

Example 3

Self Tutor

A sample of 20 juvenile lobsters was randomly selected from a tank containing several hundred. The length of each lobster was measured in cm, and the results were:

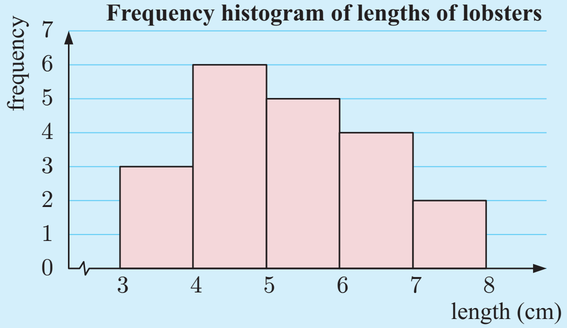
$$\begin{array}{cccccccccc} 4.9 & 5.6 & 7.2 & 6.7 & 3.1 & 4.6 & 6.0 & 5.0 & 3.7 & 7.3 \\ 6.0 & 5.4 & 4.2 & 6.6 & 4.7 & 5.8 & 4.4 & 3.6 & 4.2 & 5.4 \end{array}$$

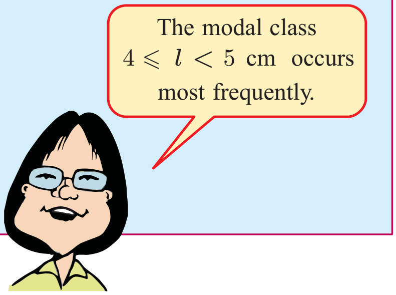
Organise the data using a frequency table, and hence graph the data.

The variable 'the length of a lobster' is *continuous* even though lengths have been rounded to the nearest mm.

The shortest length is 3.1 cm and the longest is 7.3 cm, so we will use class intervals of width 1 cm.

Length (l cm)	*Frequency*
$3 \leqslant l < 4$	3
$4 \leqslant l < 5$	6
$5 \leqslant l < 6$	5
$6 \leqslant l < 7$	4
$7 \leqslant l < 8$	2




The modal class $4 \leqslant l < 5$ cm occurs most frequently.

EXERCISE 6D

1. A frequency table for the heights of a volleyball squad is given alongside.
   a. Explain why 'height' is a continuous variable.
   b. Construct a frequency histogram for the data. Carefully mark and label the axes, and include a heading for the graph.
   c. What is the modal class? Explain what this means.
   d. Describe the distribution of the data.

Height (H cm)	*Frequency*
$170 \leqslant H < 175$	1
$175 \leqslant H < 180$	8
$180 \leqslant H < 185$	9
$185 \leqslant H < 190$	11
$190 \leqslant H < 195$	9
$195 \leqslant H < 200$	3
$200 \leqslant H < 205$	3

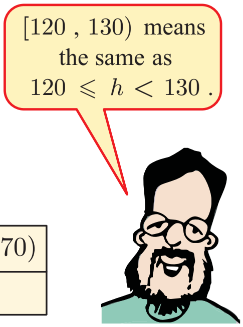
2 For the following data, state whether a frequency histogram or a column graph should be used, and draw the appropriate graph.

a The number of matches in 30 match boxes:

Number of matches per box	47	49	50	51	52	53	55
Frequency	1	1	9	12	4	2	1

b The heights of 25 hockey players (to the nearest cm):

Height (h cm)	[120, 130)	[130, 140)	[140, 150)	[150, 160)	[160, 170)
Frequency	1	2	7	14	1



3 A school has conducted a survey of 60 students to investigate the time it takes for them to travel to school. The following data gives the travel times to the nearest minute.

12	15	16	8	10	17	25	34	42	18	24	18	45	33	38
45	40	3	20	12	10	10	27	16	37	45	15	16	26	32
35	8	14	18	15	27	19	32	6	12	14	20	10	16	14
28	31	21	25	8	32	46	14	15	20	18	8	10	25	22

a Is travel time a discrete or continuous variable?

b Construct a frequency table for the data using class intervals $0 \leqslant t < 10$, $10 \leqslant t < 20$, ..., $40 \leqslant t < 50$.

c Hence draw a histogram to display the data.

d Describe the distribution of the data.

e What is the modal travelling time?

4 A group of 25 young athletes participated in a javelin throwing competition. They achieved the following distances in metres:

17.6	25.7	21.3	30.9	13.0	31.6	22.3	28.3	7.4
38.4	19.1	24.0	40.0	16.2	42.9	31.9	28.1	41.8
13.6	27.4	33.7	9.2	23.3	39.8	25.1		

a Choose suitable class intervals to group the data.

b Organise the data in a frequency table.

c Draw a frequency histogram to display the data.

d Find the modal class.

e What percentage of athletes threw the javelin 30 m or further?

5 A plant inspector takes a random sample of six month old seedlings from a nursery and measures their heights. The results are shown in the table.

Height (h mm)	*Frequency*
$300 \leqslant h < 325$	12
$325 \leqslant h < 350$	18
$350 \leqslant h < 375$	42
$375 \leqslant h < 400$	28
$400 \leqslant h < 425$	14
$425 \leqslant h < 450$	6

a Represent the data on a frequency histogram.

b How many of the seedlings are 400 mm or more?

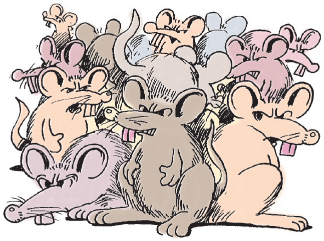
c What percentage of the seedlings are between 350 and 400 mm?

d The total number of seedlings in the nursery is 1462. Estimate the number of seedlings which measure:

i less than 400 mm **ii** between 375 and 425 mm.

6 The weights, in grams, of 50 laboratory rats are given below.

261	133	173	295	265	142	140	271	185	251
166	100	292	107	201	234	239	159	153	263
195	151	156	117	144	189	234	171	233	182
165	122	281	149	152	289	168	260	256	156
239	203	101	268	241	217	254	240	214	221



a Choose suitable class intervals to group the data.

b Organise the data in a frequency table.

c Draw a frequency histogram to display the data.

d What percentage of the rats weigh less than 200 grams?

E MEASURING THE CENTRE OF DATA

We can get a better understanding of a data set if we can locate its **middle** or **centre**, and also get an indication of its **spread** or **dispersion**. Knowing one of these without the other is often of little use.

There are *three statistics* that are used to measure the **centre** of a data set. These are the **mode**, the **mean**, and the **median**.

THE MODE

For ungrouped discrete numerical data, the **mode** is the most frequently occuring value in the data set.

For grouped numerical data, we talk about a **modal class**, which is the class that occurs most frequently.

If a set of scores has two modes we say it is **bimodal**. If there are more than two modes, we do not use mode as a measure of the centre.

THE MEAN

The **mean** of a data set is the statistical name for its arithmetic average.

$$\textbf{mean} = \frac{\textbf{sum of all data values}}{\textbf{the number of data values}}$$

The mean gives us a single number which indicates a centre of the data set. It is usually not a member of the data set.

For example, a mean test mark of 73% tells us that there are several marks below 73% and several above it. 73% is at the centre, but it does not necessarily mean that one of the students scored 73%.

We denote the mean for an entire **population** by μ, which we read as "mu".

However, in many cases we do not have data for all of the population, and so the exact value of μ is unknown. Instead we obtain data from a **sample** of the population and use the mean of the sample, $\overline{x}$, as an *approximation* for μ.

Suppose x is a numerical variable and there are n data values in the sample. We let x_i be the ith data value from the sample of values $\{x_1, x_2, x_3, ..., x_n\}$.

The mean of the sample is $$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where $\sum_{i=1}^{n} x_i$ means the **sum** of all n data values, $x_1 + x_2 + ... + x_n$.

THE MEDIAN

The **median** is the *middle value* of an ordered data set.

An ordered data set is obtained by listing the data, usually from smallest to largest.

The median splits the data in halves. Half of the data are less than or equal to the median, and half are greater than or equal to it.

For example, if the median mark for a test is 73% then you know that half the class scored less than or equal to 73%, and half scored greater than or equal to 73%.

For an **odd number** of data, the median is one of the original data values.

For an **even number** of data, the median is the average of the two middle values, and may not be in the original data set.

If there are n data values, find $\frac{n+1}{2}$. The median is the $\left(\frac{n+1}{2}\right)$th data value.

For example:

If $n = 13$, $\frac{n+1}{2} = \frac{13+1}{2} = 7$, so the median $=$ 7th ordered data value.

If $n = 14$, $\frac{n+1}{2} = \frac{14+1}{2} = 7.5$, so the median $=$ average of the 7th and 8th ordered data values.

DEMO

Example 4 Self Tutor

Find the **i** mean **ii** mode **iii** median of the following data sets:

a 3, 6, 5, 6, 4, 5, 5, 6, 7 **b** 13, 12, 15, 13, 18, 14, 16, 15, 15, 17

a **i** mean $= \frac{3+6+5+6+4+5+5+6+7}{9} = \frac{47}{9} \approx 5.22$

ii The scores 5 and 6 occur most frequently, so the data set is bimodal with modes 5 and 6.

iii Listing the set in order of size: ~~3, 4, 5, 5,~~ 5, ~~6, 6, 6, 7~~ $\{$as $n = 9$, $\frac{n+1}{2} = 5\}$

middle score

$\therefore$ the median is 5.

b **i** mean $= \frac{13+12+15+13+18+14+16+15+15+17}{10} = \frac{148}{10} = 14.8$

ii The score 15 occurs most frequently, so the mode is 15.

iii Listing the set in order of size:

~~12, 13, 13, 14,~~ $\underbrace{15,\ 15,}$ ~~15, 16, 17, 18~~ $\{$as $n = 10$, $\frac{n+1}{2} = 5.5\}$

middle scores

The median is the average of the two middle scores, which is $\frac{15+15}{2} = 15$.

Technology can be used to help find the statistics of a data set. Click on the appropriate icon to obtain instructions for your calculator or run software on the CD.

GRAPHICS CALCULATOR INSTRUCTIONS

STATISTICS PACKAGE

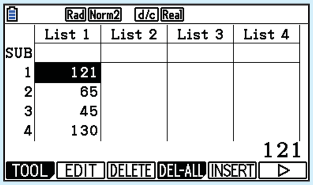
Example 5

Self Tutor

A teenager recorded the time (in minutes per day) he spent playing computer games over a 2 week holiday period: 121, 65, 45, 130, 150, 83, 148, 137, 20, 173, 56, 49, 104, 97.

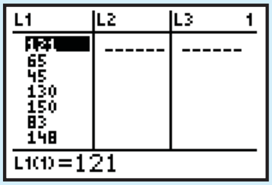
Using technology to assist, determine the mean and median daily game time the teenager recorded.

Casio fx-CG20



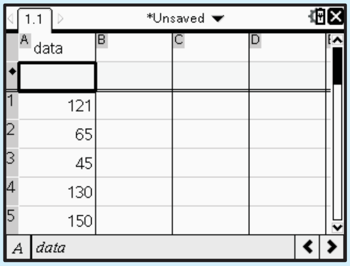
```
     List 1  List 2  List 3  List 4
SUB
  1     121
  2      65
  3      45
  4     130
                               121
```

TI-84 Plus

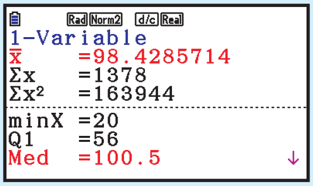


```
L1     L2     L3     1
121    ------ ------
65
45
130
150
83
148
L1(1)=121
```

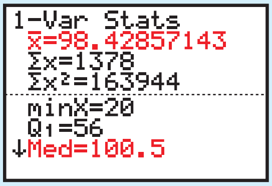
TI-nspire



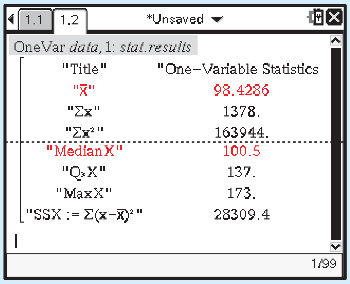
```
A data    B    C    D
1   121
2    65
3    45
4   130
5   150
A data
```



```
1-Variable
x̄      =98.4285714
Σx     =1378
Σx²    =163944
minX   =20
Q1     =56
Med    =100.5
```



```
1-Var Stats
x̄=98.42857143
Σx=1378
Σx²=163944
minX=20
Q1=56
↓Med=100.5
```



```
OneVar data,1: stat.results
 "Title"            "One-Variable Statistics"
 "x̄"                98.4286
 "Σx"               1378.
 "Σx²"              163944.
 "MedianX"          100.5
 "Q3X"              137.
 "MaxX"             173.
 "SSX := Σ(x-x̄)²"   28309.4
                                       1/99
```

The mean $\overline{x} \approx 98.4$ minutes, and the median $= 100.5$ minutes.

EXERCISE 6E.1

1. Phil kept a record of the number of cups of coffee he drank each day for 15 days:

   2, 3, 1, 1, 0, 0, 4, 3, 0, 1, 2, 3, 2, 1, 4

   Without using technology, find the **a** mode **b** median **c** mean of the data.

2. The sum of 7 scores is 63. What is their mean?

3. Find the **i** mean **ii** median **iii** mode for each of the following data sets:

   a 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 8, 8, 8, 9, 9

   b 10, 12, 12, 15, 15, 16, 16, 17, 18, 18, 18, 18, 19, 20, 21

   c 22.4, 24.6, 21.8, 26.4, 24.9, 25.0, 23.5, 26.1, 25.3, 29.5, 23.5

4. Consider the two data sets: *Data set A*: 3, 4, 4, 5, 6, 6, 7, 7, 7, 8, 8, 9, 10

   Data set B: 3, 4, 4, 5, 6, 6, 7, 7, 7, 8, 8, 9, 15

   a Find the mean of both data set A and data set B.

   b Find the median of both data set A and data set B.

   c Explain why the mean of data set A is less than the mean of data set B.

   d Explain why the median of data set A is the same as the median of data set B.

5 The scores obtained by two ten-pin bowlers over a 10 game series are:

Gordon:	160,	175,	142,	137,	151,	144,	169,	182,	175,	155
Ruth:	157,	181,	164,	142,	195,	188,	150,	147,	168,	148

Who had the higher mean score?

6 A bakery keeps a record of how many pies and pasties they sell each day for a month.

Pies

62	76	55	65	49	78	71	82
79	47	60	72	58	82	76	67
50	61	70	85	77	69	48	74
63	56	81	75	63	74	54	

Pasties

37	52	71	59	63	47	56	68
43	67	38	73	54	55	61	49
50	48	53	39	45	60	46	51
38	57	41	72	50	44	76	

a Using technology to assist, find the:

i mean number of pies and pasties sold **ii** median number of pies and pasties sold.

b Which bakery item was more popular? Explain your answer.

7 A bus and tram travel the same route many times during the day. The drivers counted the number of passengers on each trip one day, as listed below.

Bus

30	43	40	53	70	50	63
41	38	21	28	23	43	48
20	26	35	48	41	33	

Tram

58	68	43	45	70	79
38	23	30	22	63	73
25	35	60	53		

a Using technology, calculate the mean and median number of passengers for both the Bus and Tram data.

b Comment on which mode of transport is more popular. Explain your answer.

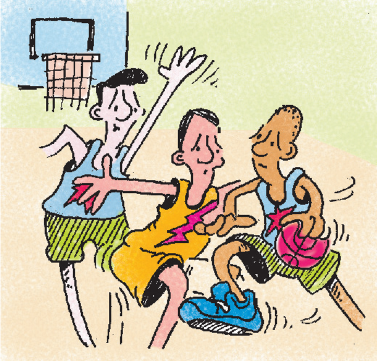
8 A basketball team scored 43, 55, 41, and 37 points in their first four matches.

a What is the mean number of points scored for the first four matches?

b What score will the team need to shoot in their next match so that they maintain the same mean score?

c The team scores only 25 points in the fifth match. Find the mean number of points scored for the five matches.

d The team then scores 41 points in their sixth and final match. Will this increase or decrease their previous mean score? What is the mean score for all six matches?



Example 6 Self Tutor

If 6 people have a mean mass of 53.7 kg, find their total mass.

$$\frac{\text{sum of masses}}{6} = 53.7 \text{ kg}$$

$$\therefore \text{ the total mass} = 53.7 \times 6 = 322.2 \text{ kg.}$$

9 This year, the mean monthly sales for a clothing store have been $15 467. Calculate the total sales for the store for the year.

10 While on an outback safari, Bill drove an average of 262 km per day for a period of 12 days. How far did Bill drive in total while on safari?

11 Towards the end of the season, a netballer had played 14 matches and had thrown an average of 16.5 goals per game. In the final two matches of the season she threw 21 goals and 24 goals. Find the netballer's new average.

12 Find x if $5, 9, 11, 12, 13, 14, 17$, and x have a mean of 12.

13 Find a given that $3, 0, a, a, 4, a, 6, a$, and 3 have a mean of 4.

14 Over the complete assessment period, Aruna averaged 35 out of 40 marks for her maths tests. However, when checking her files, she could only find 7 of the 8 tests. For these she scored 29, 36, 32, 38, 35, 34, and 39. How many marks out of 40 did she score for the eighth test?

15 A sample of 10 measurements has a mean of 15.7 and a sample of 20 measurements has a mean of 14.3. Find the mean of all 30 measurements.

16 The mean and median of a set of 9 measurements are both 12. Seven of the measurements are 7, 9, 11, 13, 14, 17, and 19. Find the other two measurements.

17 Jana took seven spelling tests, each with twelve words, but she could only find the results of five of them. These were 9, 5, 7, 9, and 10. She asked her teacher for the other two results and the teacher said that the mode of her scores was 9 and the mean was 8. Given that Jana knows her worst result was a 5, find the two missing results.

INVESTIGATION 2 EFFECTS OF OUTLIERS

We have seen that an **outlier** or **extreme value** is a value which is much greater than, or much less than, the other values.

Your task is to examine the effect of an outlier on the three measures of central tendency.

What to do:

1 Consider the set of data: 4, 5, 6, 6, 6, 7, 7, 8, 9, 10. Calculate:

a the mean **b** the mode **c** the median.

2 We now introduce the extreme value 100 to the data, so the data set is now:
4, 5, 6, 6, 6, 7, 7, 8, 9, 10, 100. Calculate:

a the mean **b** the mode **c** the median.

3 Comment on the effect that the extreme value has on:

a the mean **b** the mode **c** the median.

4 Which of the three measures of central tendency is most affected by the inclusion of an outlier?

5 Discuss with your class when it would not be appropriate to use a particular measure of the centre of a data set.

CHOOSING THE APPROPRIATE MEASURE

The mean, mode, and median can all be used to indicate the centre of a set of numbers. The most appropriate measure will depend upon the type of data under consideration. When selecting which one to use for a given set of data, you should keep the following properties in mind.

Mode:	• gives the most usual value • only takes common values into account • not affected by extreme values
Mean:	• commonly used and easy to understand • takes all values into account • affected by extreme values
Median:	• gives the halfway point of the data • only takes middle values into account • not affected by extreme values

For example:

- A shoe store is investigating the sizes of shoes sold over one month. The mean shoe size is not very useful to know, but the mode shows at a glance which size the store most commonly has to restock.
- On a particular day a computer shop makes sales of $900, $1250, $1000, $1700, $1140, $1100, $1495, $1250, $1090, and $1075. Here the mode is meaningless, the median is $1120, and the mean is $1200. The mean is the best measure of centre as the salesman can use it to predict average profit.
- When looking at real estate prices, the mean is distorted by the few sales of very expensive houses. For a typical house buyer, the median will best indicate the price they should expect to pay in a particular area.

EXERCISE 6E.2

1 The selling prices of the last 10 houses sold in a certain district were as follows:

$146 400, $127 600, $211 000, $192 500,
$256 400, $132 400, $148 000, $129 500,
$131 400, $162 500

a Calculate the mean and median selling prices and comment on the results.

b Which measure would you use if you were:

i a vendor wanting to sell your house

ii looking to buy a house in the district?

2 The annual salaries of ten office workers are:

$23 000, $46 000, $23 000, $38 000, $24 000,
$23 000, $23 000, $38 000, $23 000, $32 000

a Find the mean, median, and modal salaries of this group.

b Explain why the mode is an unsatisfactory measure of the middle in this case.

c Is the median a satisfactory measure of the middle of this data set?

