

Chapter

12

Sampling and data

Contents:

- A** Errors in sampling and data collection
- B** Sampling methods
- C** Writing surveys
- D** Types of data
- E** Simple discrete data
- F** Grouped discrete data
- G** Continuous data



OPENING PROBLEM

A supermarket sells 1 kg bags of grapes.

Things to think about:

- a Would you expect every bag of grapes to weigh *exactly* 1 kg?
- b In what range of weights would you expect most of the bags of grapes to lie?
- c Adriana weighed 50 bags of grapes which were delivered to the supermarket in one day.
 - i Do you think this sample will be *representative* of all the bags of grapes sold by the supermarket? Explain your answer.
 - ii What type of graph should Adriana use to display her results?
 - iii What would you expect Adriana's graph to look like?

In statistics we collect information about a group of individuals, then analyse this information to draw conclusions about those individuals.

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

Data:	information about the characteristics of a group of individuals
Categorical variable:	describes a particular characteristic which can be divided into categories
Quantitative variable:	describes a characteristic which has a numerical value that can be counted or measured
Population:	an entire collection of individuals about which we want to draw conclusions
Census:	the collection of information from the whole population
Parameter:	a numerical quantity measuring some aspect of a population
Sample:	a group of individuals selected from a population
Survey:	the collection of information from a sample
Statistic:	a quantity calculated from data gathered from a sample, usually used to estimate a population parameter

STATISTICAL INVESTIGATIONS

A **statistical investigation** should include the following steps.

Step 1: State the problem

We need to decide what we are investigating and then describe it exactly. We also need to decide on **variables** we can measure to help us understand our topic of interest.

Step 2: Choose a sample

In general it is impossible to obtain data from every single individual in a population. We therefore need to choose a **sample** which is representative of the whole population.

Step 3: Collect the data

Having chosen the sample, we now need to decide exactly what we are going to ask the individuals, or how we are going to measure the variables. This is important to ensure the data we collect is actually going to be appropriate or useful for our specific investigation.

Step 4: Organise and display the data

How we display data will depend on its form. Our aim is to identify features of the data more easily.

Step 5: Calculate descriptive statistics

Descriptive statistics help us to understand the **centre** and **spread** of the data.

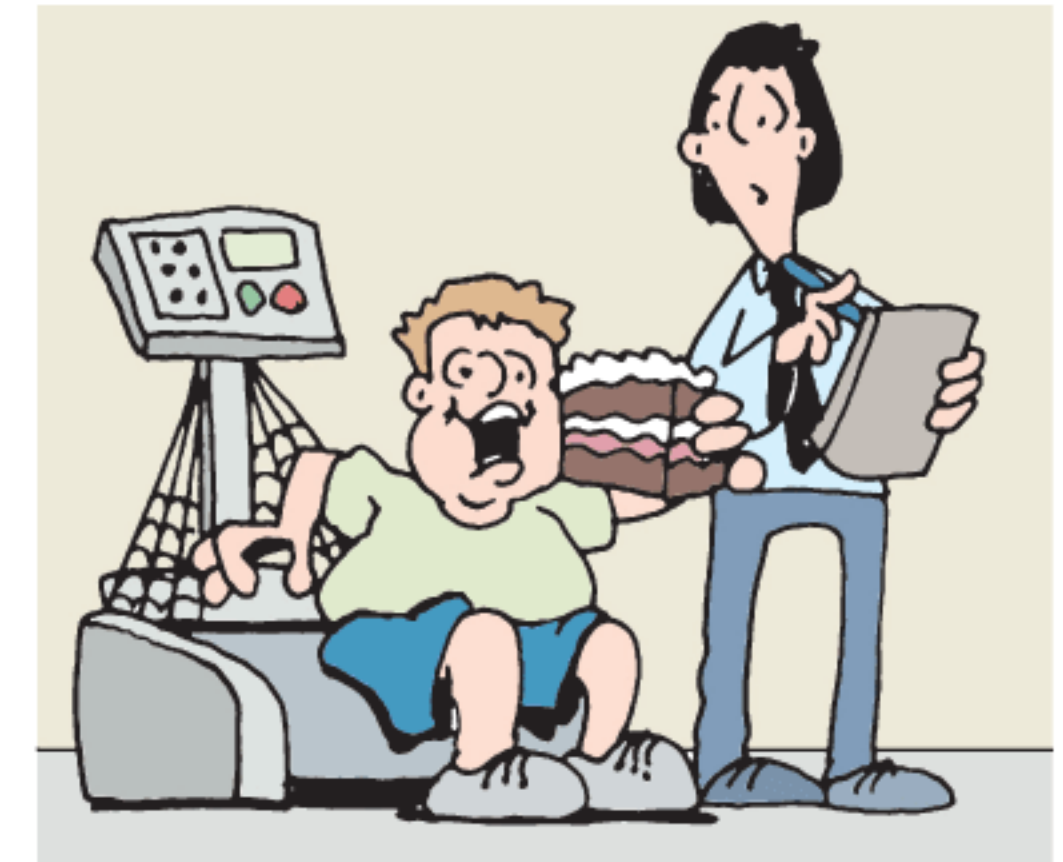
Step 6: Interpret statistics

We need to describe what the data and its statistics tell us about the problem being investigated.

DESCRIBING A PROBLEM

Before we start collecting data, we need to consider what the **goals** of our investigation are. This will help us decide what **variables** we will need to measure in order to achieve these goals.

For example, if the aim of our investigation was to study the health of primary school children, we could consider variables such as *height*, *weight*, *daily nutritional intake*, *amount of exercise*, *amount of sleep*, *lung capacity*, *eye test results*, and *medical conditions*.



Factors that should be considered when selecting variables include:

- **relevance to the investigation**

For example, if we wanted to focus on childhood obesity, the most relevant variables would be *weight*, *amount of exercise*, and *daily nutritional intake*.

- **limitations in measuring a variable**

Some variables can be very difficult or impractical to measure, so they might not be suitable for our investigation.

For example:

- ▶ Counting every hair on a person’s head within a reasonable amount of time is unrealistic.
- ▶ We may not have the equipment to properly measure *lung capacity*.
- ▶ A wooden ruler would not be suitable for measuring the distance around an athletics track.

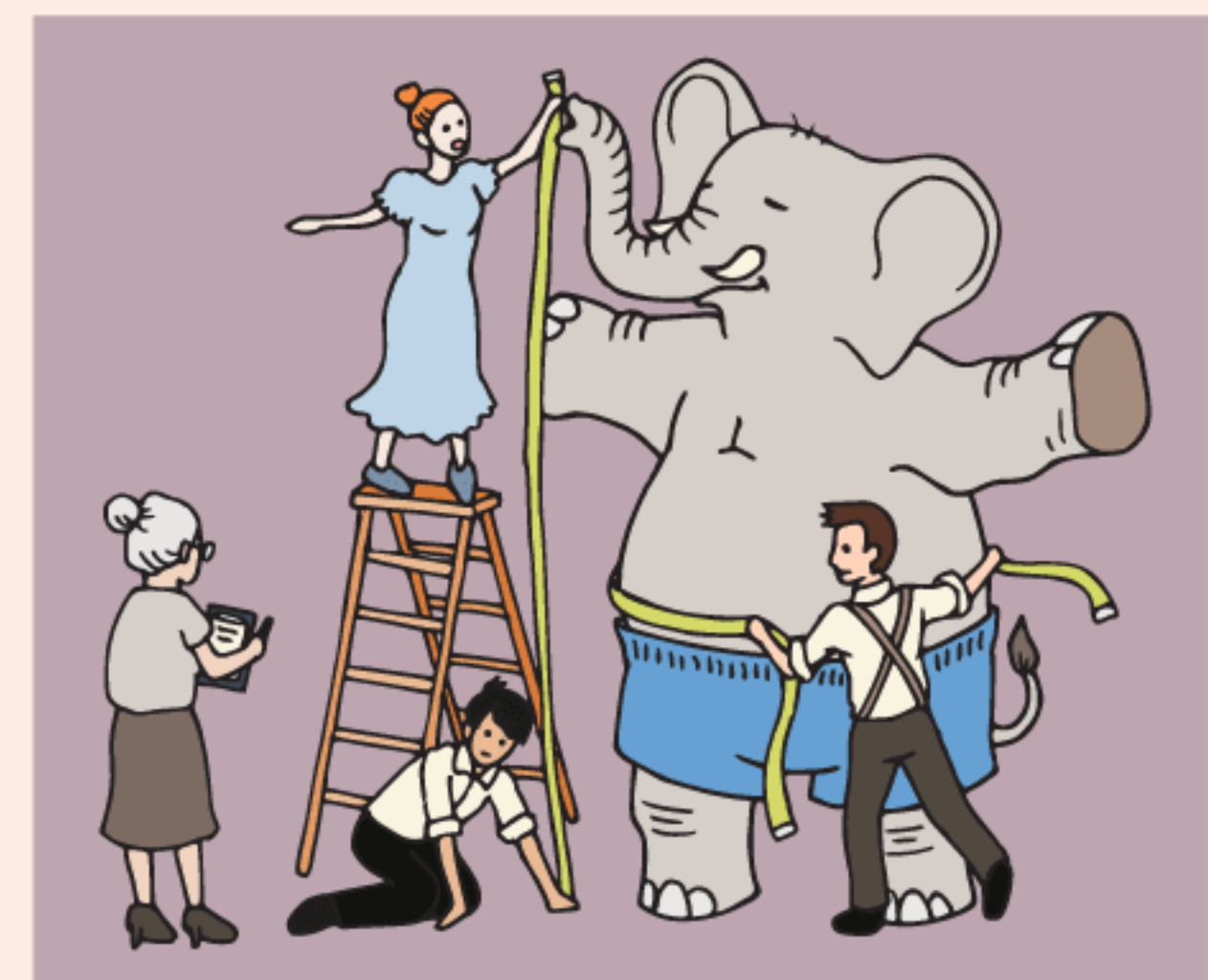
DISCUSSION

1 How big is an elephant?

Before we can answer this question, we need to decide what we mean by “big”, and what variable(s) we will use to measure it.

a Which of these variables would be suitable for measuring how big an elephant is?

- | | |
|-------------------------|----------------|
| ● height | ● length |
| ● volume | ● mass |
| ● length of trunk | ● girth |
| ● area of ears | ● size of feet |
| ● circumference of tail | |



b Is it reasonable to use more than one variable to describe how “big” an elephant is?

c Can you think of other suitable variables for this investigation?

d What other things do we need to consider to specify *exactly* what we are investigating? For example, how might gender, age, subspecies, environment, and capacity status be important?

2 How much sleep does a person get each night?

Is this question sufficient to describe a statistical investigation? Discuss the question, work out what *you* want it to mean, then rewrite the question so it better describes the goals of your investigation.

3 Should test results be used to measure academic performance?

Test and examination results are often used to assess the academic performance of students. Do you think this is necessarily fair?