3 The following raw data is the daily rainfall, to the nearest millimetre, for a month:

3, 1, 0, 0, 0, 0, 0, 2, 0, 0, 3, 0, 0, 0, 7, 1, 1, 0, 3, 8, 0, 0, 0, 42, 21, 3, 0, 3, 1, 0, 0

a Using technology, find the mean, median, and mode of the data.

b Give a reason why the median is not the most suitable measure of centre for this set of data.

c Give a reason why the mode is not the most suitable measure of centre for this set of data.

d Are there any outliers in this data set?

e On some occasions outliers are removed because they must be due to errors in observation or calculation. If the outliers in the data set were accurately found, should they be removed before finding the measures of the middle?

MEASURES OF THE CENTRE FROM OTHER SOURCES

When the same data appears several times we often summarise the data in table form.

Consider the data in the given table:

Data value (x)	*Frequency* (f)	*Product* (fx)
3	1	$1 \times 3 = 3$
4	1	$1 \times 4 = 4$
5	3	$3 \times 5 = 15$
6	7	$7 \times 6 = 42$
7	15	$15 \times 7 = 105$
8	8	$8 \times 8 = 64$
9	5	$5 \times 9 = 45$
Total	$\sum f = 40$	$\sum fx = 278$

We can find the measures of the centre directly from the table.

The mode

The data value 7 has the highest frequency.

The mode is therefore 7.

The mean

Adding a 'Product' column to the table helps to add all scores.

For example, there are 15 data of value 7 and these add to $15 \times 7 = 105$.

Remembering that the mean $= \dfrac{\text{sum of all data values}}{\text{the number of data values}}$, we find

$$\overline{x} = \frac{f_1x_1 + f_2x_2 + \dots + f_kx_k}{f_1 + f_2 + \dots + f_k} = \frac{\sum_{i=1}^{k} f_ix_i}{n}$$

where $n = \sum_{i=1}^{k} f_i$ is the total number of data,

and k is the number of *different* data values.

This formula is often abbreviated as $\overline{x} = \dfrac{\sum fx}{\sum f}$.

In this case the mean $= \dfrac{278}{40} = 6.95$.

The median

Since $\dfrac{n+1}{2} = \dfrac{41}{2} = 20.5$, the median is the average of the 20th and 21st data values.

In the table, the blue numbers show us accumulated values, or the **cumulative frequency**.

Data value	*Frequency*	*Cumulative frequency*
3	1	1 ← 1 number is 3
4	1	2 ← 2 numbers are 4 or less
5	3	5 ← 5 numbers are 5 or less
6	7	12 ← 12 numbers are 6 or less
7	15	27 ← 27 numbers are 7 or less
8	8	35 ← 35 numbers are 8 or less
9	5	40 ← all numbers are 9 or less
Total	40	

We can see that the 20th and 21st data values (in order) are both 7s.

$\therefore$ the median $= \dfrac{7+7}{2} = 7$.

Example 7

Self Tutor

The table shows the number of aces served by tennis players in their first sets of a tournament.

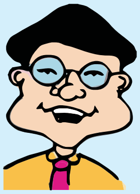
Number of aces	1	2	3	4	5	6
Frequency	4	11	18	13	7	2

Determine the:

a mean **b** median **c** mode for this data.

Number of aces (x)	*Frequency* (f)	*Product* (fx)	*Cumulative frequency*
1	4	4	4
2	11	22	15
3	18	54	33
4	13	52	46
5	7	35	53
6	2	12	55
Total	$\sum f = 55$	$\sum fx = 179$	

In this case $\dfrac{\sum fx}{\sum f}$ is short for $\dfrac{\sum_{i=1}^{6} f_i x_i}{\sum_{i=1}^{6} f_i}$.



a $$\overline{x} = \frac{\sum fx}{\sum f} = \frac{179}{55} \approx 3.25 \text{ aces}$$

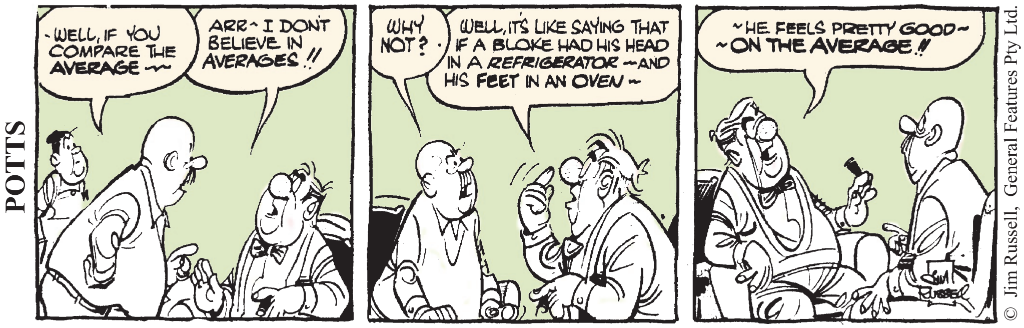
b There are 55 data values, so $n = 55$. $\dfrac{n+1}{2} = 28$, so the median is the 28th data value.

From the cumulative frequency column, the data values 16 to 33 are 3 aces.

$\therefore$ the 28th data value is 3 aces.

$\therefore$ the median is 3 aces.

c Looking down the frequency column, the highest frequency is 18. This corresponds to 3 aces, so the mode is 3 aces.


POTTS
~WELL, IF YOU COMPARE THE AVERAGE ~~
ARR~ I DON'T BELIEVE IN AVERAGES !!
WHY NOT ?
WELL, IT'S LIKE SAYING THAT IF A BLOKE HAD HIS HEAD IN A REFRIGERATOR ~AND HIS FEET IN AN OVEN ~
~HE FEELS PRETTY GOOD~ ~ON THE AVERAGE !!
JIM RUSSELL
© Jim Russell, General Features Pty Ltd.

The publishers acknowledge the late Mr Jim Russell, General Features for the reproduction of this cartoon.

We can use a graphics calculator to find the measures of centre of grouped data, by entering the data in two lists. We need to adjust the commands we give the calculator so that the calculator uses both the scores and the corresponding frequency values.

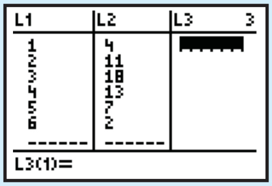
GRAPHICS CALCULATOR INSTRUCTIONS

Example 8

Self Tutor

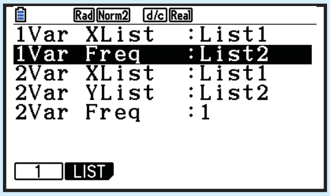
Use technology to find the mean and median of the tennis data in **Example 7**.

After entering the data in Lists 1 and 2, we calculate the descriptive statistics for the data.



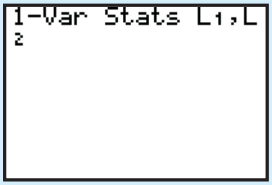
```
L1      |L2      |L3      3
1       |4       |
2       |11      |
3       |18      |
4       |13      |
5       |7       |
6       |2       |
------   ------
L3(1)=
```

Casio fx-CG20



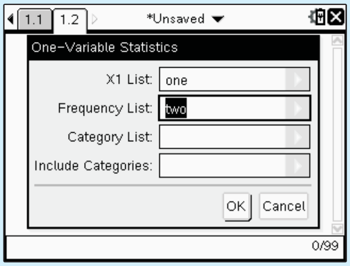
```
1Var XList   :List1
1Var Freq    :List2
2Var XList   :List1
2Var YList   :List2
2Var Freq    :1

 1  LIST
```

TI-84 Plus

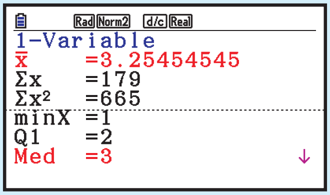


```
1-Var Stats L₁,L
₂
```

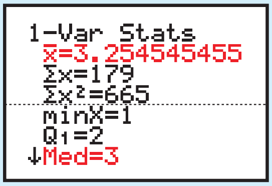
TI-nspire



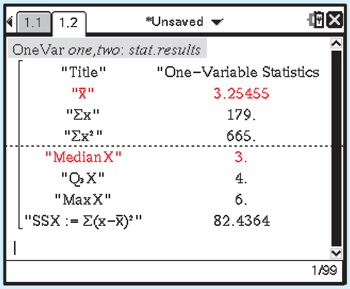
```
One-Variable Statistics
            X1 List: one
     Frequency List: two
      Category List:
Include Categories:
                         OK  Cancel
```



```
1-Variable
x̄    =3.25454545
Σx   =179
Σx²  =665
minX =1
Q1   =2
Med  =3
```



```
1-Var Stats
 x̄=3.254545455
 Σx=179
 Σx²=665
 minX=1
 Q₁=2
↓Med=3
```



```
OneVar one,two: stat.results
  "Title"            "One-Variable Statistics
  "x̄"                3.25455
  "Σx"               179.
  "Σx²"              665.
  "MedianX"          3.
  "Q₃X"              4.
  "MaxX"             6.
  "SSX := Σ(x-x̄)²"   82.4364
```

The mean $\overline{x} \approx 3.25$ aces, and the median $= 3$ aces.

EXERCISE 6E.3

1 The table alongside shows the results when 3 coins were tossed simultaneously 30 times.

Calculate the:

a mode **b** median **c** mean.

Number of heads	*Frequency*
0	4
1	12
2	11
3	3
Total	30

2 Families at a school in Australia were surveyed, and the number of children in each family recorded. The results of the survey are shown alongside.

Number of children	*Frequency*
1	5
2	28
3	15
4	8
5	2
6	1
Total	59

a Using technology, calculate the:

i mean **ii** mode **iii** median.

b The average Australian family has 2.2 children. How does this school compare to the national average?

c The data set is skewed. Is the skewness positive or negative?

d How has the skewness of the data affected the measures of its centre?

3 The following frequency table records the number of phone calls made in a day by 50 fifteen-year-olds.

Number of phone calls	*Frequency*
0	5
1	8
2	13
3	8
4	6
5	3
6	3
7	2
8	1
11	1

a For this data, find the:

i mean **ii** median **iii** mode.

b Construct a column graph for the data and show the position of the mean, median, and mode on the horizontal axis.

c Describe the distribution of the data.

d Why is the mean larger than the median for this data?

e Which measure of centre would be the most suitable for this data set?

4 A company claims that their match boxes contain, on average, 50 matches per box. On doing a survey, the Consumer Protection Society recorded the following results:

Number in a box	*Frequency*
47	5
48	4
49	11
50	6
51	3
52	1
Total	30

a Use technology to calculate the:

i mode **ii** median **iii** mean.

b Do the results of this survey support the company's claim?

c In a court for 'false advertising', the company won their case against the Consumer Protection Society. Suggest how they did this.

5 Consider again the **Opening Problem** on page **158**.

a Use a frequency table for the *Without fertiliser* data to find the:

i mean **ii** mode **iii** median number of peas per pod.

b Use a frequency table for the *With fertiliser* data to find the:

i mean **ii** mode **iii** median number of peas per pod.

c Which of the measures of centre is appropriate to use in a report on this data?

d Has the application of fertiliser significantly improved the number of peas per pod?

DATA IN CLASSES

When information has been gathered in classes, we use the **midpoint** or **mid-interval value** of the class to represent all scores within that interval.

We are assuming that the scores within each class are evenly distributed throughout that interval. The mean calculated is an **approximation** of the true value, and we cannot do better than this without knowing each individual data value.

INVESTIGATION 3 MID-INTERVAL VALUES

When mid-interval values are used to represent all scores within that interval, what effect will this have on estimating the mean of the grouped data?

Consider the following table which summarises the marks received by students for a physics examination out of 50. The exact results for each student have been lost.

Marks	*Frequency*
0 - 9	2
10 - 19	31
20 - 29	73
30 - 39	85
40 - 49	28

What to do:

1 Suppose that all of the students scored the lowest possible result in their class interval, so 2 students scored 0, 31 students scored 10, and so on.
Calculate the mean of these results, and hence complete:
"The mean score of students in the physics examination must be *at least* ...."

2 Now suppose that all of the students scored the highest possible result in their class interval.
Calculate the mean of these results, and hence complete:
"The mean score of students in the physics examination must be *at most* ...."

3 We now have two extreme values between which the actual mean must lie.
Now suppose that all of the students scored the mid-interval value in their class interval. We assume that 2 students scored 4.5, 31 students scored 14.5, and so on.

a Calculate the mean of these results.

b How does this result compare with lower and upper limits found in **1** and **2**?

c Copy and complete:
"The mean score of students in the physics examination was approximately ...."

Example 9 Self Tutor

Estimate the mean of the following *ages of bus drivers* data, to the nearest year:

Age (yrs)	21 - 25	26 - 30	31 - 35	36 - 40	41 - 45	46 - 50	51 - 55
Frequency	11	14	32	27	29	17	7

Age (yrs)	*Frequency* (f)	*Midpoint* (x)	$f\,x$
21 - 25	11	23	253
26 - 30	14	28	392
31 - 35	32	33	1056
36 - 40	27	38	1026
41 - 45	29	43	1247
46 - 50	17	48	816
51 - 55	7	53	371
Total	$\sum f = 137$		$\sum fx = 5161$

$$\overline{x} = \frac{\sum f\,x}{\sum f}$$

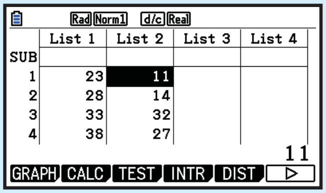
$$= \frac{5161}{137}$$

$$\approx 37.7$$

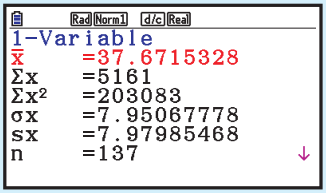
$\therefore$ the mean age of the drivers is about 38 years.

or we can find the same result using technology:

Casio fx-CG20

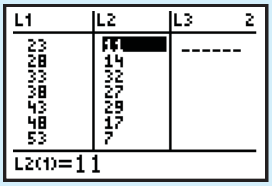


```
Rad Norm1  d/c Real
     List 1  List 2  List 3  List 4
SUB
  1      23      11
  2      28      14
  3      33      32
  4      38      27
                                  11
GRAPH CALC TEST INTR DIST ▷
```

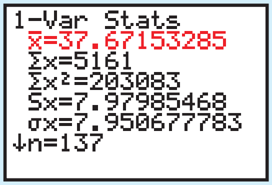


```
Rad Norm1  d/c Real
1-Variable
x̄      =37.6715328
Σx     =5161
Σx²    =203083
σx     =7.95067778
sx     =7.97985468
n      =137
```

TI-84 Plus

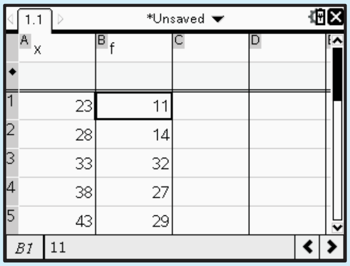


```
L1      L2      L3      2
23      11      ------
28      14
33      32
38      27
43      29
48      17
53      7
L2(1)=11
```

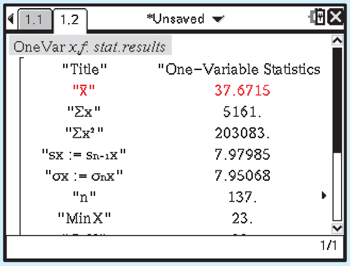


```
1-Var Stats
 x̄=37.67153285
 Σx=5161
 Σx²=203083
 Sx=7.97985468
 σx=7.950677783
↓n=137
```

TI-*n*spire



```
1.1          *Unsaved
   A x    B f    C    D
◆
1   23     11
2   28     14
3   33     32
4   38     27
5   43     29
B1  11
```



```
1.1  1.2     *Unsaved
OneVar x,f: stat.results
   "Title"            "One-Variable Statistics
   "x̄"                37.6715
   "Σx"               5161.
   "Σx²"              203083.
   "sx := sn-1x"      7.97985
   "σx := σnx"        7.95068
   "n"                137.
   "MinX"             23.
                                  1/1
```

EXERCISE 6E.4

1 50 students sit for a mathematics test. Given the results in the table, estimate the mean score.

Score	0 - 9	10 - 19	20 - 29	30 - 39	40 - 49
Frequency	2	5	7	27	9

2 The table shows the petrol sales in one day by a number of city service stations.

a How many service stations were involved in the survey?

b Estimate the total amount of petrol sold for the day by the service stations.

c Find the approximate mean sales of petrol for the day.

Petrol sold, L (litres)	*Frequency*
$2000 \leqslant L < 3000$	4
$3000 \leqslant L < 4000$	4
$4000 \leqslant L < 5000$	9
$5000 \leqslant L < 6000$	14
$6000 \leqslant L < 7000$	23
$7000 \leqslant L < 8000$	16

3 Following is a record of the number of points Chloe scored in her basketball matches.

15	8	6	10	0	9	2	16	11	14	13	17	16	12	13	12	10
3	13	5	18	14	19	4	15	15	19	19	14	6	11	8	9	3
9	7	15	19	12	17	14										

a Find the mean number of points per match.

b Estimate the mean by grouping the data into the intervals:

i 0 - 4, 5 - 9, 10 - 14, 15 - 19 **ii** 0 - 3, 4 - 7, 8 - 11, 12 - 15, 16 - 19

c Comment on the accuracy of your answers from **a** and **b**.

4 Kylie pitched a softball 50 times. The speeds of her pitches are shown in the table.
Use technology to estimate the mean speed of her pitches.

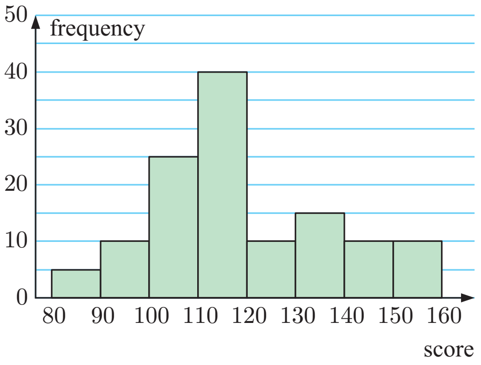
Speed (km h^{-1})	*Frequency*
80 - < 85	8
85 - < 90	14
90 - < 95	22
95 - < 100	6

5 The table shows the sizes of land blocks on a suburban street.
Use technology to estimate the mean land block size.

Land size (m^2)	*Frequency*
[500, 600)	5
[600, 700)	11
[700, 800)	23
[800, 900)	14
[900, 1000)	9

6 This frequency histogram illustrates the results of an aptitude test given to a group of people seeking positions in a company.

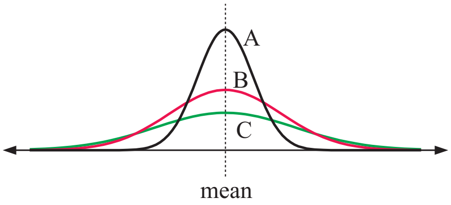
- **a** How many people sat for the test?
- **b** Estimate the mean score for the test.
- **c** What fraction of the people scored less than 100 for the test?
- **d** If the top 20% of the people are offered positions in the company, estimate the minimum mark required.



F MEASURING THE SPREAD OF DATA

To accurately describe a distribution we need to measure both its **centre** and its **spread** or **dispersion**.

The distributions shown have the same mean, but clearly they have different spreads. The A distribution has most scores close to the mean whereas the C distribution has the greatest spread.



We will examine three different measures of spread: the **range**, the **interquartile range** (**IQR**), and the **standard deviation**.

THE RANGE

The **range** is the difference between the maximum (largest) and the minimum (smallest) data value.

Example 10 Self Tutor

A library surveys 20 borrowers each day from Monday to Friday, and records the number who are not satisfied with the range of reading material. The results are: 3 7 6 8 11.

The following year the library receives a grant that enables the purchase of a large number of books. The survey is then repeated and the results are: 2 3 5 4 6.

Find the range of data in each survey.

The range is the maximum minus the minimum data value.

For the first survey, the range is $11 - 3 = 8$.

For the second survey, the range is $6 - 2 = 4$.

The **range** is not considered to be a particularly reliable measure of spread as it uses only two data values. It may be influenced by extreme values or outliers.