A**ERRORS IN SAMPLING AND DATA COLLECTION**

A **census** is the most accurate way to investigate a population of interest. However, in most situations it is impractical or impossible to obtain data from the entire population. Instead, we can conduct a **survey** of a well-chosen **sample** of the population.

When we collect data to estimate a characteristic of a population, our estimate will almost certainly be different from the actual characteristic of the population. This difference is referred to as **error**.

There are four main categories of error: **sampling error**, **coverage error**, **non-response error**, and **measurement error**.

Sampling error occurs when a characteristic of a sample differs from that of the whole population. This error is random, and will occur even for samples which are well-chosen to avoid bias.

Coverage errors occur when a sample does not truly reflect the population we are trying to find information about.

To avoid coverage errors, samples should be **sufficiently large** and **unbiased**.

For example, suppose you are interested in the health of bees on a particular island.

- If you only collect data from 10 bees, you will not get a reliable idea of the health of all bees on the island.
- If you only collect data from one particular bee hive, the sample may not be **representative** of all of the bees on the island. For example, the hive you pick may be stressed and preparing to swarm, whereas its neighbouring hives may be healthy. The sample would therefore be a **biased sample**, and would be unreliable for forming conclusions about the whole population.



Non-response errors occur when a large number of people selected for a survey choose not to respond to it. For example:

- An online survey is less likely to be completed by elderly people who are unfamiliar with technology. This means that elderly people will be under-represented in the survey.
- In surveys on customer satisfaction, people are more likely to respond if they are dissatisfied.

Measurement error refers to inaccuracies in measurement at the data collection stage. For example:

- When we record a person's height to the nearest centimetre, the recorded height is slightly different from the person's *exact* height.
- If the questions in a survey are not well worded, they may be misunderstood and produce answers which are not relevant to the question. Survey construction is discussed more thoroughly in **Section C**.

EXERCISE 12A

- 1** A new drug called Cobrasyl has been developed for the treatment of high blood pressure in humans. A derivative of cobra venom, it is able to reduce blood pressure to an acceptable level. Before its release, a research team treated 7 high blood pressure patients with the drug, and in 5 cases it reduced their blood pressure to an acceptable level.

Do you think this sample can be used to draw reliable conclusions about the drug's effectiveness for all patients? Explain your answer.

- 2** 50 people in a Toronto shopping mall were surveyed. It was found that 20 of them had been to an ice hockey game in the past year. From this survey, it was concluded that "40% of people living in Canada have been to an ice hockey game in the past year".

Give *two* reasons why this conclusion is unreliable.

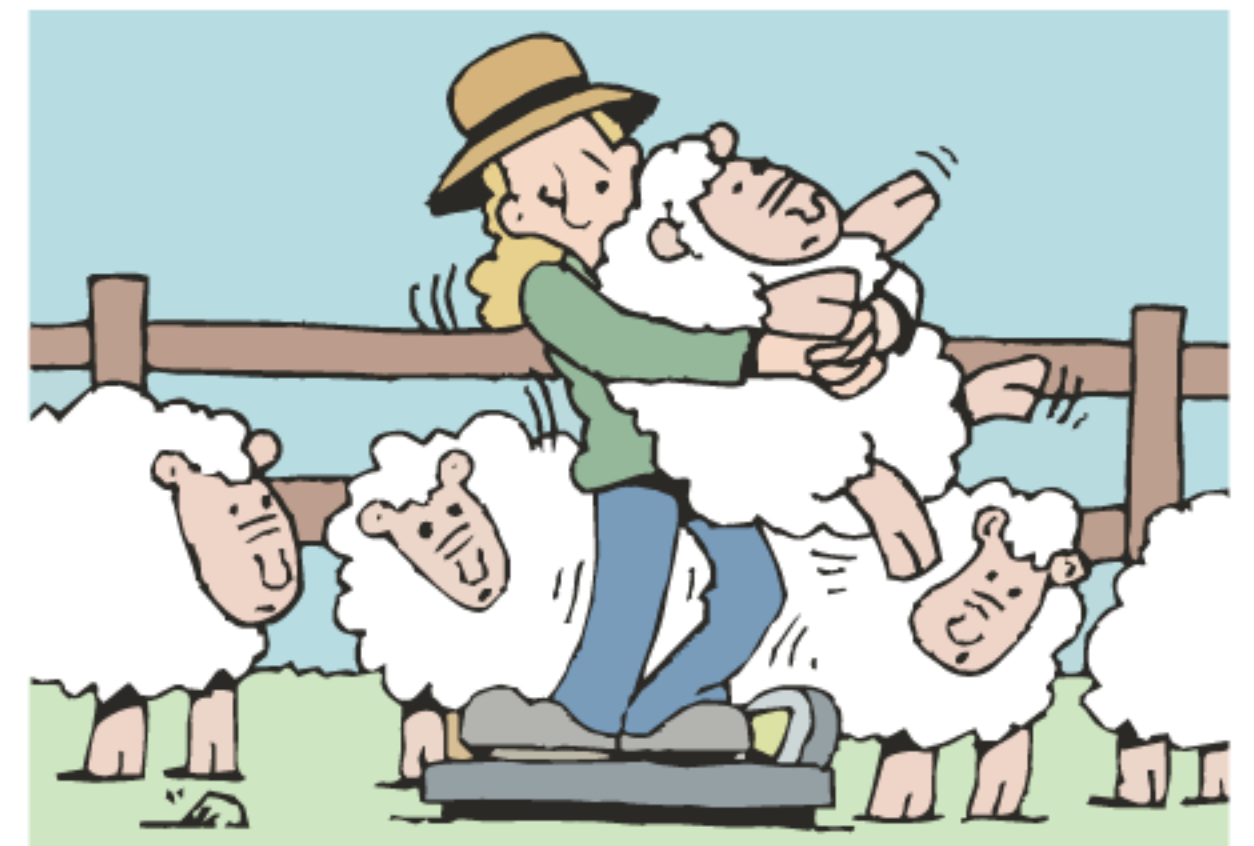
- 3** A polling agency is employed to investigate the voting intention of residents in a particular electorate. From the data collected, they want to predict the election result for that electorate in the next election. Explain why each of the following situations may produce a biased sample:

- a** A random selection of people in the local large shopping complex is surveyed between 1 pm and 3 pm on a weekday.
- b** The members of the local golf club are surveyed.
- c** A random sample of people at the local train station between 7 am and 9 am are surveyed.
- d** A door to door visit is undertaken, surveying every voter in a particular street.