THE QUARTILES AND THE INTERQUARTILE RANGE

The median divides the ordered data set into two halves and these halves are divided in half again by the **quartiles**.

The middle value of the lower half is called the **lower quartile** or **25th percentile**. One quarter or 25% of the data have values less than or equal to the lower quartile. 75% of the data have values greater than or equal to the lower quartile.

The middle value of the upper half is called the **upper quartile** or **75th percentile**. One quarter or 25% of the data have values greater than or equal to the upper quartile. 75% of the data have values less than or equal to the upper quartile.

The **interquartile range** is the range of the middle half or 50% of the data.

interquartile range = upper quartile − lower quartile

The data set is thus divided into quarters by the lower quartile (Q_1), the median (Q_2), and the upper quartile (Q_3).

So, the interquartile range, $\mathbf{IQR = Q_3 - Q_1}$.

Example 11 Self Tutor

For the data set: 7, 3, 1, 7, 6, 9, 3, 8, 5, 8, 6, 3, 7, 1, 9 find the:

a median **b** lower quartile **c** upper quartile **d** interquartile range.

The ordered data set is: ~~1, 1, 3, 3, 3, 5, 6,~~ 6, ~~7, 7, 7, 8, 8, 9, 9~~ (15 of them)

a As $n = 15$, $\dfrac{n+1}{2} = 8$ $\therefore$ the median = 8th data value = 6

b/c As the median is a data value we now ignore it and split the remaining data into two:

$$\overbrace{1\ 1\ 3\ 3\ 3\ 5\ 6}^{\text{lower}}\quad\overbrace{7\ 7\ 7\ 8\ 8\ 9\ 9}^{\text{upper}}$$

Q_1 = median of lower half = 3

Q_3 = median of upper half = 8

d $\text{IQR} = Q_3 - Q_1 = 8 - 3 = 5$

Example 12 **Self Tutor**

For the data set: 6, 4, 9, 15, 5, 13, 7, 12, 8, 10, 4, 1, 13, 1, 6, 4, 5, 2, 8, 2 find:

a the median **b** Q_1 **c** Q_3 **d** the IQR.

The ordered data set is:

~~1 1 2 2 4 4 4 5 5~~ 6 6 ~~7 8 8 9 10 12 13 13 15~~ (20 of them)

a As $n = 20$, $\dfrac{n+1}{2} = 10.5$

$$\therefore \text{ the median} = \frac{\text{10th value } + \text{ 11th value}}{2} = \frac{6+6}{2} = 6$$

b/c As we have an even number of data values, we split the data into two:

$$\overbrace{1\ 1\ 2\ 2\ 4\ 4\ 4\ 5\ 5\ 6}^{\text{lower}}\quad \overbrace{6\ 7\ 8\ 8\ 9\ 10\ 12\ 13\ 13\ 15}^{\text{upper}}$$

$$\therefore \quad Q_1 = \frac{4+4}{2} = 4, \quad Q_3 = \frac{9+10}{2} = 9.5$$

d $\text{IQR} = Q_3 - Q_1$
$= 9.5 - 4$
$= 5.5$

EXERCISE 6F

1 For each of the following data sets, make sure the data is ordered and then find:

i the median **ii** the upper and lower quartiles
iii the range **iv** the interquartile range.

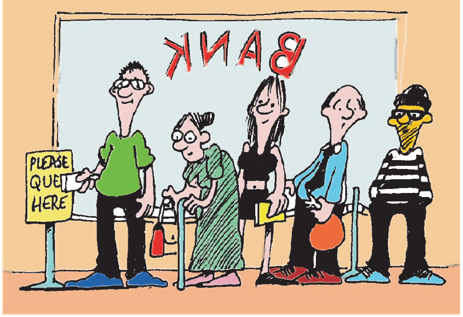
a 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 8, 8, 8, 9, 9

b 10, 12, 15, 12, 24, 18, 19, 18, 18, 15, 16, 20, 21, 17, 18, 16, 22, 14

c 21.8, 22.4, 23.5, 23.5, 24.6, 24.9, 25, 25.3, 26.1, 26.4, 29.5

2 The times spent (in minutes) by 20 people waiting in a queue at a bank for a teller were:

3.4 2.1 3.8 2.2 4.5 1.4 0 0 1.6 4.8
1.5 1.9 0 3.6 5.2 2.7 3.0 0.8 3.8 5.2



a Find the median waiting time and the upper and lower quartiles.

b Find the range and interquartile range of the waiting times.

c Copy and complete the following statements:

i "50% of the waiting times were greater than ...... minutes."

ii "75% of the waiting times were less than ...... minutes."

iii "The minimum waiting time was ...... minutes and the maximum waiting time was ...... minutes. The waiting times were spread over ...... minutes."

Example 13 Self Tutor

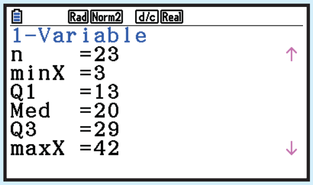
Consider the data set:
20, 31, 4, 17, 26, 9, 29, 37, 13, 42, 20, 18, 25, 7, 14, 3, 23, 16, 29, 38, 10, 33, 29

Use technology to find the:

a range **b** interquartile range.

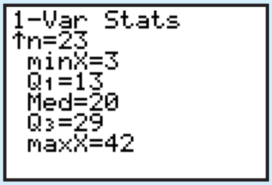
GRAPHICS CALCULATOR INSTRUCTIONS

Casio fx-CG20



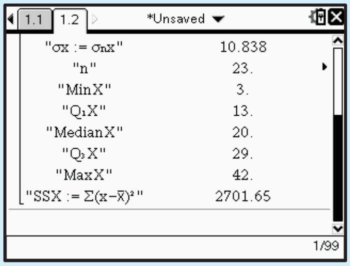
```
Rad Norm2  d/c Real
1-Variable
n    =23
minX =3
Q1   =13
Med  =20
Q3   =29
maxX =42
```

TI-84 Plus



```
1-Var Stats
↑n=23
 minX=3
 Q1=13
 Med=20
 Q3=29
 maxX=42
```

TI-*n*spire



```
1.1  1.2   *Unsaved
    "σx := σnx"        10.838
         "n"            23.
       "MinX"            3.
       "Q1X"            13.
     "MedianX"          20.
       "Q3X"            29.
       "MaxX"           42.
"SSX := Σ(x−x̄)²"      2701.65
                              1/99
```

a range = maximum − minimum
= 42 − 3
= 39

b IQR = $Q_3 - Q_1$
= 29 − 13
= 16

3 For the data set given, find using technology:

a the minimum value **b** the maximum value
c the median **d** the lower quartile
e the upper quartile **f** the range
g the interquartile range.

15	22	19	8	14	11
12	25	20	10	9	16
24	21	15	12	28	13
26	19	11	14	6	18
22	14	13	20	25	10

4 The heights of 20 ten year olds were recorded in centimetres:

109	111	113	114	114	118	119	122	122	124
124	126	128	129	129	131	132	135	138	138

a Using technology, find the:
i median height **ii** upper and lower quartiles of the data.

b Copy and complete the following statements:
i "Half of the children are no more than ...... cm tall."
ii "75% of the children are no more than ...... cm tall."

c Find the: **i** range **ii** interquartile range for the height of the ten year olds.

d Copy and complete: "The middle 50% of the children have heights spread over ...... cm."

5 Revisit the **Opening Problem** on page **158**.

a For the *Without fertiliser* data, find:
i the range **ii** the median **iii** the lower quartile
iv the upper quartile **v** the interquartile range.

b Repeat **a** for the *With fertiliser* data.

c Consider again the questions posed in the **Opening Problem**. Amend your solutions where appropriate.

G BOX AND WHISKER PLOTS

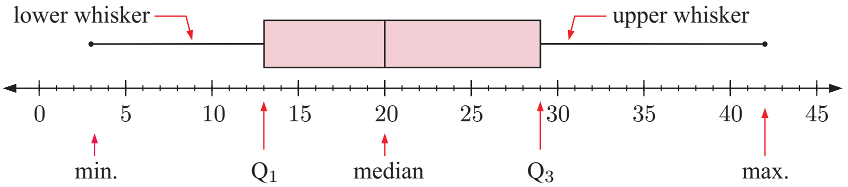
A **box and whisker plot** or simply **boxplot** is a visual display of some of the descriptive statistics of a data set. It shows:

- the minimum value
- the lower quartile (Q_1)
- the median (Q_2)
- the upper quartile (Q_3)
- the maximum value

These five numbers form the **five-number summary** of the data set.

For the data set in **Example 13** on page **185**, the five-number summary and boxplot are:

minimum $= 3$
$Q_1 = 13$
median $= 20$
$Q_3 = 29$
maximum $= 42$



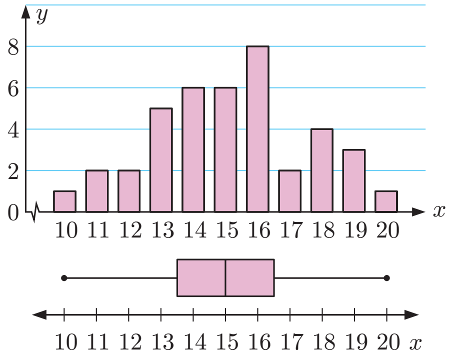
The rectangular box represents the 'middle' half of the data set.

The lower whisker represents the 25% of the data with smallest values.

The upper whisker represents the 25% of the data with greatest values.

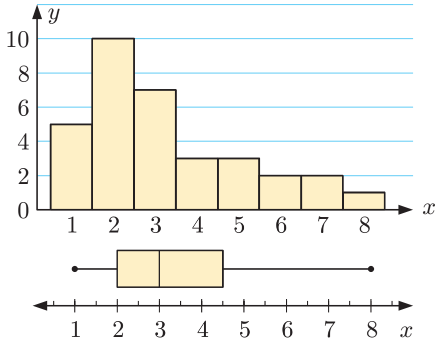
INTERPRETING A BOXPLOT

A set of data with a **symmetric distribution** will have a symmetric boxplot.



The whiskers of the boxplot are the same length and the median line is in the centre of the box.

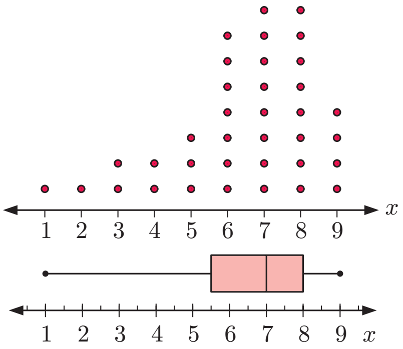
A set of data which is **positively skewed** will have a positively skewed boxplot.



The right whisker is longer than the left whisker and the median line is to the left of the box.

A set of data which is **negatively skewed** will have a negatively skewed boxplot.

The left whisker is longer than the right whisker and the median line is to the right of the box.



Click on the icons to explore boxplots further.

GAME

STATISTICS PACKAGE

Example 14

Self Tutor

Consider the data set: 8 2 3 9 6 5 3 2 2 6 2 5 4 5 5 6

a Construct the five-number summary for this data.

b Draw a boxplot.

c Find the: **i** range **ii** interquartile range of the data.

d Find the percentage of data values less than 3.

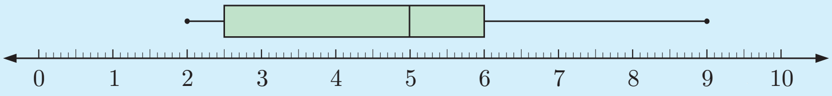
a The ordered data set is:

2 2 2 2 3 3 4 5 5 5 5 6 6 6 8 9 {16 data values}

$Q_1 = 2.5$ median $= 5$ $Q_3 = 6$

So the 5-number summary is: $\begin{cases} \text{minimum} = 2 & Q_1 = 2.5 \\ \text{median} = 5 & Q_3 = 6 \\ \text{maximum} = 9 & \end{cases}$

b

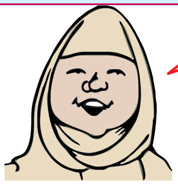


c **i** range = maximum − minimum
$= 9 - 2$
$= 7$

ii $\text{IQR} = Q_3 - Q_1$
$= 6 - 2.5$
$= 3.5$

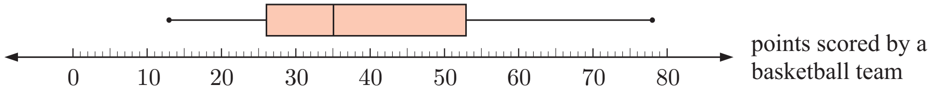
d 25% of the data values are less than 3.

This can be seen from the original data set. We cannot read it straight from the boxplot because the boxplot does not tell us that all of the data values are integers.



EXERCISE 6G.1

1 The boxplot below summarises the points scored by a basketball team.



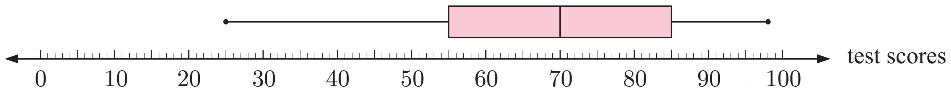
a Locate:

i the median **ii** the maximum value **iii** the minimum value

iv the upper quartile **v** the lower quartile.

b Calculate: **i** the range **ii** the interquartile range.

2 The boxplot below summarises the class results for a test out of 100 marks.



a Copy and complete the following statements about the test results:

i The highest mark scored for the test was ...., and the lowest mark was ....

ii Half of the class scored a mark greater than or equal to ....

iii The top 25% of the class scored at least .... marks for the test.

iv The middle half of the class had scores between .... and .... for this test.

b Find the range of the data set.

c Find the interquartile range of the data set.

d Estimate the mean mark for these test scores.

3 For the following data sets:

i construct a 5-number summary **ii** draw a boxplot

iii find the range **iv** find the interquartile range.

a 3, 4, 5, 5, 5, 6, 6, 6, 7, 7, 8, 8, 9, 10

b 3, 7, 0, 1, 4, 6, 8, 8, 8, 9, 7, 5, 6, 8, 7, 8, 8, 2, 9

c 23, 44, 31, 33, 26, 17, 30, 35, 47, 31, 51, 47, 20, 31, 28, 49, 26, 49

4 Enid examines a new variety of bean and counts the number of beans in 33 pods. Her results were:

5, 8, 10, 4, 2, 12, 6, 5, 7, 7, 5, 5, 5, 13, 9, 3, 4, 4, 7, 8, 9, 5, 5, 4, 3, 6, 6, 6, 6, 9, 8, 7, 6

a Find the median, lower quartile, and upper quartile of the data set.

b Find the interquartile range of the data set.

c Draw a boxplot of the data set.

5 Ranji counts the number of bolts in several boxes and tabulates the data as follows:

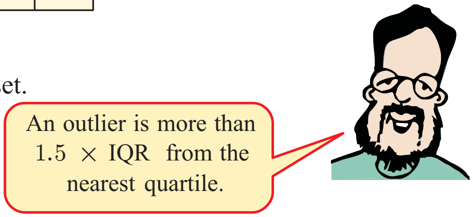
Number of bolts	33	34	35	36	37	38	39	40
Frequency	1	5	7	13	12	8	0	1

a Find the five-number summary for this data set.

b Find the **i** range **ii** IQR for this data set.

c Draw a boxplot of the data set.

d Are there any outliers in this data?


An outlier is more than 1.5 × IQR from the nearest quartile.

PARALLEL BOXPLOTS

A parallel boxplot enables us to make a *visual comparison* of the distributions of two data sets. We can easily compare descriptive statistics such as their median, range, and interquartile range.

Example 15

Self Tutor

A hospital is trialling a new anaesthetic drug and has collected data on how long the new and old drugs take before the patient becomes unconscious. They wish to know which drug acts faster and which is more reliable.

Old drug times (s): 8, 12, 9, 8, 16, 10, 14, 7, 5, 21,
13, 10, 8, 10, 11, 8, 11, 9, 11, 14

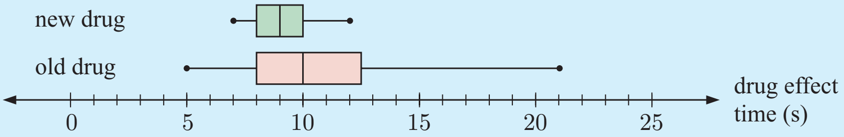
New drug times (s): 8, 12, 7, 8, 12, 11, 9, 8, 10, 8,
10, 9, 12, 8, 8, 7, 10, 7, 9, 9

Prepare a parallel boxplot for the data sets and use it to compare the two drugs for speed and reliability.

The 5-number summaries are:

For the old drug: $\text{min}_x = 5$, $Q_1 = 8$, $\text{median} = 10$, $Q_3 = 12.5$, $\text{max}_x = 21$

For the new drug: $\text{min}_x = 7$, $Q_1 = 8$, $\text{median} = 9$, $Q_3 = 10$, $\text{max}_x = 12$



Using the median, 50% of the time the new drug takes 9 seconds or less, compared with 10 seconds for the old drug. We conclude that the new drug is generally a little quicker.

Comparing the spread:

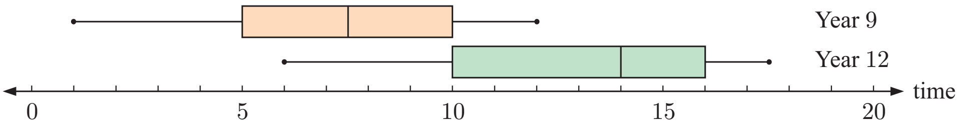
$$\begin{aligned}\text{range for old drug} &= 21 - 5\\ &= 16\\ \text{IQR} &= Q_3 - Q_1\\ &= 12.5 - 8\\ &= 4.5\end{aligned}$$

$$\begin{aligned}\text{range for new drug} &= 12 - 7\\ &= 5\\ \text{IQR} &= Q_3 - Q_1\\ &= 10 - 8\\ &= 2\end{aligned}$$

The new drug times are less 'spread out' than the old drug times. They are more predictable or reliable.

EXERCISE 6G.2

1 The following side-by-side boxplots compare the times students in years 9 and 12 spend on homework.



a Copy and complete:

Statistic	*Year 9*	*Year 12*
minimum		
Q_1		
median		
Q_3		
maximum		

b For each group, determine the:

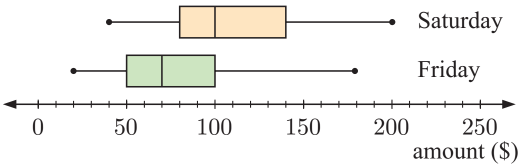
i range **ii** interquartile range.

c Are the following true or false, or is there not enough information to tell?

i On average, Year 12 students spend about twice as much time on homework as Year 9 students.

ii Over 25% of Year 9 students spend less time on homework than all Year 12 students.

2 The amounts of money withdrawn from an ATM were recorded on a Friday and a Saturday. The results are displayed on the parallel boxplot alongside.

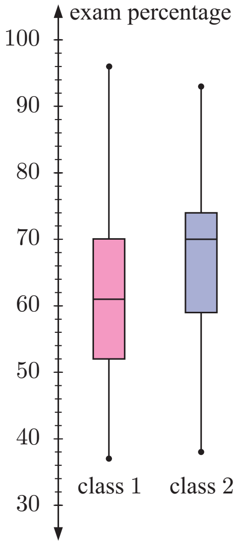


a Find the five-number summary for each set of data.

b For each data set, determine the

i range **ii** interquartile range.

3 After the final examination, two classes studying the same subject compiled this parallel boxplot to show their results.



a In which class was:

i the highest mark **ii** the lowest mark

iii there a larger spread of marks?

b Find the interquartile range of class 1.

c Find the range of class 2.

d If students who scored at least 70% received an achievement award, what percentage of students received an award in:

i class 1 **ii** class 2?

e Describe the distribution of marks in:

i class 1 **ii** class 2.

f Copy and complete:
The students in class ...... generally scored higher marks.
The marks in class ...... were more varied.

4 Below are the durations, in minutes, of Paul and Redmond's last 25 mobile phone calls.

Paul: 1.7, 2.0, 3.9, 3.4, 0.9, 1.4, 2.5, 1.1, 5.1, 4.2, 1.5, 2.6, 0.8, 4.0, 1.5, 1.0, 2.9, 3.2, 2.5, 0.8, 1.8, 3.1, 6.9, 2.3, 1.2

Redmond: 2.0, 4.8, 1.2, 7.5, 3.2, 5.7, 3.9, 0.2, 2.7, 6.8, 3.4, 5.2, 3.2, 7.2, 1.7, 11.5, 4.0, 2.4, 3.7, 4.2, 10.7, 3.0, 2.0, 0.9, 5.7

a Find the five-number summary for each of the data sets.

b Display the data in a parallel boxplot.

c Compare and comment on the distributions of the data.

5 Shane and Brett play in the same cricket team and are fierce but friendly rivals when it comes to bowling. During a season the number of wickets taken in each innings by the bowlers was:

Shane:	1	6	2	0	3	4	1	4	2	3	0	3	2	4	3	4	3	3
	3	4	2	4	3	2	3	3	0	5	3	5	3	2	4	3	4	3
Brett:	7	2	4	8	1	3	4	2	3	0	5	3	5	2	3	1	2	0
	4	3	4	0	3	3	0	2	5	1	1	2	2	5	1	4	0	1

a Is the variable discrete or continuous?

b Enter the data into a graphics calculator or statistics package.

c Produce a vertical column graph for each data set.

d Describe the shape of each distribution.

e Compare the measures of the centre of each distribution.

f Compare the spreads of each distribution.

g Obtain a side-by-side boxplot.

h What conclusions can be drawn from the data?

6 A manufacturer of light globes claims that their new design has a 20% longer life than those they are presently selling. Forty of each globe are randomly selected and tested. Here are the results to the nearest hour:

	103	96	113	111	126	100	122	110	84	117	103	113	104	104
Old type:	111	87	90	121	99	114	105	121	93	109	87	118	75	111
	87	127	117	131	115	116	82	130	113	95	108	112		
	146	131	132	160	128	119	133	117	139	123	109	129	109	131
New type:	191	117	132	107	141	136	146	142	123	144	145	125	164	125
	133	124	153	129	118	130	134	151	145	131	133	135		

a Is the variable discrete or continuous?

b Enter the data into a graphics calculator or statistics package.

c Compare the measures of centre and spread.

d Obtain a side-by-side boxplot.

e Describe the shape of each distribution.

f What conclusions, if any, can be drawn from the data?

OUTLIERS (EXTENSION)

We have seen that **outliers** are extraordinary data that are separated from the main body of the data.

A commonly used test to identify outliers involves the calculation of upper and lower boundaries:

- **The upper boundary = upper quartile + 1.5 × IQR**.
  Any data larger than the upper boundary is an outlier.
- **The lower boundary = lower quartile − 1.5 × IQR**.
  Any data smaller than the lower boundary is an outlier.

Outliers are marked with an asterisk on a boxplot. It is possible to have more than one outlier at either end.

Each whisker extends to the last value that is not an outlier.

Example 16 Self Tutor

Test the following data for outliers and hence construct a boxplot for the data:

3, 7, 8, 8, 5, 9, 10, 12, 14, 7, 1, 3, 8, 16, 8, 6, 9, 10, 13, 7

The ordered data set is:

1 3 3 5 6 7 7 7 8 8 8 8 9 9 10 10 12 13 14 16 $\{n = 20\}$

$\text{Min}_x = 1$ $\quad Q_1 = 6.5$ $\quad$ median $= 8$ $\quad Q_3 = 10$ $\quad \text{Max}_x = 16$

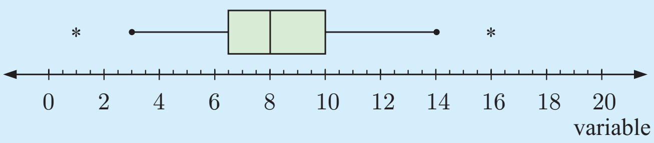
$\text{IQR} = Q_3 - Q_1 = 3.5$

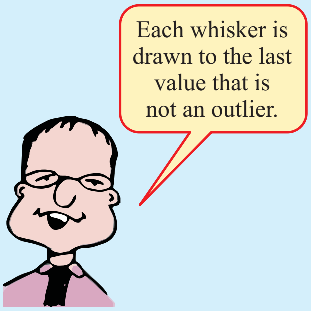
Test for outliers: upper boundary and lower boundary

$$\begin{aligned} &= \text{upper quartile} + 1.5 \times \text{IQR} & &= \text{lower quartile} - 1.5 \times \text{IQR} \\ &= 10 + 1.5 \times 3.5 & &= 6.5 - 1.5 \times 3.5 \\ &= 15.25 & &= 1.25 \end{aligned}$$

16 is above the upper boundary, so it is an outlier.
1 is below the lower boundary, so it is an outlier.

So, the boxplot is:





Each whisker is drawn to the last value that is not an outlier.

EXERCISE 6G.3

1 A set of data has a lower quartile of 31.5, a median of 37, and an upper quartile of 43.5 .

- **a** Calculate the interquartile range for this data set.
- **b** Calculate the boundaries that identify outliers.
- **c** The smallest values of the data set are 13 and 20. The largest values are 52 and 55. Which of these would be outliers?
- **d** Draw a boxplot of the data set.

2 James goes bird watching for 25 days. The number of birds he sees each day are:

12, 5, 13, 16, 8, 10, 12, 18, 9, 11, 14, 14, 22, 9, 10, 7, 9, 11, 13, 7, 10, 6, 13, 3, 8

- **a** Find the median, lower quartile, and upper quartile of the data set.
- **b** Find the interquartile range of the data set.
- **c** What are the lower and upper boundaries for outliers?
- **d** Are there any outliers?
- **e** Draw a boxplot of the data set.

H CUMULATIVE FREQUENCY GRAPHS

Sometimes, in addition to finding the median, it is useful to know the number or proportion of scores that lie above or below a particular value. In such situations we can construct a **cumulative frequency distribution table** and use a graph called a **cumulative frequency graph** to represent the data.

The cumulative frequencies are plotted and the points joined by a smooth curve. This differs from an ogive or cumulative frequency polygon where two points are joined by straight lines.

PERCENTILES

A **percentile** is the score below which a certain percentage of the data lies.

For example:

- the 85th percentile is the score below which 85% of the data lies.
- If your score in a test is the 95th percentile, then 95% of the class have scored less than you.

Notice that:

- the **lower quartile** ($\mathbf{Q_1}$) is the 25th percentile
- the **median** ($\mathbf{Q_2}$) is the 50th percentile
- the **upper quartile** ($\mathbf{Q_3}$) is the 75th percentile.

A cumulative frequency graph provides a convenient way to find percentiles.

Example 17 Self Tutor

The data shows the results of the women's marathon at the 2008 Olympics, for all competitors who finished the race.

Finishing time t	*Frequency*
2 h 26 $\leqslant t <$ 2 h 28	8
2 h 28 $\leqslant t <$ 2 h 30	3
2 h 30 $\leqslant t <$ 2 h 32	9
2 h 32 $\leqslant t <$ 2 h 34	11
2 h 34 $\leqslant t <$ 2 h 36	12
2 h 36 $\leqslant t <$ 2 h 38	7
2 h 38 $\leqslant t <$ 2 h 40	5
2 h 40 $\leqslant t <$ 2 h 48	8
2 h 48 $\leqslant t <$ 2 h 56	6

- **a** Construct a cumulative frequency distribution table.
- **b** Represent the data on a cumulative frequency graph.
- **c** Use your graph to estimate the:
  - **i** median finishing time
  - **ii** number of competitors who finished in less than 2 hours 35 minutes
  - **iii** percentage of competitors who took more than 2 hours 39 minutes to finish
  - **iv** time taken by a competitor who finished in the top 20% of runners completing the marathon.

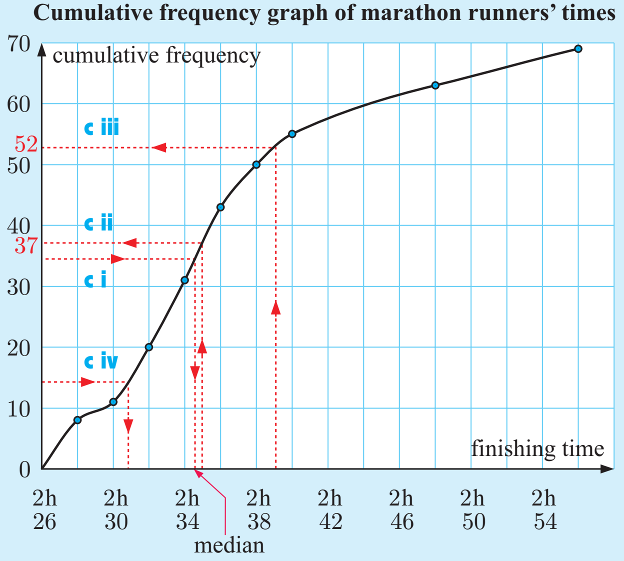
a

Finishing time t	*Frequency*	*Cumulative frequency*
2 h 26 $\leqslant t <$ 2 h 28	8	8
2 h 28 $\leqslant t <$ 2 h 30	3	11
2 h 30 $\leqslant t <$ 2 h 32	9	20
2 h 32 $\leqslant t <$ 2 h 34	11	31
2 h 34 $\leqslant t <$ 2 h 36	12	43
2 h 36 $\leqslant t <$ 2 h 38	7	50
2 h 38 $\leqslant t <$ 2 h 40	5	55
2 h 40 $\leqslant t <$ 2 h 48	8	63
2 h 48 $\leqslant t <$ 2 h 56	6	69

$8 + 3 = 11$ competitors completed the marathon in less than 2 hours 30 minutes.

50 competitors completed the marathon in less than 2 hours 38 minutes.

b



The cumulative frequency gives a *running total* of the number of runners finishing by a given time.

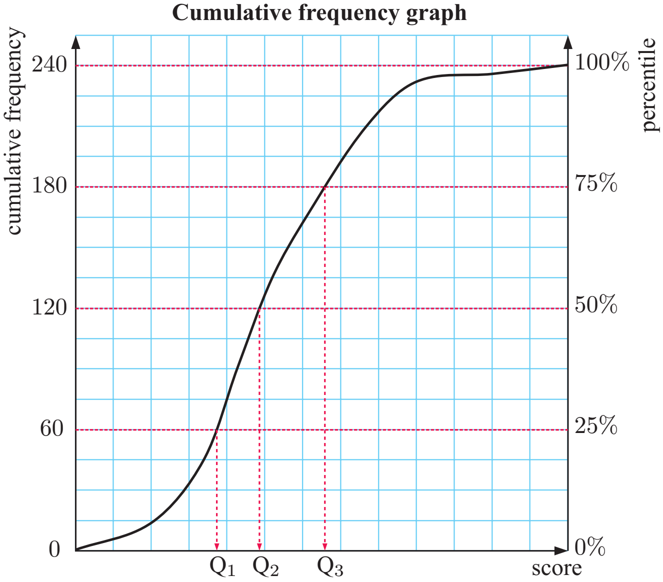
c **i** The median is estimated using the 50th percentile. As 50% of 69 is 34.5, we start with the cumulative frequency of 34.5 and find the corresponding time.
The median is approximately 2 hours 34.5 min.

ii There are approximately 37 competitors who took less than 2 h 35 min to complete the race.

iii There are $69 - 52 = 17$ competitors who took more than 2 hours 39 min.
So $\frac{17}{69} \approx 26.4\%$ took more than 2 hours 39 min.

iv The time taken is estimated using the 20th percentile. As 20% of 69 is 13.8, we find the time corresponding to a cumulative frequency of approximately 14.
The top 20% of competitors took less than 2 hours 31 minutes.

Another way to calculate percentiles is to add a separate scale to a cumulative frequency graph. On the graph alongside, the cumulative frequency is read from the axis on the left side, and each value corresponds to a percentile on the right side.



EXERCISE 6H

1 The examination scores of a group of students are shown in the table. Draw a cumulative frequency graph for the data and use it to find:

a the median examination mark

b how many students scored less than 65 marks

c how many students scored between 50 and 70 marks

d how many students failed, given that the pass mark was 45

e the credit mark, given that the top 16% of students were awarded credits.

Score	*Frequency*
$10 \leqslant x < 20$	2
$20 \leqslant x < 30$	5
$30 \leqslant x < 40$	7
$40 \leqslant x < 50$	21
$50 \leqslant x < 60$	36
$60 \leqslant x < 70$	40
$70 \leqslant x < 80$	27
$80 \leqslant x < 90$	9
$90 \leqslant x < 100$	3

2 A botanist has measured the heights of 60 seedlings and has presented her findings on the cumulative frequency graph below.

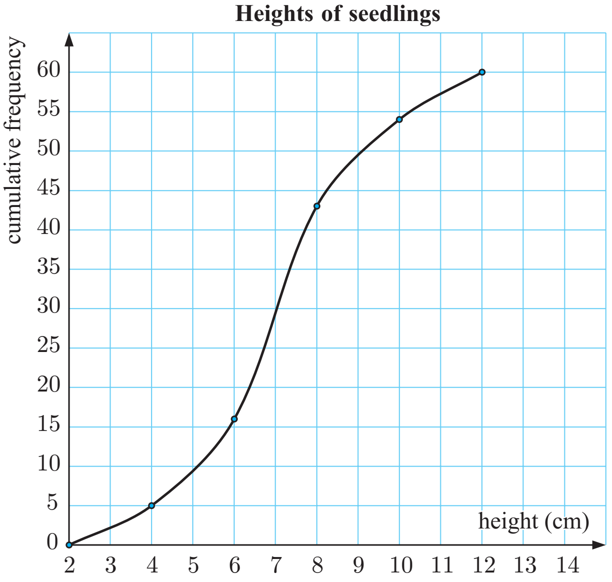
a How many seedlings have heights of 5 cm or less?

b What percentage of seedlings are taller than 8 cm?

c Find the median height.

d Find the interquartile range for the heights.

e Copy and complete:
"90% of the seedlings are shorter than ......"



3 The following table summarises the age groups of car drivers involved in accidents in a city for a given year. Draw a cumulative frequency graph for the data and use it to estimate:

a the median age of the drivers involved in accidents

b the percentage of drivers involved in accidents who had an age of 23 or less

c the probability that a driver involved in an accident is:

i aged 27 years or less **ii** aged 27 years.

Age (in years)	*Number of accidents*
$16 \leqslant x < 20$	59
$20 \leqslant x < 25$	82
$25 \leqslant x < 30$	43
$30 \leqslant x < 35$	21
$35 \leqslant x < 40$	19
$40 \leqslant x < 50$	11
$50 \leqslant x < 60$	24
$60 \leqslant x < 80$	41

4 The following data shows the lengths of 30 trout caught in a lake during a fishing competition. The measurements were rounded down to the next centimetre.

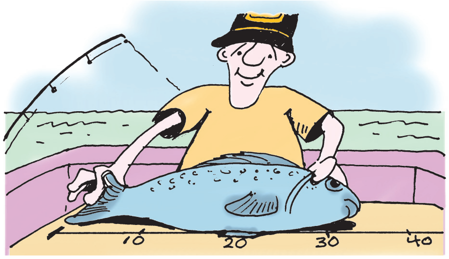
31 38 34 40 24 33 30 36 38 32 35 32 36 27 35
40 34 37 44 38 36 34 33 31 38 35 36 33 33 28

a Construct a cumulative frequency table for trout lengths, x cm, using the intervals $24 \leqslant x < 27$, $27 \leqslant x < 30$, and so on.

b Draw a cumulative frequency graph for the data.

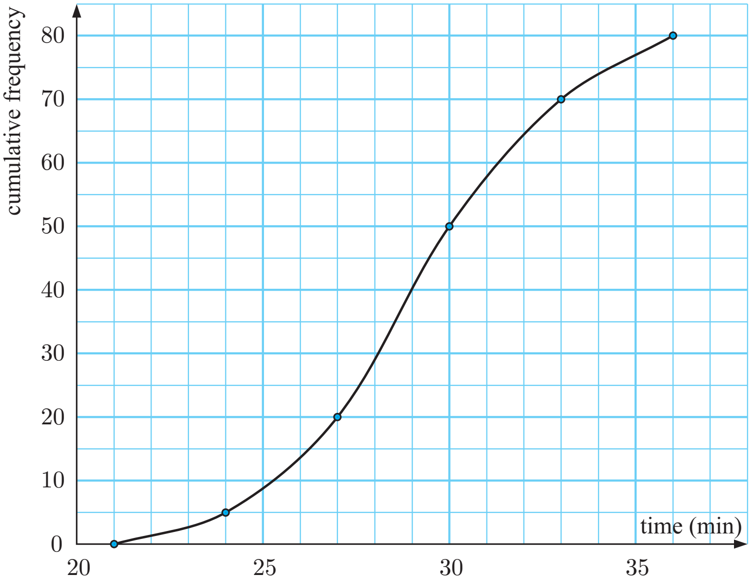
c Hence estimate the median length.

d Use the original data to find its median and compare your answer with **c**. Comment on your results.



5 The following cumulative frequency graph displays the performance of 80 competitors in a cross-country race.

Cross-country race times



Find:

a the lower quartile time

b the median

c the upper quartile

d the interquartile range

e an estimate of the 40th percentile.

6 The table shows the lifetimes of a sample of electric light globes.
Draw a cumulative frequency graph for the data and use it to estimate:

a the median life of a globe

b the percentage of globes which had a life of 2700 hours or less

c the number of globes which had a life between 1500 and 2500 hours.

Life (hours)	*Number of globes*
$0 \leqslant l < 500$	5
$500 \leqslant l < 1000$	17
$1000 \leqslant l < 2000$	46
$2000 \leqslant l < 3000$	79
$3000 \leqslant l < 4000$	27
$4000 \leqslant l < 5000$	4

7 The following frequency distribution was obtained by asking 50 randomly selected people to measure the lengths of their feet. Their answers were given to the nearest centimetre.

Foot length (cm)	20	21	22	23	24	25	26	27	28	29	30
Frequency	1	1	0	3	5	13	17	7	2	0	1

a Between what limits are scores rounded to 20 cm?

b Rewrite the frequency table to show the data in class intervals like the one found in **a**.

c Hence draw a cumulative frequency graph for the data.

d Estimate: **i** the median foot length

ii the number of people with foot length 26 cm or more.

I STANDARD DEVIATION

The problem with using the range and the IQR as measures of spread or dispersion of scores is that both of them only use two values in their calculation. Some data sets have their spread characteristics hidden when the range or IQR are quoted, and so we need a better way of describing spread.

The **standard deviation** of a distribution takes into account the **deviation** of **each score** from the mean. It is therefore a good measure of the **dispersion** of the data.

Consider a data set of n values: $x_1, x_2, x_3, x_4, ..., x_n,$ with mean $\overline{x}$.

For a data set of n values, $s_n = \sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}$ is called the **standard deviation**.

Notice in this formula that:

- $(x_i - \overline{x})^2$ is a measure of how far x_i deviates from $\overline{x}$.
- If $\sum_{i=1}^{n}(x_i - \overline{x})^2$ is small, it will indicate that most of the data values are close to $\overline{x}$.
- Dividing by n gives an indication of how far, on average, the data is from the mean.
- The square root is used to correct the units.

The standard deviation is a **non-resistant** measure of spread. This is due to its dependence on the mean of the sample and because extreme data values will give large values for $(x_i - \overline{x})^2$. It is only a useful measure if the distribution is close to symmetrical. The IQR and percentiles are more appropriate tools for measuring spread if the distribution is considerably skewed.

INVESTIGATION 4 STANDARD DEVIATION

A group of 5 students is chosen from each of three schools, to test their ability to solve puzzles.

The 15 students are each given a series of puzzles and two hours to solve as many as they can individually.

The results were:	School A:	7, 7, 7, 7, 7
	School B:	5, 6, 7, 8, 9
	School C:	3, 5, 7, 9, 11

What to do:

1 Show that the mean and median for each school is 7.

2 Given the mean $\overline{x} = 7$ for each group, complete a table like the one following, for each school:

School A

Score (x_i)	*Deviation* $(x_i - \overline{x})$	*Square of deviation* $(x_i - \overline{x})^2$
7		
7		
7		
7		
7		
Sum		

3 Calculate the standard deviation $\sqrt{\frac{\sum(x_i - \overline{x})^2}{n}}$ for each group.

Check that your results match the following table:

School	*Mean*	*Standard deviation*
A	7	0
B	7	$\sqrt{2}$
C	7	$\sqrt{8}$

4 Use the table above to compare the performances of the different schools.

5 A group of 5 students from a higher year level at school C are given the same test. They each score 2 more than the students in the lower year group, so their scores are: 5, 7, 9, 11, 13.

a Find the mean and standard deviation for this set.

b Comment on the effect of adding 2 to each member of a data set.