- 4** Jennifer wants to estimate the average weight of the 2000 sheep on her farm. She selects a sample of 10 sheep, and weighs them.

Explain why this approach may produce a:

- a** coverage error
- b** measurement error.



- 5** Jack owns 800 apple trees. To determine how many apples the trees are producing, he instructs his four sons to each count the apples from 200 trees.

- a** Explain why there will be no sampling error in this process.
- b** Two of the sons only count the apples on the tree itself, whilst the other two sons also count the apples on the ground beneath the tree. What type of error is this?

- 6** A survey company is interested in whether people feel overworked at their jobs. They mail out a survey to 5000 workers, and ask the workers to mail back the survey.

- a** Explain why this survey may produce a significant non-response error.
- b** What would be the advantages and disadvantages of conducting the survey online instead of by mail?

- 7** A national sporting organisation has over 300 000 members. Every member is invited to complete an online survey regarding the management structure of the organisation. Only 16% of the members responded.

- a** Do you think the non-response error in this situation is likely to produce a biased sample? Explain your answer.
- b** Does such a high non-response error necessarily invalidate findings from the survey? Discuss your answer.

DISCUSSION

- Why do you think companies offer incentives for people to complete their surveys?
- Which of the following incentives for completing a survey would be more effective?
 - ▶ A chance to win a prize as shown alongside.
 - ▶ A guaranteed discount or promotional code for the participant to use on their next purchase.
- Is it ethical to offer monetary compensation for completing a survey?



B

SAMPLING METHODS

In general, the best way to avoid bias when selecting a sample is to make sure the sample is **randomly selected**. This means that each member of the population has the same chance of being selected in the sample.

We will look at five sampling methods:

- **simple random sampling**
- **stratified sampling**
- **systematic sampling**
- **quota sampling**
- **convenience sampling**

SIMPLE RANDOM SAMPLING

Suppose 3 students are to be sampled from a class of 30 students. The names of all students in the class are placed in a barrel, and 3 names are drawn from the barrel.

Notice that:

- Each student has the same chance ($\frac{1}{30}$) of being selected.
- Each set of 3 students is just as likely to be selected as any other. For example, the selection {Bruce, Jane, Sean} is just as likely to occur as {Jane, Peter, Vanessa}.



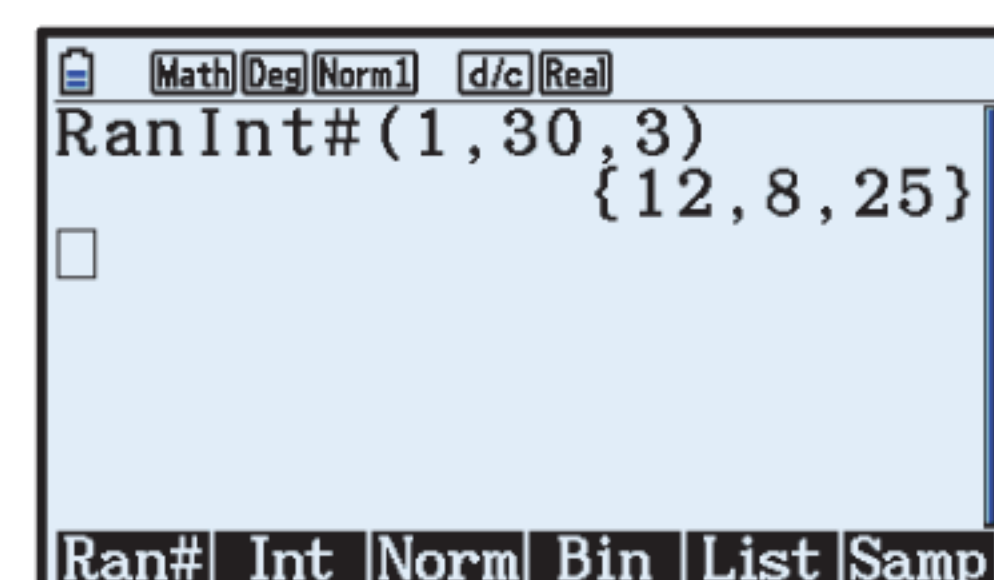
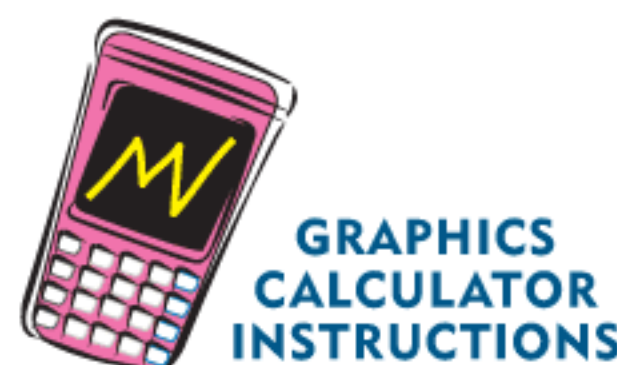
This type of sampling is called **simple random sampling**.

For a **simple random sample** of size n from a population:

- Each member of the population has the same chance of being selected in the sample.
- Each set of n members of the population has the same chance of being selected as any other set of n members.

Instead of drawing names from a barrel, it is usually more practical to number the members of the population, and use a random number generator to select the sample.

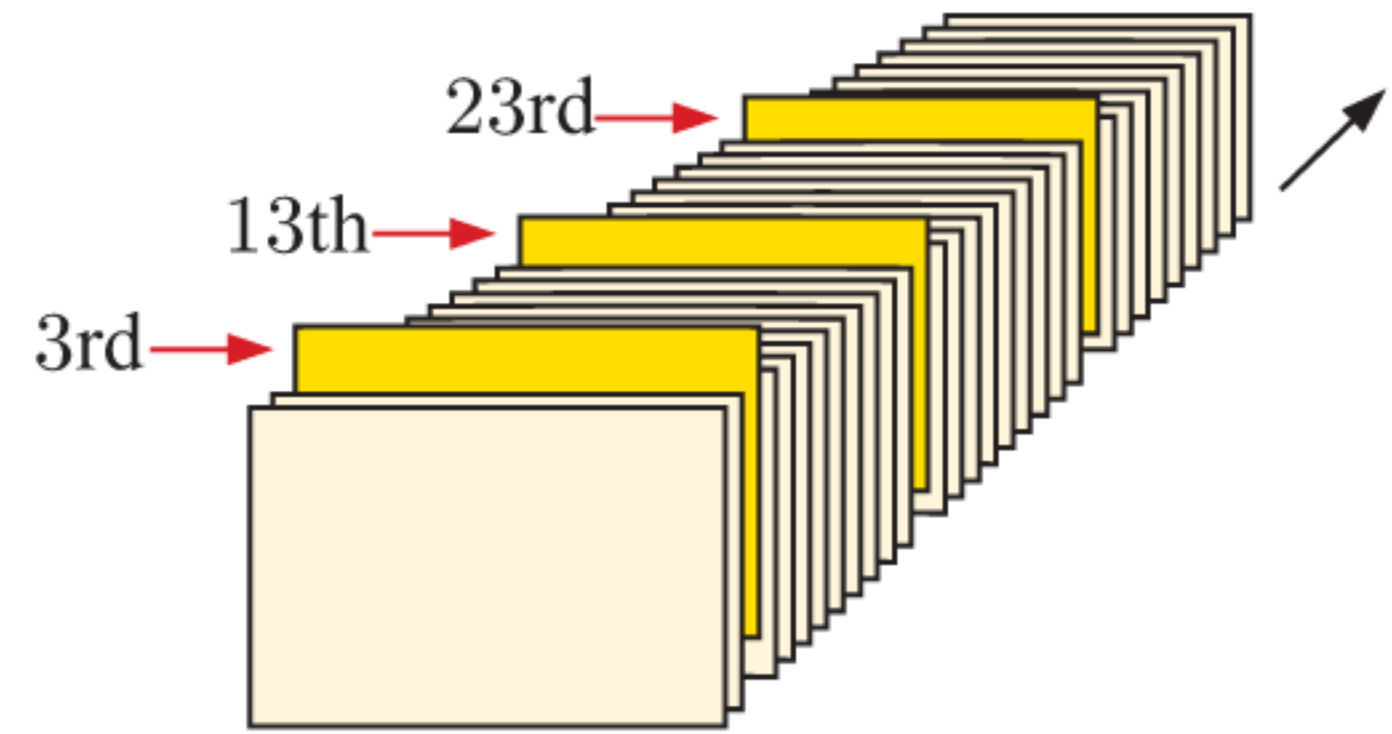
You can use your calculator to generate random numbers. In this case, the 8th, 12th, and 25th students would be selected for the sample.



SYSTEMATIC SAMPLING

In **systematic sampling**, the sample is created by selecting members of the population at regular intervals.

For example, an accountancy firm may wish to sample the files of $\frac{1}{10}$ th of their clients. They choose a starting file from 1 to 10 (for example, 3), and then select every 10th file after that. So, they would select the 3rd file, then the 13th, 23rd, 33rd, and so on.



Systematic sampling is useful when not all members of the population are available for sampling at the same time. An example of this is the sampling of cars which pass through a particular intersection during the day.

Example 1



Management of a large city store wishes to find out whether potential customers like the look of a new product. They decide to sample 5% of the customers using a systematic sample. Show how this sample would be selected.

$$5\% = \frac{5}{100} = \frac{1}{20}$$

So, every 20th customer will be sampled.

A starting customer is selected from 1 to 20. In this case it is customer 7.

So, the store would select the 7th customer, then the 27th, 47th, 67th, and so on.

```
NORMAL FLOAT AUTO REAL DEGREE MP
randInt(1,20,1)
.....{7}
```

CONVENIENCE SAMPLING

In many situations, people are chosen simply because they are easier to select or more likely to respond.

For example, consider a researcher conducting a survey regarding environmental issues. The researcher decides to stand in a pedestrian mall and ask people walking past. It is easiest for the researcher to ask people who are:

- walking closest to them
- walking slowly
- not already in a conversation or using their phone.

These types of samples are known as **convenience samples** because they are convenient for the experimenter.

DISCUSSION

Do you think convenience samples will often be biased?