6 A group of 5 teachers from B decide to show their students how clever they are. They complete twice as many puzzles as each of their students, so their scores are: 10, 12, 14, 16, 18.

a Find the mean and standard deviation for this set.

b Comment on the effect of doubling each member of a data set.

In this course you are only expected to use technology to calculate standard deviation. However, we present both methods in the following example so you can see how it works!

GRAPHICS CALCULATOR INSTRUCTIONS

STATISTICS PACKAGE

SPREADSHEET

Example 18 Self Tutor

Calculate the standard deviation of the data set: 2, 5, 4, 6, 7, 5, 6.

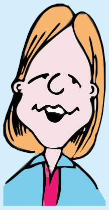
$$\overline{x} = \frac{2+5+4+6+7+5+6}{7} = 5$$

$$s = \sqrt{\frac{\sum(x-\overline{x})^2}{n}}$$

$$= \sqrt{\frac{16}{7}} \approx 1.51$$

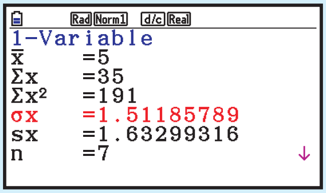
Score (x)	$x - \overline{x}$	$(x-\overline{x})^2$
2	-3	9
4	-1	1
5	0	0
5	0	0
6	1	1
6	1	1
7	2	4
35		16

Make sure you always use the standard deviation of the **population** as highlighted in the screenshots.



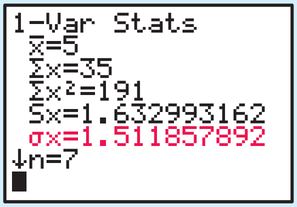
The following screendumps indicate the result when we calculate the standard deviation for this data set:

Casio fx-CG20



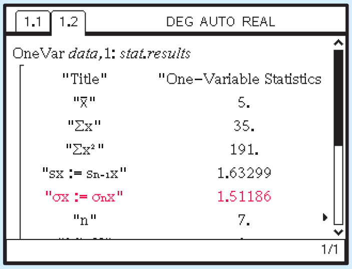
```
Rad Norm1 d/c Real
1-Variable
x̄    =5
Σx   =35
Σx²  =191
σx   =1.51185789
sx   =1.63299316
n    =7
```

TI-84 Plus



```
1-Var Stats
 x̄=5
 Σx=35
 Σx²=191
 Sx=1.632993162
 σx=1.511857892
↓n=7
```

TI-*n*spire



```
1.1  1.2        DEG AUTO REAL
OneVar data,1: stat.results
    "Title"        "One-Variable Statistics
    "x̄"             5.
    "Σx"            35.
    "Σx²"           191.
  "sx := sn-1x"     1.63299
  "σx := σnx"       1.51186
    "n"             7.
                                   1/1
```

EXERCISE 6I.1

1 Use technology to find the standard deviation of the following data sets:

a 5, 8, 6, 9, 6, 6, 4, 7 **b** 22, 19, 28, 20, 15, 27, 23, 26, 32, 26, 21, 30

2 A company recorded the following weekly petrol usage (in litres) by its salespersons:

62, 40, 52, 48, 64, 55, 44, 75, 40, 68, 60, 42, 70, 49, 56

Use technology to find the mean and standard deviation of this data.

3 The weights of a group of cooking chickens in kilograms are:

1.5, 1.8, 1.7, 1.4, 1.7, 1.8, 2.0, 1.5, 1.6, 1.6, 1.9, 1.7, 1.4, 1.7, 1.8, 2.0

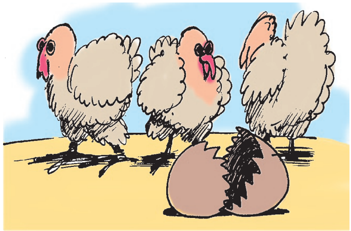
Use technology to find the mean and standard deviation of weights.

4 The heights in cm of seven junior footballers are: 179, 164, 159, 171, 168, 168, 174.

a Find the mean and standard deviation for this group.

b When measured one year later, each footballer had grown by exactly 5 cm. Find the new mean and standard deviation.

c Comment on your results in general terms.



5 The weights of ten young turkeys to the nearest 0.1 kg are:
0.8, 1.1, 1.2, 0.9, 1.2, 1.2, 0.9, 0.7, 1.0, 1.1

- **a** Find the mean and standard deviation for the weights of the turkeys.
- **b** After being fed a special diet for one month, the weights of the turkeys doubled. Find the new mean and standard deviation.
- **c** Comment on your results.

6 The following table shows the decrease in cholesterol levels in 6 volunteers after a two week trial of special diet and exercise.

Volunteer	A	B	C	D	E	F
Decrease in cholesterol	0.8	0.6	0.7	0.8	0.4	2.8

- **a** Find the standard deviation of the data.
- **b** Which of the data values is an outlier?
- **c** Recalculate the standard deviation with the outlier removed.
- **d** Discuss the effect of an extreme value on the standard deviation.

STANDARD DEVIATION FOR GROUPED DATA

For **continuous** data, or data that has been grouped in **classes**, we use the **mid-interval values** to represent all data in that interval.

GRAPHICS CALCULATOR INSTRUCTIONS

Example 19 Self Tutor

Use technology to estimate the standard deviation for this distribution of examination scores:

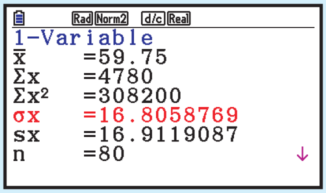
Mark	*Frequency*	*Mark*	*Frequency*
0 - 9	1	50 - 59	16
10 - 19	1	60 - 69	24
20 - 29	2	70 - 79	13
30 - 39	4	80 - 89	6
40 - 49	11	90 - 99	2

In order to estimate the standard deviation of already grouped data, the mid-interval values are used to represent all data in that interval.

We then use technology to estimate the standard deviation.

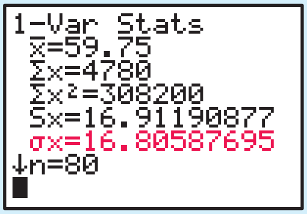
Class interval	*Mid-interval value*	*Frequency*	*Class interval*	*Mid-interval value*	*Frequency*
0 - 9	4.5	1	50 - 59	54.5	16
10 - 19	14.5	1	60 - 69	64.5	24
20 - 29	24.5	2	70 - 79	74.5	13
30 - 39	34.5	4	80 - 89	84.5	6
40 - 49	44.5	11	90 - 99	94.5	2

Casio fx-CG20



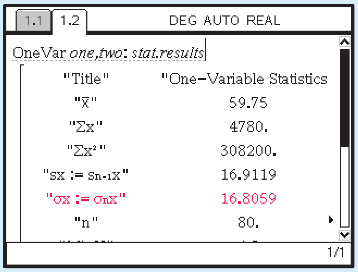
```
1-Variable
x̄       =59.75
Σx      =4780
Σx²     =308200
σx      =16.8058769
sx      =16.9119087
n       =80
```

TI-84 Plus



```
1-Var Stats
 x̄=59.75
 Σx=4780
 Σx²=308200
 Sx=16.91190877
 σx=16.80587695
↓n=80
```

TI-*n*spire



```
OneVar one,two: stat.results
   "Title"          "One-Variable Statistics"
    "x̄"                   59.75
    "Σx"                  4780.
    "Σx²"                 308200.
"sx := sn-1x"             16.9119
"σx := σnx"               16.8059
    "n"                   80.
```

The standard deviation $s \approx 16.8$.

EXERCISE 6I.2

1 The workers at a factory were asked how many children they had. The results are shown in the table below.

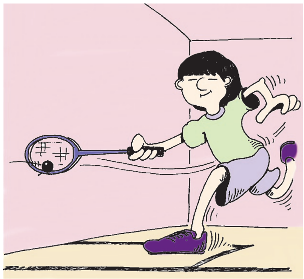
Number of children	0	1	2	3	4	5	6	7
Frequency	14	18	13	5	3	2	2	1

Find the mean and standard deviation of the data.

2 The ages of squash players at the Junior National Squash Championship are given below.

Age	11	12	13	14	15	16	17	18
Frequency	2	1	4	5	6	4	2	1

Find the mean and standard deviation of the ages.



3 The local Health and Fitness Centre recorded the following number of clients per week during the last year:

Calculate the average number of clients per week and the standard deviation from this number.

Number of clients	*Frequency*
36	2
39	5
44	9
45	11
46	15
48	5
50	4
52	1
Total	52

4 The lengths of 30 randomly selected 12-day old babies were measured and the following data obtained:

Length (cm)	[40, 42)	[42, 44)	[44, 46)	[46, 48)	[48, 50)	[50, 52)	[52, 54)
Frequency	1	1	3	7	11	5	2

Estimate the mean length and the standard deviation of the lengths.

5 The weekly wages (in dollars) of 200 steel workers are given alongside.
Estimate the mean and the standard deviation of the data.

Wage ($)	*Number of workers*
360 - 369.99	17
370 - 379.99	38
380 - 389.99	47
390 - 399.99	57
400 - 409.99	18
410 - 419.99	10
420 - 429.99	10
430 - 439.99	3

6 The hours worked last week by 40 employees of a local clothing factory were as follows:

38 40 46 32 41 39 44 38 40 42 38 40 43 41
47 36 38 39 34 40 48 30 49 40 40 43 45 36
35 39 42 44 48 36 38 42 46 38 39 40

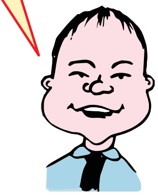
a Calculate the mean and standard deviation for this data.

b Now group the data into classes 30 - 33, 34 - 37, and so on. Calculate the mean and standard deviation using these groups. Examine any differences in the two sets of answers.

c Draw a cumulative frequency graph for the data and determine its interquartile range.

d Represent this data on a boxplot.

Since the data is continuous, we use the intervals 29.5 - 33.5, 33.5 - 37.5, ... for the cumulative frequency graph.



7 A traffic survey by the highways department revealed that the following numbers of vehicles passed through a suburban intersection in 15 minute intervals during the day.

a Estimate the mean and the standard deviation for the data.

b Draw a cumulative frequency graph of the data and determine its interquartile range.

Number of vehicles	*Frequency*
1 - 5	4
6 - 10	16
11 - 15	22
16 - 20	28
21 - 25	14
26 - 30	9
31 - 35	5
36 - 40	2

COMPARING THE SPREAD OF TWO DATA SETS

We have seen how the **mean** of two data sets is a useful comparison of their centres. To compare the spread or dispersion of two data sets we can use their **standard deviations**.

Example 20 Self Tutor

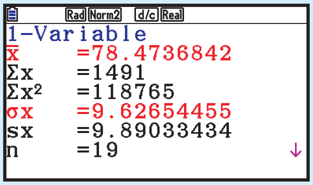
The following exam results were recorded by two classes of students studying Spanish:

Class A: 64 69 74 67 78 88 76 90 89 84 83 87 78 80 95 75 55 78 81
Class B: 94 90 88 81 86 96 92 93 88 72 94 61 87 90 97 95 77 77 82 90

Compare the results of the two classes including their spread.

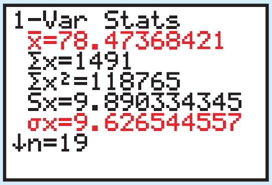
Class A:

Casio fx-CG20



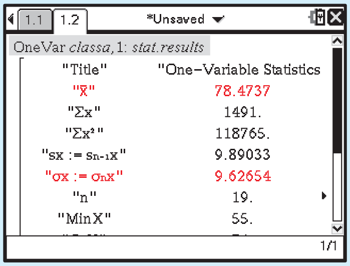
```
1-Variable
x̄    =78.4736842
Σx   =1491
Σx²  =118765
σx   =9.62654455
sx   =9.89033434
n    =19
```

TI-84 Plus



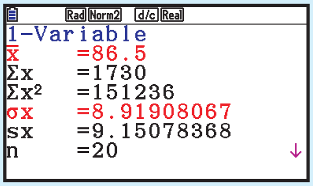
```
1-Var Stats
 x̄=78.47368421
 Σx=1491
 Σx²=118765
 Sx=9.890334345
 σx=9.626544557
↓n=19
```

TI-*nspire*

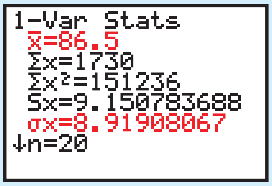


```
OneVar classa,1: stat.results
    "Title"        "One-Variable Statistics
    "x̄"            78.4737
    "Σx"           1491.
    "Σx²"          118765.
  "sx := sn-1x"    9.89033
  "σx := σnx"      9.62654
    "n"            19.
    "MinX"         55.
```

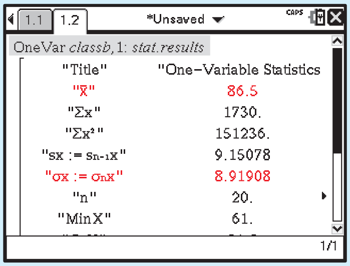
Class B:



```
1-Variable
x̄    =86.5
Σx   =1730
Σx²  =151236
σx   =8.91908067
sx   =9.15078368
n    =20
```



```
1-Var Stats
 x̄=86.5
 Σx=1730
 Σx²=151236
 Sx=9.150783688
 σx=8.91908067
↓n=20
```



```
OneVar classb,1: stat.results
    "Title"        "One-Variable Statistics
    "x̄"            86.5
    "Σx"           1730.
    "Σx²"          151236.
  "sx := sn-1x"    9.15078
  "σx := σnx"      8.91908
    "n"            20.
    "MinX"         61.
```

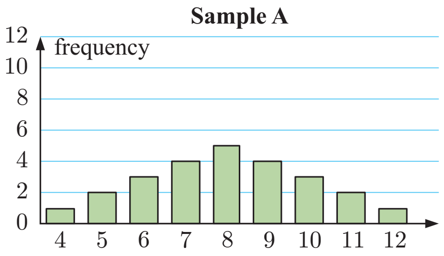
	Mean	*Standard deviation*
Class A	78.5	9.63
Class B	86.5	8.92

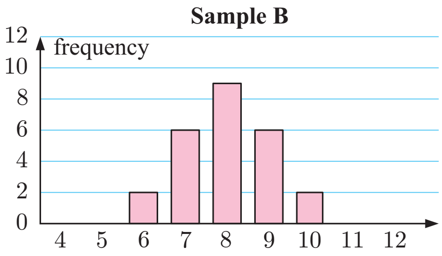
Class B has a higher mean than class A, indicating that the students in class B generally performed better in the exam.

Class A has a higher standard deviation than class B, indicating that the results in class A were more dispersed.

EXERCISE 6I.3

1 The column graphs show two distributions:





a By looking at the graphs, which distribution appears to have wider spread?

b Find the mean of each sample.

c Find the standard deviation of each sample. Comment on your answers.

2 The number of points scored by Andrew and Brad in the last 8 basketball matches are shown below.

Points by Andrew	23	17	31	25	25	19	28	32
Points by Brad	9	29	41	26	14	44	38	43

a Find the mean and standard deviation of the number of points scored by each player.

b Which of the two players is more consistent?

3 Two baseball coaches compare the number of runs scored by their teams in their last ten matches:

Rockets	0	10	1	9	11	0	8	5	6	7
Bullets	4	3	4	1	4	11	7	6	12	5

a Show that each team has the same mean and range of runs scored.

b Which team's performance do you suspect is more variable over the period?

c Check your answer to **b** by finding the standard deviation for each distribution.

d Does the range or the standard deviation give a better indication of variability?

4 A manufacturer of soft drinks employs a statistican for quality control. He needs to check that 375 mL of drink goes into each can, but realises the machine which fills the cans will slightly vary each delivery.

a Would you expect the standard deviation for the whole production run to be the same for one day as it is for one week? Explain your answer.

b If samples of 125 cans are taken each day, what measure would be used to:

i check that an average of 375 mL of drink goes into each can

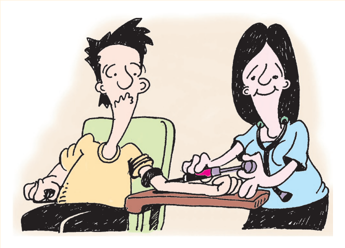
ii check the variability of the volume of drink going into each can?

c What is the significance of a low standard deviation in this case?

INVESTIGATION 5 HEART STOPPERS

A new drug is claimed to lower the cholesterol level in humans. To test this claim, a heart specialist enlisted the help of 50 of his patients.

The patients agreed to take part in an experiment in which 25 of them would be randomly allocated to take the new drug and the other 25 would take an identical looking pill that was actually a *placebo* with no effect.



All participants had their cholesterol level measured before starting the course of pills, with the following results:

7.1 8.2 8.4 6.5 6.5 7.1 7.2 7.1 6.1 6.0 8.5 5.0 6.3 6.7 7.3 8.9 6.2
6.3 7.1 8.4 7.4 7.6 7.5 6.6 8.1 6.2 6.2 7.0 8.1 8.4 6.4 7.6 8.6 7.5
7.9 6.2 6.8 7.5 6.0 5.0 8.3 7.9 6.7 7.3 6.0 7.4 7.4 8.6 6.5 7.6

Two months later, the cholesterol levels of the participants were again measured, but this time they were divided into two groups.

The cholesterol levels of the 25 participants who took the drug were:

4.8 5.6 4.7 4.2 4.8 4.6 4.8 5.2 4.8 5.0 4.7 5.1 4.7
4.4 4.7 4.9 6.2 4.7 4.7 4.4 5.6 3.2 4.4 4.6 5.2

The cholesterol levels of the 25 participants who took the placebo were:

7.0 8.4 8.8 6.1 6.6 7.6 6.5 7.9 6.2 6.8 7.5 6.0 8.2
5.7 8.3 7.9 6.7 7.3 6.1 7.4 8.4 6.6 6.5 7.6 6.1

What to do:

1 Use the data to complete the table:

Cholesterol level	*Before the experiment*	25 *participants taking the drug*	25 *participants taking the placebo*
$4.0 \leqslant l < 4.5$			
$4.5 \leqslant l < 5.0$			
$5.0 \leqslant l < 5.5$			
$5.5 \leqslant l < 6.0$			
$\vdots$			
$8.5 \leqslant l < 9.0$			

2 Produce histograms showing the cholesterol levels of the three groups in the table.

STATISTICS PACKAGE

3 Calculate the mean and standard deviation for each group in the table.

4 Write a report presenting your findings.

PROJECT IDEAS

You should now have enough knowledge to be able to conduct your own statistical investigation.

1 Choose a problem or issue that you find interesting. Find a question that you can investigate, making sure that you can find useful data for it. Some ideas to get you started can be found by clicking on the icon alongside.

PROJECT IDEAS

2 Think about how you will organise and display your data when you have collected it.

3 Discuss your question and plans for analysis with your teacher, and make changes to the problem or your research plan if necessary.

4 Collect your data, making sure that it is randomly selected, and that you have enough to make a fair conclusion. Use technology to produce appropriate graphs or statistical calculations. In your analysis, you may need to consider:

- Is the data categorical, quantitative discrete, or quantitative continuous?
- Do you need to group any of the data?
- Are there any outliers? If so, are they legitimate data?
- Should you find measures for the centre or spread? If so, which ones should you use?

5 Write a report of your investigation as a newspaper article, a slideshow presentation, or a word processed document. Your report should include:

- an explanation of the problem you researched
- a simple description of your method of investigation
- the analysis you carried out including raw data and any summary statistics, graphs, or tables that you produced
- your conclusion, with the reasons you came to that decision
- a discussion of any flaws in your method that might weaken your conclusion.

REVIEW SET 6A

1 Classify the following data as categorical, quantitative discrete, or quantitative continuous:

a the number of pages in a daily newspaper

b the maximum daily temperature in the city

c the manufacturer of a television

d the preferred football code

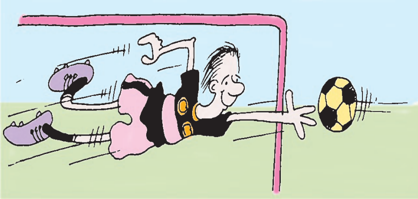
e the position taken by a player on a lacrosse field

f the time it takes to run one kilometre

g the length of people's feet

h the number of goals shot by a soccer player

i the cost of a bicycle.