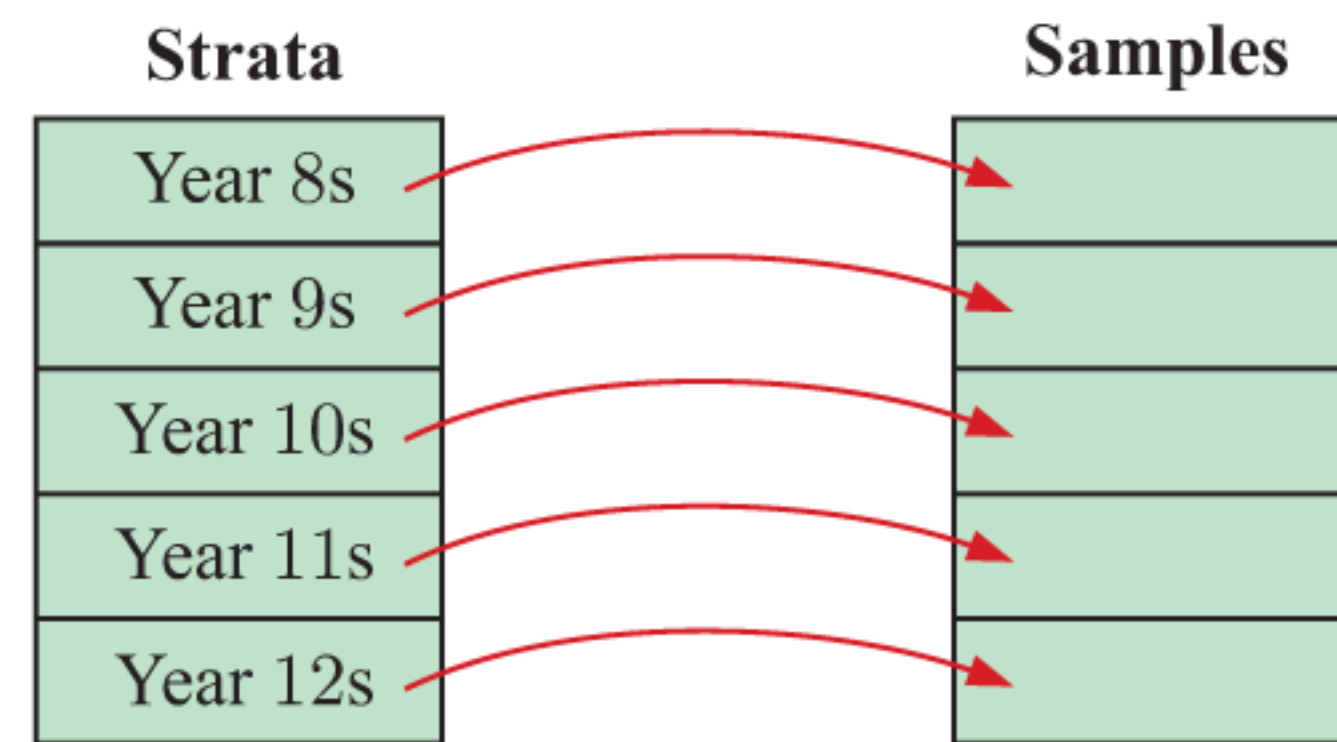
Discuss any possible bias if the researcher in the mall was studying:

- mobile internet usage
- social media
- personal relationships
- mental health issues.

STRATIFIED SAMPLING AND QUOTA SAMPLING

Stratified sampling and **quota sampling** are useful when the population can be divided into subgroups, and you want to make sure each subgroup is represented fairly in the sample.

For example, a school may want to know the opinions of its students on which charities it should support in the school fun run. To make sure each year level is represented fairly, the number of students sampled from each year level should be proportional to the fraction of the total number of students that year level represents.



Example 2

Self Tutor

In our school there are 137 students in Year 8, 152 in Year 9, 174 in Year 10, 168 in Year 11, and 121 in Year 12. A sample of 50 students is needed. How many should be randomly selected from each year?

Total number of students in the school = $137 + 152 + 174 + 168 + 121 = 752$

For the sample, we want:

$$\text{number of Year 8 students} = \frac{137}{752} \times 50 \approx 9$$

$$\text{number of Year 9 students} = \frac{152}{752} \times 50 \approx 10$$

$$\text{number of Year 10 students} = \frac{174}{752} \times 50 \approx 12$$

$$\text{number of Year 11 students} = \frac{168}{752} \times 50 \approx 11$$

$$\text{number of Year 12 students} = \frac{121}{752} \times 50 \approx 8$$

We should select 9 students from Year 8, 10 from Year 9, 12 from Year 10, 11 from Year 11, and 8 from Year 12.

Year 8 students represent $\frac{137}{752}$ of the school, so they should also represent $\frac{137}{752}$ of the sample.



Ideally, we would want the individuals from each strata to be randomly selected to minimise bias. If this can be done, the sample is a **stratified sample**. Otherwise, if the individuals are specifically selected by the experimenter (such as in a convenience sample) then the sample is a **quota sample**.

EXERCISE 12B

- 1 Use your calculator to select a random sample of:
 - a 6 different numbers between 5 and 25 inclusive
 - b 10 different numbers between 1 and 25 inclusive
 - c 6 different numbers between 1 and 45 inclusive
 - d 5 different numbers between 100 and 499 inclusive.

You may need to generate additional random numbers if a number appears more than once.



- 2 A chocolate factory produces 80 000 blocks of chocolate per day. Today, the factory operator wants to sample 2% of the blocks for quality testing. He uses a systematic sample, starting from the 17th block.
 - a List the first five blocks to be sampled.
 - b Find the total size of the sample.

- 3** Click on the icon to obtain a printable calendar for 2019 showing the weeks of the year. Each day is numbered.

CALENDAR



Using a random number generator, choose a sample from the calendar of:

- a** five different dates
- b** a complete week starting with a Monday
- c** a month
- d** three different months
- e** three consecutive months
- f** four different Wednesdays.

Explain your method of selection in each case.

January	February	March	April	May
1 Tu (1) Wk 1	1 Fr (32)	1 Fr (60)	1 Mo (91)	1 We (121)
2 We (2)	2 Sa (33)	2 Sa (61)	2 Tu (92) Wk 14	2 Th (122)
3 Th (3)	3 Su (34)	3 Su (62)	3 We (93)	3 Fr (123)
4 Fr (4)	4 Mo (35)	4 Mo (63)	4 Th (94)	4 Sa (124)
5 Sa (5)	5 Tu (36) Wk 6	5 Tu (64) Wk 10	5 Fr (95)	5 Su (125)
6 Su (6)	6 We (37)	6 We (65)	6 Sa (96)	6 Mo (126)
7 Mo (7)	7 Th (38)	7 Th (66)	7 Su (97)	7 Tu (127) Wk 19
8 Tu (8) Wk 2	8 Fr (39)	8 Fr (67)	8 Mo (98)	8 We (128)
9 We (9)	9 Sa (40)	9 Sa (68)	9 Tu (99) Wk 15	9 Th (129)
...

- 4** An annual dog show averages 3540 visitors. The catering manager is conducting a survey to investigate the proportion of visitors who will spend more than €20 on food and drinks at the show. He decides to survey the first 40 people through the gate.

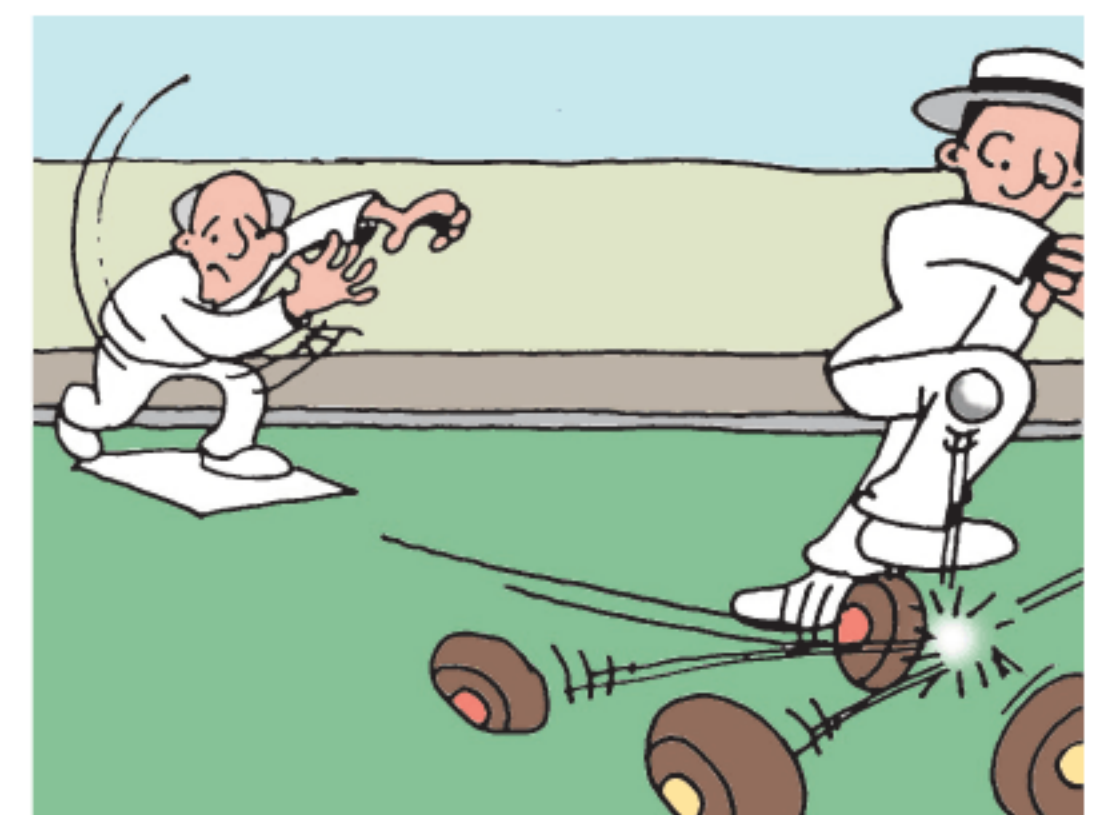
- a** Identify the sampling method used.
- b** Discuss any problems with the sampling method.
- c** Suggest a better sampling method that includes a suitable sample size and which better represents the population.

- 5** A library manager is interested in the number of people using the library each day. She decides to perform a count every 28th day for one year, starting next Monday.

- a** What type of sampling method is this?
- b** How many days will be in her sample?
- c** Explain why the sample may be biased.

- 6** A sporting club wants to ask its members some questions about the clubhouse. The club has 80 tennis members, 60 lawn bowls members, and 20 croquet members.

- a** How many members does the club have in total?
- b** The club decides to use a sample of 40. How many members of each sport should be sampled?



- 7** A large retail store has 10 departmental managers, 24 supervisors, 65 senior sales staff, 98 junior sales staff, and 28 shelf packers. The company director wishes to interview a sample of 30 staff to obtain their view of operating procedures. How many of each group should be selected for the sample?

- 8** Mona wants to gauge the opinions of her peers on the design of the school's yearbook. She uses her own home room class as her sample.