2 The data below are the lengths, in metres, of yachts competing in a sailing race.

14.7	14.1	21.6	16.2	15.7	12.8	10.1	13.9	14.4	13.0
11.7	14.6	17.2	13.4	12.1	11.3	13.1	21.6	23.5	16.4
14.4	15.8	12.6	19.7	18.0	16.2	27.4	21.9	14.4	12.4

a Produce a frequency histogram of the data.

b Find the **i** median **ii** range of the yacht lengths.

c Comment on the skewness of the data.

3 Find a given that the data set $2, a, 5, 4, 1, 2, 3, 5$ has a mean of 3.

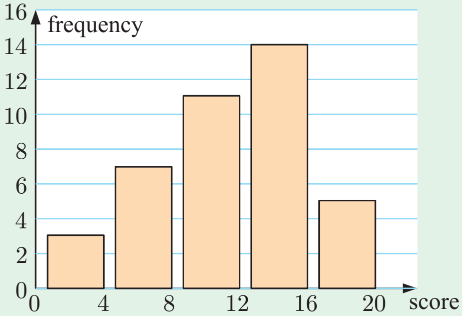
4 The column graph shows the marks out of 20 that were scored for a test.

a Describe the distribution of the data.

b What percentage of the students scored 13 or more marks?

c What percentage of the students scored less than 5 marks?

d Explain why we cannot display the data in this graph in a box and whisker plot.



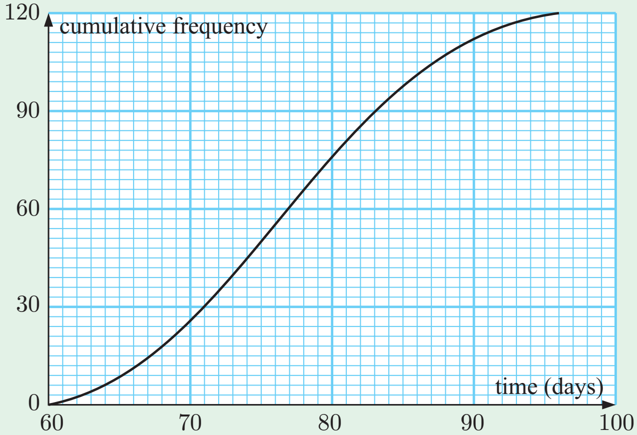
5 Draw a box and whisker plot for the data: 11, 12, 12, 13, 14, 14, 15, 15, 15, 16, 17, 17, 18.

6 120 people caught whooping cough in an outbreak. The times for them to recover were recorded and the results were used to produce the cumulative frequency graph shown.

Estimate:

a the median

b the interquartile range.



7 Find, using your calculator, the mean and standard deviation of these sets of data:

a 117, 129, 105, 124, 123, 128, 131, 124, 123, 125, 108

b 6.1, 5.6, 7.2, 8.3, 6.6, 8.4, 7.7, 6.2

8 Consider this set of data:

19, 7, 22, 15, 14, 10, 8, 28, 14, 18, 31, 13, 18, 19, 11, 3, 15, 16, 19, 14

a Find the 5-number summary for the data. **b** Find the range and IQR of the data.

c Draw a boxplot of the data set.

REVIEW SET 6B

1 A sample of lamp-posts was surveyed for the following data. Classify the data as categorical, quantitative discrete, or quantitative continuous:

a the diameter of the lamp-post measured 1 metre from its base

b the material from which the lamp-post is made

c the location of the lamp-post (inner, outer, North, South, East, or West)

d the height of the lamp-post

e the time since the last inspection

f the number of inspections since installation

g the condition of the lamp-post (very good, good, fair, unsatisfactory).

2 The data below are the distances in metres that Thabiso threw a baseball:

71.2	65.1	68.0	71.1	74.6	68.8	83.2	85.0	74.5	87.4
84.3	77.0	82.8	84.4	80.6	75.9	89.7	83.2	97.5	82.9
90.5	85.5	90.7	92.9	95.6	85.5	64.6	73.9	80.0	86.5

a Determine the highest and lowest value for the data set.

b Determine: **i** the mean **ii** the median.

c Choose between 6 and 12 groups into which all the data values can be placed.

d Prepare a frequency distribution table.

e Draw a frequency histogram for the data.

3 Consider the following distribution of continuous grouped data:

Scores (x)	$0 \leqslant x < 10$	$10 \leqslant x < 20$	$20 \leqslant x < 30$	$30 \leqslant x < 40$	$40 \leqslant x < 50$
Frequency	1	13	27	17	2

a Construct a cumulative frequency graph for the data.

b Estimate the:

i median **ii** interquartile range **iii** mean **iv** standard deviation.

4 The daily profits of a shop over the last 20 days, in pounds, are:

324	348	352	366	346	329	375	353	336	368
336	375	356	358	353	311	365	376	343	331

a Find the: **i** median **ii** lower quartile **iii** upper quartile.

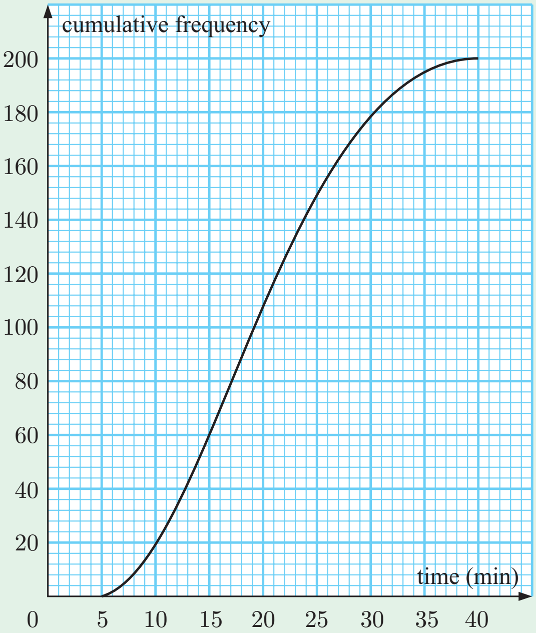
b Find the interquartile range of the data set.

c Find the mean and standard deviation of the daily profits.

5 This cumulative frequency curve shows the times taken for 200 students to travel to school by bus.

a Estimate how many of the students spent between 10 and 20 minutes travelling to school.

b 30% of the students spent more than m minutes travelling to school. Estimate the value of m.



6 The playing time, in minutes, of CDs in a shop is shown alongside.

a Estimate the mean and standard deviation of the playing time.

b Draw a histogram to present this data.

c Comment on the shape of the distribution.

Playing time (minutes)	*Number of CDs*
$30 \leqslant t < 35$	5
$35 \leqslant t < 40$	13
$40 \leqslant t < 45$	17
$45 \leqslant t < 50$	29
$50 \leqslant t < 55$	27
$55 \leqslant t < 60$	18
$60 \leqslant t < 65$	7

7 Find the range, lower quartile, upper quartile, and standard deviation for the following data:

120, 118, 132, 127, 135, 116, 122, 128.

8 A confectioner claims to sell an average of 30 liquorice allsorts per bag. The results from a survey of bags are shown in the table below.

Number of allsorts	27	28	29	30	31	32
Frequency	23	29	41	37	22	32

a Find the mean and standard deviation for this data.

b Is the confectioner's claim justified?

REVIEW SET 6C

1 A set of 14 data is: 6, 8, 7, 7, 5, 7, 6, 8, 6, 9, 6, 7, p, q.
The mean and mode of the set are both 7.
Find p and q.

2 The winning margins in 100 rugby games were recorded as follows:

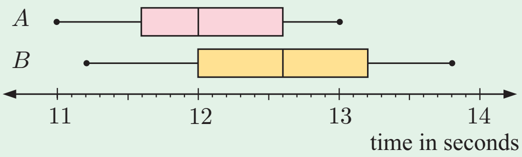
Margin (points)	1 - 10	11 - 20	21 - 30	31 - 40	41 - 50
Frequency	13	35	27	18	7

Draw a column graph to present this information.

3 The table alongside shows the number of patrons visiting an art gallery on various days.
Estimate the mean number of patrons per day.

Number of patrons	*Frequency*
250 - 299	14
300 - 349	34
350 - 399	68
400 - 449	72
450 - 499	54
500 - 549	23
550 - 599	7

4 The parallel boxplots show the 100 metre sprint times for the members of two athletics squads.



a Determine the 5-number summaries for both A and B.

b Determine: **i** the range **ii** the interquartile range for each group.

c Copy and complete:

i We know the members of squad ...... generally ran faster because ......

ii We know the times in squad ...... are more varied because ......

5 The supermarket bills for a number of families was recorded in the table given.
Estimate the mean bill and the standard deviation of the bills.

Bill (€)	*Frequency*
70 - 79.99	27
80 - 89.99	32
90 - 99.99	48
100 - 109.99	25
110 - 119.99	37
120 - 129.99	21
130 - 139.99	18
140 - 149.99	7

6 An examination worth 100 marks was given to 800 biology students. The cumulative frequency graph for the students' results is shown on the following page.

a Find the number of students who scored 45 marks or less for the test.

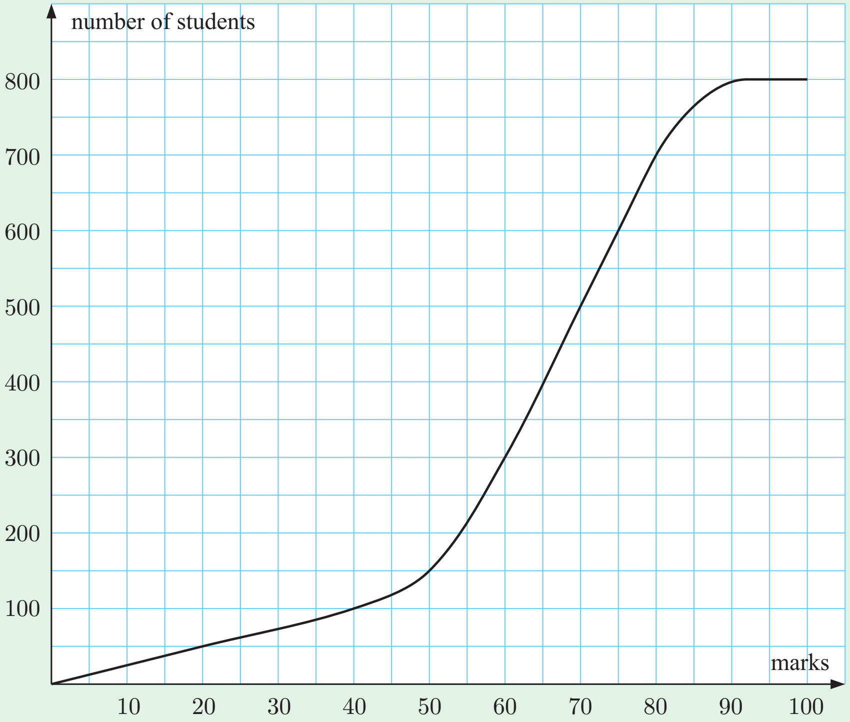
b Find the median score.

c Between what values do the middle 50% of test results lie?

d Find the interquartile range of the data.

e What percentage of students obtained a mark of 55 or more?

f If a 'distinction' is awarded to the top 10% of students, what score is required to receive this honour?



7 The number of peanuts in a jar varies slightly from jar to jar. Samples of 30 jars were taken for each of two brands X and Y, and the number of peanuts in each jar was recorded.

Brand X

871	885	878	882	889	885
916	913	886	905	907	898
874	904	901	894	897	899
908	901	898	894	895	895
910	904	896	893	903	888

Brand Y

909	906	913	891	898	901
894	894	928	893	924	892
927	907	901	900	907	913
921	904	903	896	901	895
917	903	910	903	909	904

a Copy and complete this table:

	Brand X	*Brand Y*
min		
Q_1		
median		
Q_3		
max		
IQR		

b Display the data on a parallel boxplot.

c Comment on which brand:

- **i** has more peanuts per jar
- **ii** has a more consistent number of peanuts per jar.

6 **a** 5, 8, 11, 14, 17, ... **b** 5, 8, 11, 14, 17, ...

c 100, 93, 86, 79, 72, ... **d** 100, 93, 86, 79, 72, ...

e 5, 10, 20, 40, 80, ... **f** 5, 10, 20, 40, 80, ...

g 48, 24, 12, 6, 3, ... **h** 48, 24, 12, 6, 3, ...

Each pair describes the same sequence. (**a** and **b**, **c** and **d**, ...)

EXERCISE 5C

1 **a** 73 **b** 65 **c** 21.5

2 **a** 101 **b** -107 **c** $a + 14d$

3 **b** $u_n = 11n - 5$ **c** 545 **d** Yes **e** No

4 **b** $u_n = 91 - 4n$ **c** -69 **d** the 97th term

5 **b** $u_1 = 1,\ \ d = 3$ **c** 169 **d** $u_{150} = 448$

6 **b** $u_1 = 32,\ \ d = -\frac{7}{2}$ **c** -227 **d** $n \geqslant 68$

7 **a** $k = 17\frac{1}{2}$ **b** $k = 4$ **c** $k = 4$ **d** $k = 0$

e $k = 7$ **f** $k = -4$

8 **a** $k = 3$ or -2 **b** $k = 1$ or -3 **c** $k = 3$ or -1

9 **a** $u_n = 6n - 1$ **b** $u_n = -\frac{3}{2}n + \frac{11}{2}$

c $u_n = -5n + 36$ **d** $u_n = -\frac{3}{2}n + \frac{1}{2}$

10 **a** 6.25, 7.5, 8.75 **b** $3\frac{5}{7}, 8\frac{3}{7}, 13\frac{1}{7}, 17\frac{6}{7}, 22\frac{4}{7}, 27\frac{2}{7}$

11 **a** $u_1 = 36,\ \ d = -\frac{2}{3}$ **b** u_{100} **12** $u_{7692} = 100\,006$

13 **a** Month 1 = 5 cars Month 4 = 44 cars
Month 2 = 18 cars Month 5 = 57 cars
Month 3 = 31 cars Month 6 = 70 cars

b The constant difference $d = 13$. **c** 148 cars

d 20 months

14 **b** 111 online friends **c** 18 weeks

15 **a** Day 1 = 97.3 tonnes, Day 2 = 94.6 tonnes
Day 3 = 91.9 tonnes

b $d = -2.7$, the cattle eat 2.7 tonnes of hay each day.

c $u_{25} = 32.5$. After 25 days (i.e., July 25th) there will be 32.5 tonnes of hay left.

d 16.3 tonnes

EXERCISE 5D.1

1 **a** $b = 18,\ \ c = 54$ **b** $b = \frac{5}{2},\ \ c = \frac{5}{4}$ **c** $b = 3,\ \ c = -\frac{3}{2}$

2 **a** 96 **b** 6250 **c** 16

3 **a** 6561 **b** $\frac{19\,683}{64}$ $(307\frac{35}{64})$ **c** 16 **d** ar^8

4 **a** $u_1 = 5,\ \ r = 2$ **b** $u_n = 5 \times 2^{n-1}$ **c** $u_{15} = 81\,920$

5 **a** $u_1 = 12,\ \ r = -\frac{1}{2}$ **b** $u_n = 12 \times (-\frac{1}{2})^{n-1}$

c $u_{13} = \frac{3}{1024}$

6 $u_{10} \approx -0.601$ **7** $u_n = 8\left(\frac{1}{\sqrt{2}}\right)^{n-1}$

8 **a** $k = \pm 14$ **b** $k = 2$ **c** $k = -2$ or 4

9 **a** $u_n = 3 \times 2^{n-1}$ **b** $u_n = 32 \times (-\frac{1}{2})^{n-1}$

c $u_n = 3 \times (\sqrt{2})^{n-1}$ or $u_n = 3 \times (-\sqrt{2})^{n-1}$

d $u_n = 10 \times \left(\frac{1}{\sqrt{2}}\right)^{n-1}$ or $u_n = 10\left(-\frac{1}{\sqrt{2}}\right)^{n-1}$

10 **a** $u_9 = 13\,122$ **b** $u_{14} = 2916\sqrt{3}$ **c** $u_{18} \approx 0.000\,091\,6$

EXERCISE 5D.2

1 **a** **i** ≈ 1550 ants **ii** ≈ 4820 ants **b** ≈ 12.2 weeks

2 **a** **i** ≈ 73 **ii** ≈ 167 **b** 30.5 years

3 **a** ≈ 220 members **b** ≈ 11.2 years

4 **a** **i** ≈ 2860 **ii** $\approx 184\,000$ **b** 14.5 years

5 **a** ≈ 319 **b** ≈ 52 years

EXERCISE 5E.1

1 **a** 91 **b** 91 **c** 91 **2** 203 **3** $-115\frac{1}{2}$

4 **a** 160 **b** 820 **c** $3087\frac{1}{2}$ **d** -1460

e -150 **f** -740

5 **a** $d = 6$ **b** $n = 12$ **c** $S_{12} = 504$

6 **a** 1749 **b** 2115 **c** $1410\frac{1}{2}$

7 **a** 65 **b** 1914 **c** 47 850

8 **a** 14 025 **b** 71 071 **c** 3367

9 **a** **i** $u_{10} = 38$ **ii** $u_{30} = 78$ **b** 1470

10 8 terms **11** **a** $d = 3$ **b** $n = 11$

12 15 terms **13** 18 layers **14** -2, 4, 10 or 10, 4, -2

EXERCISE 5E.2

1 **a** 93 **b** 93

2 **a** 6560 **b** 5115 **c** $\frac{3069}{128}$ **d** $\approx 189\,134$

e ≈ 4.00 **f** ≈ 0.585

3 **a** $S_n = \dfrac{\sqrt{3}\left((\sqrt{3})^n - 1\right)}{\sqrt{3} - 1}$ **b** $S_n = 24\left(1 - (\frac{1}{2})^n\right)$

c $S_n = 1 - (0.1)^n$ **d** $S_n = \frac{40}{3}\left(1 - (-\frac{1}{2})^n\right)$

4 **c** \$26 361.59 **5** $n = 5$

6 **a** $u_8 = 1.25$ **b** $S_8 = 318.75$ **c** 12 terms

EXERCISE 5F.1

1 \$6945.75 **2** **a** £17 496 **b** £2496

3 **a** ¥1 295 718.36 **b** ¥407 718.36 **4** £13 373.53

5 \$546 **6** Bank A

EXERCISE 5F.2

2 £6629.65 **3** \$4079.77 **4** €4159.08 **5** ¥199 713.08

6 \$20 836.86 **7** €80 000 **8** 2 years 5 months

9 2 years 9 months **10** 13 years 3 months **11** 14.5% p.a.

12 6.00% p.a. **13** 5.25% p.a.

EXERCISE 5G

1 €1280 **2** **a** €26 103.52 **b** €83 896.48

3 **a** ¥30 012.5 **b** ¥57 487.5 **4** 24.8% **5** 18.4%

REVIEW SET 5A

1 **a** arithmetic **b** arithmetic and geometric

c geometric **d** neither **e** arithmetic

2 $k = -\frac{11}{2}$ **3** $u_n = 33 - 5n,\ \ S_n = \dfrac{n}{2}(61 - 5n)$

4 $k = 4$ or -4 **5** $u_n = \frac{1}{6} \times 2^{n-1}$ or $u_n = -\frac{1}{6} \times (-2)^{n-1}$

6 21, 19, 17, 15, 13, 11

7 **a** $u_8 = \frac{1}{15\,625}$ **b** $u_8 = 6\frac{1}{2}$ **c** $u_8 = a - 7d$

8 **a** Week 1: 2817 L Week 3: 2451 L
Week 2: 2634 L Week 4: 2268 L

b Amount in tank decreases by the same amount (183 L) each week.

c after 17 weeks

9 **a** -492 **b** 7 324 218

10 **a** $u_8 = 61$ **b** $S_{10} = 435$ **c** $n = 15$

11 Bank A: \$231 995.25, Bank B: \$220 787.17
Val should deposit her money in Bank A.

12 \$657.26 **13** **a** \$59 900.22 **b** \$75 099.78

14 in 3 years

REVIEW SET 5B

1 **b** $u_1 = 6, \ \ r = \frac{1}{2}$ **c** $0.000\,183$

2 **a** 81 **b** $u_{35} = -1\frac{1}{2}$ **c** -486 **3** **a** 1587 **b** $47\frac{253}{256}$

4 **a** $\frac{1}{3}, \frac{1}{9}, \frac{1}{27}, \frac{1}{81}, \frac{1}{243}$ **b** $17, 22, 27, 32, 37$
c $\frac{4}{3}, 1, \frac{4}{5}, \frac{2}{3}, \frac{4}{7}$

5 **a** €8415.31 **b** €2415.31 **6** $u_{11} \approx 0.000\,406$

7 **a** $u_n = (\frac{3}{4})2^{n-1}$ **b** $49\,152$ **c** $24\,575\frac{1}{4}$

8 **a** 33 **b** $4(n+1) - 7 - (4n - 7) = 4$
c The difference between terms is always the same. **d** 1328

9 **a** 17 terms **b** $255\frac{511}{512}$

10 **a** $r = \frac{1}{3}$ **b** $u_6 = \frac{20}{27}$ **c** $n = 8$ **11** €970.26

12 4.80% p.a. **13** **a** 12.5% per year **b** $10 966.45

REVIEW SET 5C

1 **b** $u_1 = 63, \ \ d = -5$ **c** $u_{37} = -117$ **d** $u_{54} = -202$

2 **b** $u_n = 3 \times 4^{n-1}, \ \ u_9 = 196\,608$

3 $u_n = 73 - 6n, \ \ u_{34} = -131$

4 **a** $a = 15$ **b** $a = 12$ or -12

5 **a** $u_{10} = -31$ **b** $u_{10} = 243$

6 **a** ≈ 3470 **b** 11 years

7 **a** $u_n = 89 - 3n$ **b** $u_n = \dfrac{2n+1}{n+3}$
c $u_n = 100 \times (0.9)^{n-1}$

8 $u_{12} = 10\,240$ **9** $k = -2$ or 4

10 **a** 2001: 630 000, 2002: 567 000 **b** $\approx 4\,560\,000$ sheets

11 €9838.99 **12** £13 125.36 **13** $\approx -13.3\%$

EXERCISE 6A

1 **a** quantitative discrete **b** categorical
c quantitative continuous **d** quantitative continuous
e categorical **f** quantitative discrete **g** categorical
h quantitative discrete **i** quantitative continuous
j quantitative continuous **k** quantitative continuous
l categorical **m** quantitative discrete

2 **a** 0, 1, 2, 3, ..., 8 **b** red, yellow, orange, green, ...
c 0 - 15 minutes **d** 0 - 25 m
e Ford, BMW, Renault, ... **f** 1, 2, 3, ..., 20
g Australia, Hawaii, Dubai, ... **h** 0.0 - 10.0
i 0 - 4 L **j** 0 - 80 hours **k** −20°C - 35°C
l cereal, toast, fruit, rice, eggs, ... **m** 0, 1, 2, ..., 10

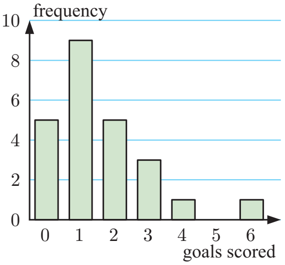
EXERCISE 6B

1 **a** the number of goals scored in a game
b variable is counted, not measured

c

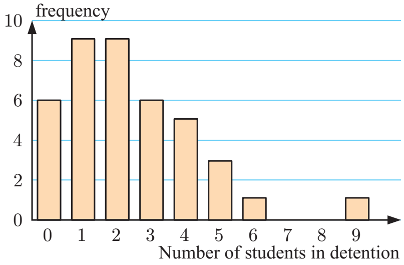
Goals scored	*Tally*	*Frequency*	*Rel. Frequency*
0	𝍸	5	0.208
1	𝍸 ||||	9	0.375
2	𝍸	5	0.208
3	|||	3	0.125
4	|	1	0.042
5		0	0
6	|	1	0.042
	Total	24	

d



e 1 goal
f positively skewed, one outlier, (6 goals)
g $\approx 20.8\%$

2 **a**

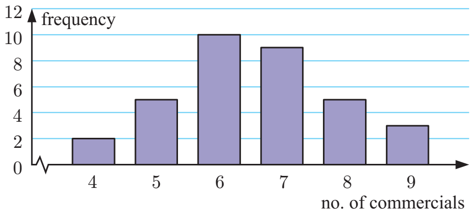


b 1 and 2 **c** positively skewed, one outlier, (9 detentions)
d $12\frac{1}{2}\%$

3 **a**

No. of commercials	*Tally*	*Frequency*
4	||	2
5	𝍸	5
6	𝍸 𝍸	10
7	𝍸 ||||	9
8	𝍸	5
9	|||	3
	Total	34

b



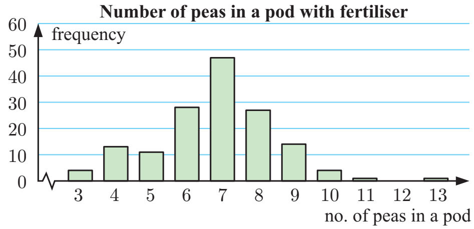
c 6 commercials **d** symmetrical, no outliers **e** $\approx 79.4\%$

4 **a** 45 **b** 1 time **c** 8 **d** 20%
e positively skewed, no outliers

5 **a**

Peas in pod	*Tally*	*Freq.*
3	||||	4
4	𝍸 𝍸 |||	13
5	𝍸 𝍸 |	11
6	𝍸 𝍸 𝍸 𝍸 𝍸 |||	28
7	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 ||	47
8	𝍸 𝍸 𝍸 𝍸 𝍸 ||	27
9	𝍸 𝍸 ||||	14
10	||||	4
11	|	1
12		0
13	|	1
	Total	150

b Number of peas in a pod with fertiliser



c Symmetrical, one outlier (13 peas) **d** Yes

e Not necessarily. (Consider factors like the cost of the fertiliser, changing prices, etc.)

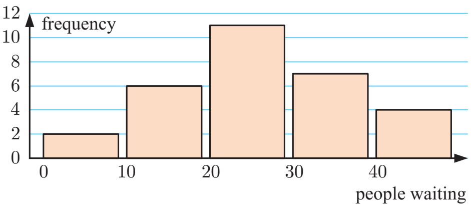
EXERCISE 6C

1 **a**

People waiting	*Tally*	*Frequency*	*Rel. Freq.*
0 - 9	||	2	0.067
10 - 19	卌 |	6	0.200
20 - 29	卌 卌 |	11	0.367
30 - 39	卌 ||	7	0.233
40 - 49	||||	4	0.133
	Total	30	

b 2 days **c** ≈ 36.7% **e** 20 - 29 people

d



2 **a** 37 **b** 40 - 49 employees **c** negatively skewed

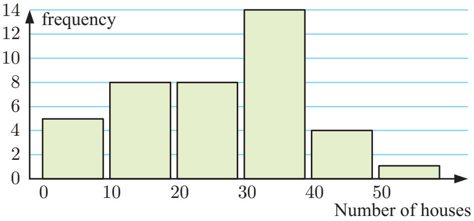
d ≈ 37.8%

e No, only that it was in the interval 50 - 59 employees.

3 **a**

Number of houses	*Tally*	*Frequency*
0 - 9	卌	5
10 - 19	卌 |||	8
20 - 29	卌 |||	8
30 - 39	卌 卌 ||||	14
40 - 49	||||	4
50 - 59	|	1
	Total	40

b

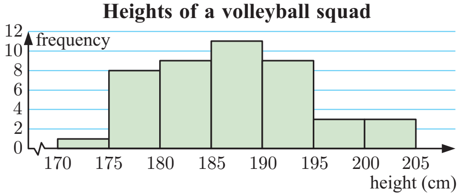


c 30 - 39 houses **d** 67.5%

EXERCISE 6D

1 **a** Height is measured on a continuous scale.

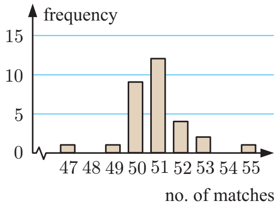
b Heights of a volleyball squad

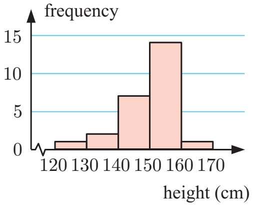


c $185 \leqslant H < 190$ cm. This is the class of values that appears most often.

d slightly positively skewed

2 **a** column graph **b** frequency histogram



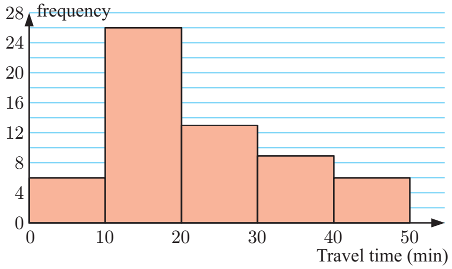


3 **a** continuous

b

Travel time (min)	*Tally*	*Frequency*
$0 \leqslant t < 10$	卌 |	6
$10 \leqslant t < 20$	卌 卌 卌 卌 卌 |	26
$20 \leqslant t < 30$	卌 卌 |||	13
$30 \leqslant t < 40$	卌 ||||	9
$40 \leqslant t < 50$	卌 |	6
	Total	60

c Travel times to school

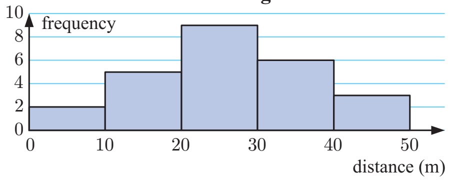


d positively skewed **e** $10 \leqslant t < 20$ minutes

4 **a**, **b**

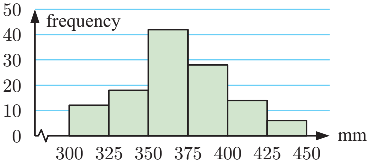
Distance (m)	*Tally*	*Frequency*
$0 \leqslant d < 10$	||	2
$10 \leqslant d < 20$	卌	5
$20 \leqslant d < 30$	卌 ||||	9
$30 \leqslant d < 40$	卌 |	6
$40 \leqslant d < 50$	|||	3
	Total	25

c Javelin throwing distances



d $20 \leqslant d < 30$ m **e** 36%

5 **a** Heights of 6-month old seedlings at a nursery

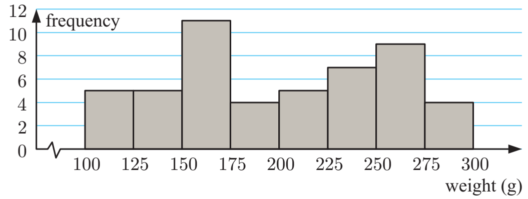


b 20 **c** ≈ 58.3% **d** **i** 1218 **ii** 512

6 **a, b**

Weight (g)	*Tally*	*Frequency*
$100 \leqslant w < 125$	卌	5
$125 \leqslant w < 150$	卌	5
$150 \leqslant w < 175$	卌 卌 |	11
$175 \leqslant w < 200$	||||	4
$200 \leqslant w < 225$	卌	5
$225 \leqslant w < 250$	卌 ||	7
$250 \leqslant w < 275$	卌 ||||	9
$275 \leqslant w < 300$	||||	4
	Total	50

c **Weights of laboratory rats**



d 50%

EXERCISE 6E.1

1 **a** 1 cup **b** 2 cups **c** 1.8 cups **2** 9

3 **a** **i** 5.61 **ii** 6 **iii** 6 **b** **i** 16.3 **ii** 17 **iii** 18
c **i** 24.8 **ii** 24.9 **iii** 23.5

4 **a** data set A: 6.46, data set B: 6.85
b data set A: 7, data set B: 7
c The data are the same except for the last value, which pushes the mean of set B up.
d 7 is the middle value in both data sets. It is not affected by extreme values.

5 Ruth (164)

6 **a** **i** Pies: 67.1, Pasties: 53.6
ii Pies: 69, Pasties: 52
b Pies, higher mean (more sold), higher median (higher data values)

7 **a** Bus: mean = 39.7, median = 40.5,
Tram: mean ≈ 49.1, median = 49
b Tram has higher mean and median, but there are more bus trips per day and more people travel by bus in a day, so bus is more popular.

8 **a** 44 points **b** 44 points **c** 40.2 points
d increase, 40.3 points

9 \$185 604 **10** 3144 km **11** 17.25 goals **12** $x = 15$
13 $a = 5$ **14** 37 **15** 14.8 **16** 6, 12 **17** 7, 9

EXERCISE 6E.2

1 **a** Mean: \$163 770, median: \$147 200
Mean has been affected by the extreme values (the two values greater than \$200k).
b **i** the mean **ii** the median

2 **a** mean: \$29 300, median: \$23 500, mode: \$23 000
b It is the lowest value in the data set.
c No, it is too close to the lower end of the distribution.

3 **a** mean: 3.19 mm, median: 0 mm, mode: 0 mm
b The median is not in the centre as the data is positively skewed.
c The mode is the lowest value.
d Yes, 42 and 21. **e** No

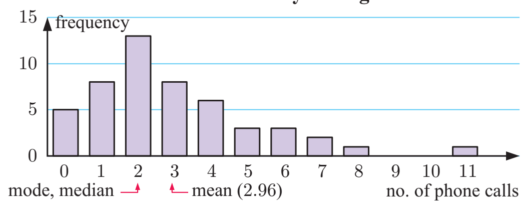
EXERCISE 6E.3

1 **a** 1 head **b** 1 head **c** 1.43 heads

2 **a** **i** 2.61 children **ii** 2 children **iii** 2 children
b This school has more children per family than average.
c positive **d** mean is higher than the median, mode

3 **a** **i** 2.96 calls **ii** 2 calls **iii** 2 calls

b **Phone calls made by teenagers**



c positively skewed **d** Because of the skewness.
e mean

4 **a** **i** 49 matches **ii** 49 matches **iii** 49.0 matches
b No **c** Need a larger sample.

5 **a** **i** 5.63 peas **ii** 6 peas **iii** 6 peas
b **i** 6.81 peas **ii** 7 peas **iii** 7 peas
c all of them **d** It has improved it.

EXERCISE 6E.4

1 31.7 **2** **a** 70 **b** ≈ 411 000 L **c** ≈ 5870 L

3 **a** 11.5 points **b** **i** 11.3 points **ii** 11.4 points
c **ii** is closer to the actual mean than **i**. Smaller class intervals give better estimates.

4 90.1 km h^{-1} **5** 768 m^2

6 **a** 125 people **b** 119 marks **c** $\frac{3}{25}$ **d** 137

EXERCISE 6F

1 **a** **i** 6 **ii** $Q_1 = 4$, $Q_3 = 7$ **iii** 7 **iv** 3
b **i** 17.5 **ii** $Q_1 = 15$, $Q_3 = 19$ **iii** 14 **iv** 4
c **i** 24.9 **ii** $Q_1 = 23.5$, $Q_3 = 26.1$ **iii** 7.7 **iv** 2.6

2 **a** median = 2.45 min, Q_1 = 1.45 min, Q_3 = 3.8 min
b range = 5.2 minutes, IQR = 2.35 minutes
c **i** 2.45 min **ii** 3.8 min **iii** 0, 5.2, 5.2

3 **a** 6 **b** 28 **c** 15 **d** 12 **e** 21 **f** 22 **g** 9

4 **a** **i** 124 cm **ii** Q_1 = 116 cm, Q_3 = 130 cm
b **i** 124 cm **ii** 130 cm **c** **i** 29 cm **ii** 14 cm
d 14 cm

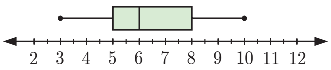
5 **a** **i** 7 peas **ii** 6 peas **iii** 5 peas **iv** 7 peas **v** 2 peas
b **i** 10 peas **ii** 7 peas **iii** 6 peas **iv** 8 peas
v 2 peas
c The fertiliser does improve the yield of peas.

EXERCISE 6G.1

1 **a** **i** 35 points **ii** 78 points **iii** 13 points
iv 53 points **v** 26 points
b **i** 65 points **ii** 27 points

2 **a** **i** 98, 25 marks **ii** 70 marks **iii** 85 marks
iv 55, 85 marks
b 73 marks **c** 30 marks **d** 67 marks

3 **a** **i** min = 3; Q_1 = 5; med = 6; Q_3 = 8; max = 10 ii

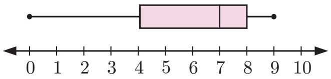


iii 7

iv 3

b i min = 0, $Q_1 = 4$; med = 7; $Q_3 = 8$, max = 9

ii

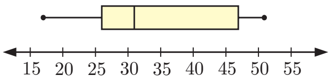


iii 9

iv 4

c i min = 17, $Q_1 = 26$; med = 31; $Q_3 = 47$, max = 51

ii

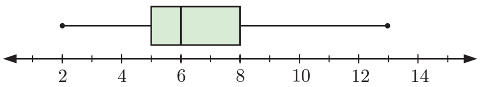


iii 34

iv 21

4 **a** median = 6, $Q_1 = 5$, $Q_3 = 8$ **b** 3

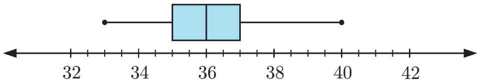
c



5 **a** min = 33, $Q_1 = 35$, med = 36, $Q_3 = 37$, max = 40

b i 7 ii 2

c



d No

EXERCISE 6G.2

1 **a**

Statistic	*Year 9*	*Year 12*
minimum	1	6
Q_1	5	10
median	7.5	14
Q_3	10	16
maximum	12	17.5

b i Year 9: 11, Year 12: 11.5

ii Year 9: 5, Year 12: 6

c i cannot tell ii true since Year 9 $Q_1 <$ Year 12 min.

2 **a** Friday: min = \$20, Q_1 = \$50, med = \$70, Q_3 = \$100, max = \$180

Saturday: min = \$40, Q_1 = \$80, med = \$100, Q_3 = \$140, max = \$200

b i Friday: \$160, Saturday: \$160

ii Friday: \$50, Saturday: \$60

3 **a** i Class 1 (96%) ii Class 1 (37%) iii Class 1

b 18 **c** 55 **d** i 25% ii 50%

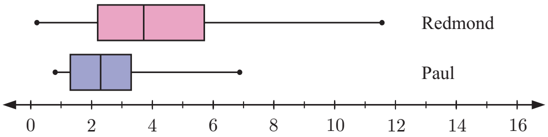
e i slightly positively skewed ii negatively skewed

f .... class 2, ... class 1

4 **a** Paul: min = 0.8; $Q_1 = 1.3$; med = 2.3; $Q_3 = 3.3$; max = 6.9

Redmond: min = 0.2; $Q_1 = 2.2$; med = 3.7; $Q_3 = 5.7$; max = 11.5

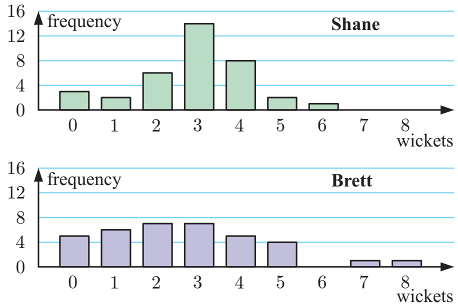
b **Mobile Phone call duration**



c Both are positively skewed (Redmond's more so than Paul's). Redmond's phone calls were more varied in duration.

5 **a** discrete

c



d Shane: approximately symmetrical Brett: positively skewed.

e Shane: mean ≈ 2.89, median = 3, mode = 3

Brett: mean ≈ 2.67, median = 2.5, mode = 2, 3

Shane's mean and median are slightly higher.

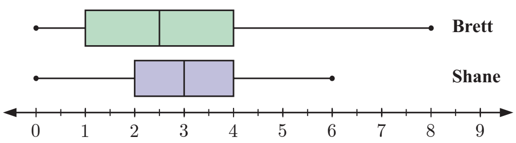
Shane has a clear mode of 3, whereas Brett has two modes (2 and 3)

f Shane: Range = 6, IQR = 2

Brett: Range = 8, IQR = 3

Shane's data set demonstrates less variability than Brett's.

g



h Shane is more consistent with his bowling (in terms of wickets taken) than Brett.

6 **a** continuous (the data is measured)

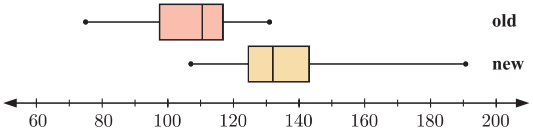
c Old: mean = 107, median = 110.5, range = 56, IQR = 19, min = 75, max = 131

New: mean = 134, median = 132, range = 84, IQR = 18.5, min = 107, max = 191

The 'new' type of light globe has a higher mean and median than the 'old' type.

The IQR is relatively unchanged going from 'old' to 'new', however, the range of the 'new' type is greater, suggesting greater variability.

d



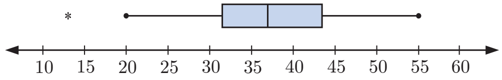
e Old type: negatively skewed, New type: positively skewed

f The 'new' type of light globes do last longer than the old type. Each number in the 5-number summary is at least 20% greater in the 'new' type. The manufacturer's claim appears to be valid.

EXERCISE 6G.3

1 **a** 12 **b** lower: 13.5, upper: 61.5 **c** 13

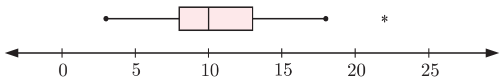
d



2 **a** median = 10, $Q_1 = 8$, $Q_3 = 13$ **b** 5

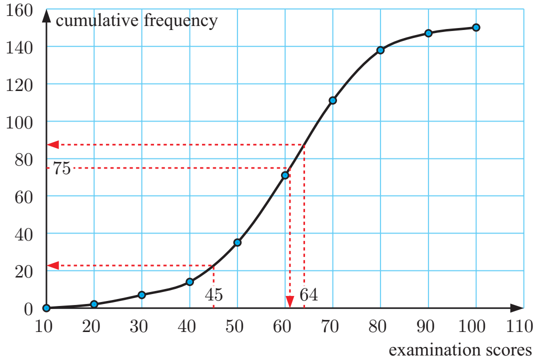
c lower = 0.5, upper = 20.5 **d** Yes, 22

e



EXERCISE 6H

1



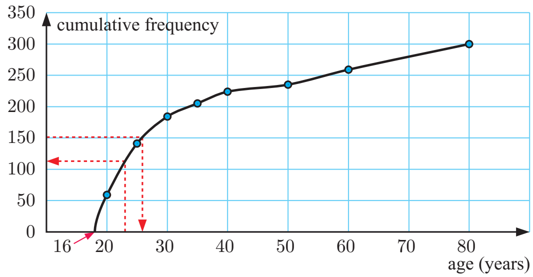
a ≈ 61 marks **b** ≈ 87 students **c** ≈ 76 students

d ≈ 23 students **e** 79 marks

2 **a** 9 **b** ≈ 28.3% **c** 7.1 cm **d** ≈ 2.4 cm

e 90% of the seedlings are shorter than 10 cm.

3

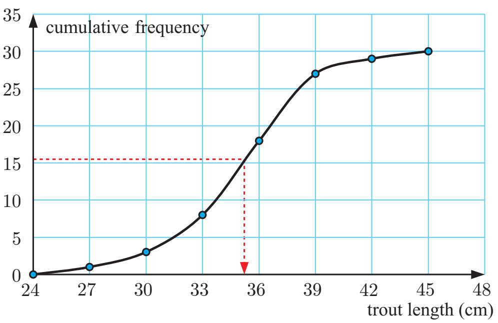


a 26 years **b** 36% **c** **i** 0.527 **ii** 0.0267

4 **a**

Length (cm)	*Frequency*	*Cumulative frequency*
$24 \leqslant x < 27$	1	1
$27 \leqslant x < 30$	2	3
$30 \leqslant x < 33$	5	8
$33 \leqslant x < 36$	10	18
$36 \leqslant x < 39$	9	27
$39 \leqslant x < 42$	2	29
$42 \leqslant x < 45$	1	30

b



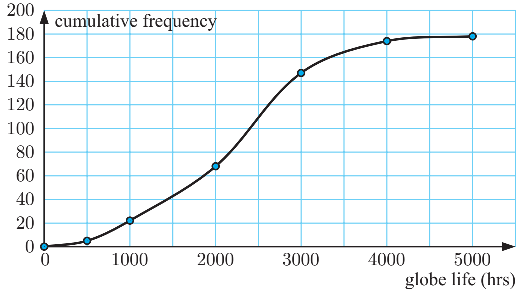
c median ≈ 35 cm

d median = 34.5. Median from graph is a good approximation.

5 **a** 27 min **b** 29 min **c** 31.3 min

d 4.3 min **e** ≈ 28 min

6



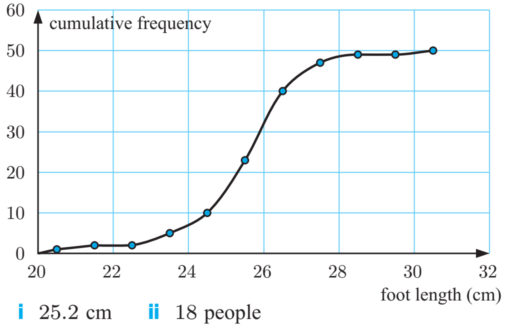
a ≈ 2270 hours **b** ≈ 69% **c** ≈ 63

7 **a** $19.5 \leqslant l < 20.5$

b

Foot length (cm)	*Frequency*	*Cumulative frequency*
$19.5 \leqslant l < 20.5$	1	1
$20.5 \leqslant l < 21.5$	1	2
$21.5 \leqslant l < 22.5$	0	2
$22.5 \leqslant l < 23.5$	3	5
$23.5 \leqslant l < 24.5$	5	10
$24.5 \leqslant l < 25.5$	13	23
$25.5 \leqslant l < 26.5$	17	40
$26.5 \leqslant l < 27.5$	7	47
$27.5 \leqslant l < 28.5$	2	49
$28.5 \leqslant l < 29.5$	0	49
$29.5 \leqslant l < 30.5$	1	50

c



d **i** 25.2 cm **ii** 18 people

EXERCISE 6I.1

1 **a** 1.49 **b** 4.73

2 mean = 55 L, standard deviation ≈ 10.9 L

3 mean ≈ 1.69 kg, standard deviation ≈ 0.182 kg

4 **a** $\overline{x} = 169$, $s \approx 6.05$ **b** $\overline{x} = 174$, $s \approx 6.05$

c The distribution has simply shifted by 5 cm. The mean increases by 5 cm and the standard deviation remains the same.

5 **a** $\overline{x} = 1.01$ kg; $s = 0.17$ **b** $\overline{x} = 2.02$ kg; $s = 0.34$

c Doubling the values doubles the mean and standard deviation.

6 **a** 0.809 **b** 2.8, from volunteer F **c** 0.150

d the extreme value greatly increases the standard deviation

EXERCISE 6I.2

1 $\overline{x} \approx 1.72$ children, $s_n \approx 1.67$ children

2 $\overline{x} \approx 14.5$ years, $s_n \approx 1.75$ years

3 $\overline{x} = 45$ clients, $s_n \approx 3.28$ clients

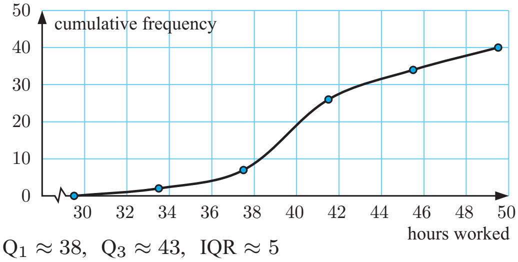
4 $\overline{x} \approx 48.3$ cm, $s_n \approx 2.66$ cm **5** $\overline{x} \approx \$390.30$, $s_n \approx \$15.87$

6 **a** $\overline{x} \approx 40.4$ hours $s_n \approx 4.23$ hours

b $\overline{x} = 40.6$ hours $s_n \approx 4.10$ hours

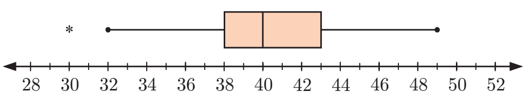
The mean increases slightly, the standard deviation decreases slightly. These are good approximations.

c



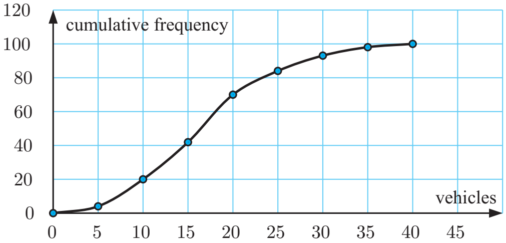
$Q_1 \approx 38$, $Q_3 \approx 43$, $\text{IQR} \approx 5$

d



7 **a** $\overline{x} \approx 17.5$ cars, $s_n \approx 7.87$ cars

b



$Q_1 \approx 11$, $Q_3 \approx 22$, $\text{IQR} \approx 11$

EXERCISE 6I.3

1 **a** Sample A

b Sample A: mean = 8, Sample B: mean = 8

c Sample A: $s_n = 2$, Sample B: $s_n \approx 1.06$
Sample B's standard deviation is smaller than Sample A's. The graph shows the data to be less 'spread out' in Sample B.

2 **a** Andrew: $\overline{x} = 25$, $s_n \approx 4.97$
Brad: $\overline{x} = 30.5$, $s_n \approx 12.6$

b Andrew

3 **a** Rockets: mean = 5.7, range = 11
Bullets: mean = 5.7, range = 11

b We suspect the Rockets, they have two zeros.

c Rockets: $s_n = 3.9$ ⟵ greater variability
Bullets: $s_n \approx 3.29$

d Standard deviation, as it takes into account all data values.

4 **a** No, because of random variation

b **i** the sample mean $\overline{x}$

ii the sample standard deviation s_n

c Less variability in the volume of soft drink per can.

REVIEW SET 6A

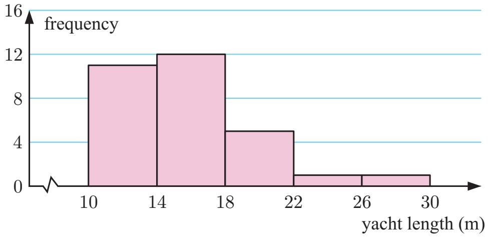
1 **a** quantitative discrete **b** quantitative continuous

c categorical **d** categorical **e** categorical

f quantitative continuous **g** quantitative continuous

h quantitative discrete **i** quantitative discrete

2 **a**



b **i** median = 14.5 m **ii** range = 17.3 m

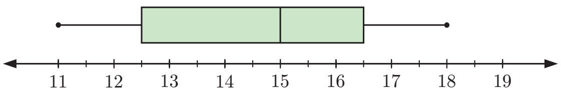
c The data is positively skewed.

3 $a = 2$

4 **a** negatively skewed **b** 47.5% **c** 7.5%

d We do not know all the data values exactly, only the class intervals they fall into.

5



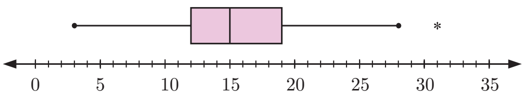
6 **a** 77 days **b** 12 days

7 **a** $\overline{x} \approx 122$, $s_n \approx 7.94$ **b** $\overline{x} \approx 7.01$, $s_n \approx 0.984$

8 **a** min = 3; $Q_1 = 12$; med = 15; $Q_3 = 19$; max = 31

b range = 28; IQR = 7

c



REVIEW SET 6B

1 **a** quantitative continous **b** categorical **c** categorical

d quantitative continous **e** quantitative continous

f quantitative discrete **g** categorical

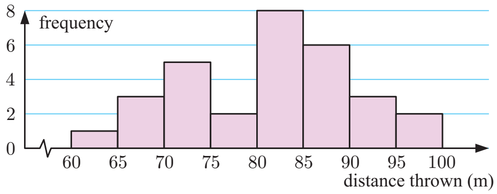
2 **a** minimum = 64.6 m, maximum = 97.5 m

b **i** mean ≈ 81.1 m **ii** median ≈ 83.1 m

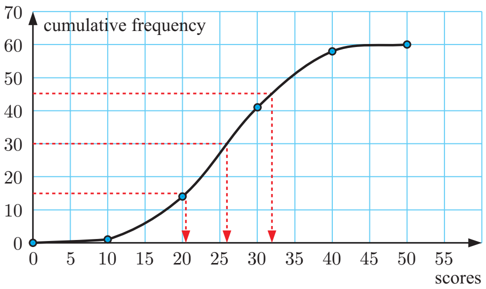
c, d

Distance (m)	*Tally*	*Frequency*
$60 \leqslant d < 65$	|	1
$65 \leqslant d < 70$	|||	3
$70 \leqslant d < 75$	𝍸	5
$75 \leqslant d < 80$	||	2
$80 \leqslant d < 85$	𝍸 |||	8
$85 \leqslant d < 90$	𝍸 |	6
$90 \leqslant d < 95$	|||	3
$95 \leqslant d < 100$	||	2
	Total	30

e



3 **a**



b **i** median ≈ 26.0 **ii** IQR ≈ 12

iii $\overline{x} \approx 26.0$ **iv** $s_n \approx 8.31$

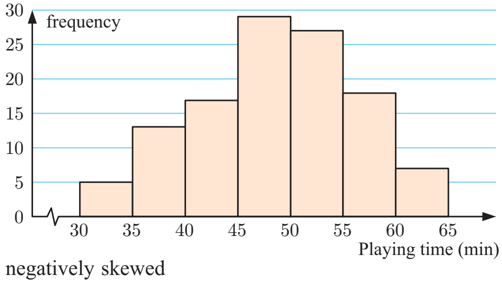
4 **a** **i** £352.50 **ii** £336 **iii** £365.50

b £29.50 **c** $\overline{x} \approx$ £350, $s_n \approx$ £17.80

5 **a** 88 students **b** $m = 24$

6 a $\overline{x} \approx 48.6$ min, $s_n \approx 7.63$ min

b



c negatively skewed

7 range $= 19$; $\text{Q}_1 = 119$; $\text{Q}_3 = 130$; $s_n \approx 6.38$

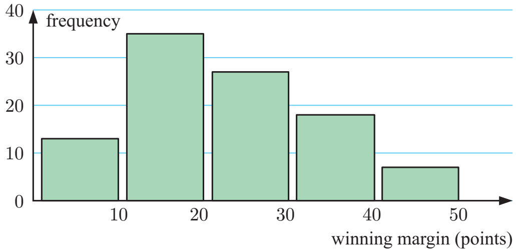
8 a $\overline{x} \approx 29.6$ allsorts, $s_n \approx 1.61$ allsorts

b More investigation is needed.

REVIEW SET 6C

1 $p = 7$, $q = 9$ (or $p = 9$, $q = 7$)

2



3 $\overline{x} \approx 414$ patrons

4 a A: min $= 11$ s; $\text{Q}_1 = 11.6$ s; med $= 12$ s; $\text{Q}_3 = 12.6$ s; max $= 13$ s

B: min $= 11.2$ s; $\text{Q}_1 = 12$ s; med $= 12.6$ s; $\text{Q}_3 = 13.2$ s; max $= 13.8$ s

b i A: range $= 2.0$ s ii A: IQR $= 1.0$ s

B: range $= 2.6$ s B: IQR $= 1.2$ s

c i A, the median time is lower.

ii B, the range and IQR are higher.

5 $\overline{x} \approx$ €104, $s_n \approx$ €19.40

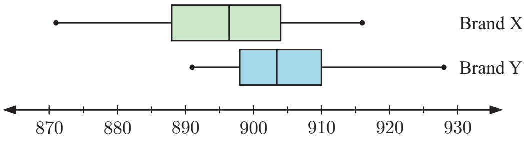
6 a 120 students b 65 marks c 54 and 75

d 21 marks e $\approx 73\%$ f 82 marks

7 a

	Brand X	*Brand Y*
min	871	891
Q_1	888	898
median	896.5	903.5
Q_3	904	910
max	916	928
IQR	16	12

b



c i Brand Y, as the median is higher.

ii Brand X, as the IQR is lower, so less variations.

EXERCISE 7A

1 a $5 \in D$ b $6 \notin G$ c $d \notin \{a, e, i, o, u\}$

d $\{2, 5\} \subseteq \{1, 2, 3, 4, 5, 6\}$

e $\{3, 8, 6\} \not\subseteq \{1, 2, 3, 4, 5, 6\}$

2 a i $\{9\}$ ii $\{5, 6, 7, 8, 9, 10, 11, 12, 13\}$

b i $\varnothing$ ii $\{1, 2, 3, 4, 5, 6, 7, 8\}$

c i $\{1, 3, 5, 7\} = A$ ii $\{1, 2, 3, 4, 5, 6, 7, 8, 9\} = B$

3 a 5 b 6 c 2 d 9

4 a True b True c False d True

e False f True g True h False

5 a finite b infinite c infinite d infinite

6 a True b True c False d True

7 a disjoint b not disjoint 8 True

9 a i $\varnothing$, $\{a\}$, $\{b\}$, $\{c\}$, $\{a, b\}$, $\{a, c\}$, $\{b, c\}$, $\{a, b, c\}$, so 8 subsets

ii $\varnothing$, $\{a\}$, $\{b\}$, $\{c\}$, $\{d\}$, $\{a, b\}$, $\{a, c\}$, $\{a, d\}$, $\{b, c\}$, $\{b, d\}$, $\{c, d\}$, $\{a, b, c\}$, $\{a, b, d\}$, $\{a, c, d\}$, $\{b, c, d\}$, $\{a, b, c, d\}$, so 16 subsets

b 2^n, $n \in \mathbb{Z}^+$

EXERCISE 7B

1 a finite b infinite c infinite d infinite

2 a i A is the set of all x such that x is an integer between -1 and 7, including -1 and 7.

ii $\{-1, 0, 1, 2, 3, 4, 5, 6, 7\}$ iii 9

b i A is the set of all x such that x is a natural number between -2 and 8.

ii $\{0, 1, 2, 3, 4, 5, 6, 7\}$ iii 8

c i A is the set of all x such that x is a real number between 0 and 1, including 0 and 1.

ii not possible iii infinite

d i A is the set of all x such that x is a rational number between 5 and 6, including 5 and 6.

ii not possible iii infinite

3 a $A = \{x \mid -100 < x < 100,\ x \in \mathbb{Z}\}$

b $A = \{x \mid x > 1000,\ x \in \mathbb{R}\}$

c $A = \{x \mid 2 \leqslant x \leqslant 3,\ x \in \mathbb{Q}\}$

4 a $A \subseteq B$ b $A \not\subseteq B$ c $A \subseteq B$ d $A \subseteq B$

e $A \not\subseteq B$ f $A \not\subseteq B$

EXERCISE 7C

1 a $C' = \{\text{consonants}\}$ b $C' = \mathbb{N}$

c $C' = \{x \mid x \geqslant -4,\ x \in \mathbb{Z}\}$

d $C' = \{x \mid 2 < x < 8,\ x \in \mathbb{Q}\}$

2 a $\{2, 3, 4, 5, 6, 7\}$ b $\{0, 1, 8\}$ c $\{5, 6, 7, 8\}$

d $\{0, 1, 2, 3, 4\}$ e $\{5, 6, 7\}$ f $\{2, 3, 4, 5, 6, 7, 8\}$

g $\{2, 3, 4\}$

3 a 9 b 11 4 a False b True

5 a $\{1, 2, 10, 11, 12\}$ b $\{1, 2, 3, 4, 12\}$

c $\{1, 8, 9, 10, 11, 12\}$ d $\{3, 4, 5, 6, 7\}$

e $\{1, 2, 8, 9, 10, 11, 12\}$ f $\{8, 9, 10, 11\}$

g $\{1, 2, 5, 6, 7, 8, 9, 10, 11, 12\}$ h $\{2, 10, 11\}$

6 a $P = \{2, 3, 5, 7, 11, 13, 17, 19, 23\}$ b $\{2, 5, 11\}$

c $\{2, 3, 4, 5, 7, 11, 12, 13, 15, 17, 19, 23\}$

d $12 = 9 + 6 - 3$ ✓

7 a $P = \{1, 2, 4, 7, 14, 28\}$, $Q = \{1, 2, 4, 5, 8, 10, 20, 40\}$

b $\{1, 2, 4\}$ c $\{1, 2, 4, 5, 7, 8, 10, 14, 20, 28, 40\}$

d $11 = 6 + 8 - 3$ ✓

8 a $M = \{32, 36, 40, 44, 48, 52, 56\}$, $N = \{36, 42, 48, 54\}$