- a** Explain why Mona's sample is a convenience sample.
- b** In what ways will Mona's sample be biased?
- c** Suggest a more appropriate sampling method that Mona should use.

- 9 Lucian is a school counsellor. He wants to raise awareness of student cyber-bullying with the students' parents. Lucian therefore wants to find out whether students at the school have discussed the issue with their parents.
- a Why might it be impractical for Lucian to use a simple random sample or systematic sample?
 - b Lucian wants to make sure that each gender is appropriately represented in his sample. Should he use a stratified sample or quota sample?

10 The 200 students in Years 11 and 12 at a high school were asked whether or not they had ever smoked a cigarette. The replies received were:

nnny nnnyn ynnnn yynyy ynyny ynnyn nyynn yynyn ynynn nyynn
 ynnyn yynyy nnyyy yyyyy nnyyy nnnnn nnyny ynyny nnyyy ynynn
 nynnn ynyyn nnyny ynyyy ynnnn yyyyn ynnnn nynyn yyyny ynnyy
 nynnn yynny nyynn yynyn ynynn nyyyn ynnyy nyyny nnyny ynnnn

- a Why is this considered to be a census?
- b Find the actual proportion of all students who said they had smoked.
- c Discuss the validity and usefulness of the following sampling methods which could have been used to estimate the proportion in **b**:
 - i sampling the first five replies
 - ii sampling the first ten replies
 - iii sampling every second reply
 - iv sampling the fourth member of every group of five
 - v randomly selecting 30 numbers from 1 to 200 and choosing the response corresponding to that number
 - vi sampling 20% of Year 11 students and 20% of Year 12 students.
- d Are any of the methods in **c** examples of simple random sampling, systematic sampling, stratified sampling, or quota sampling?

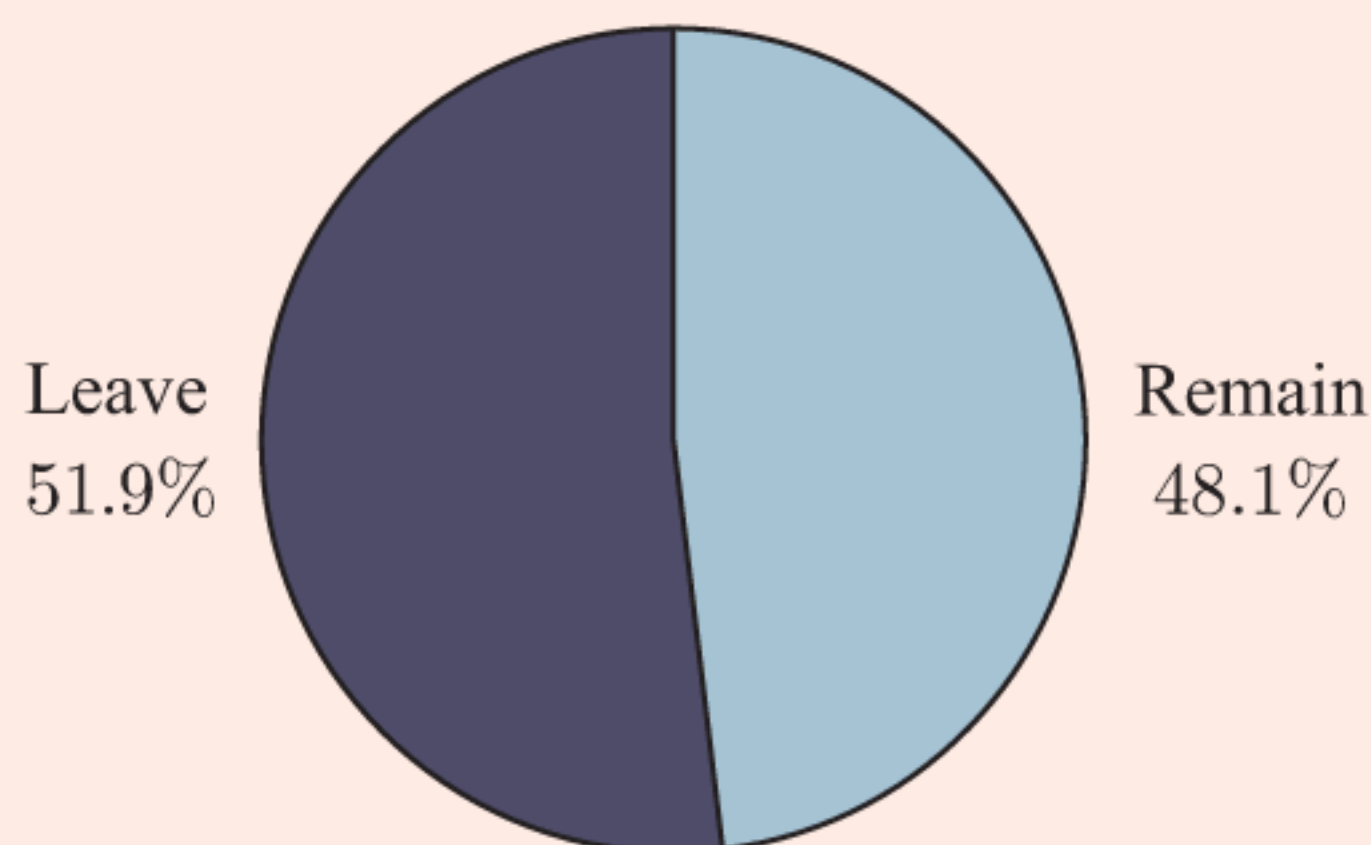


DISCUSSION

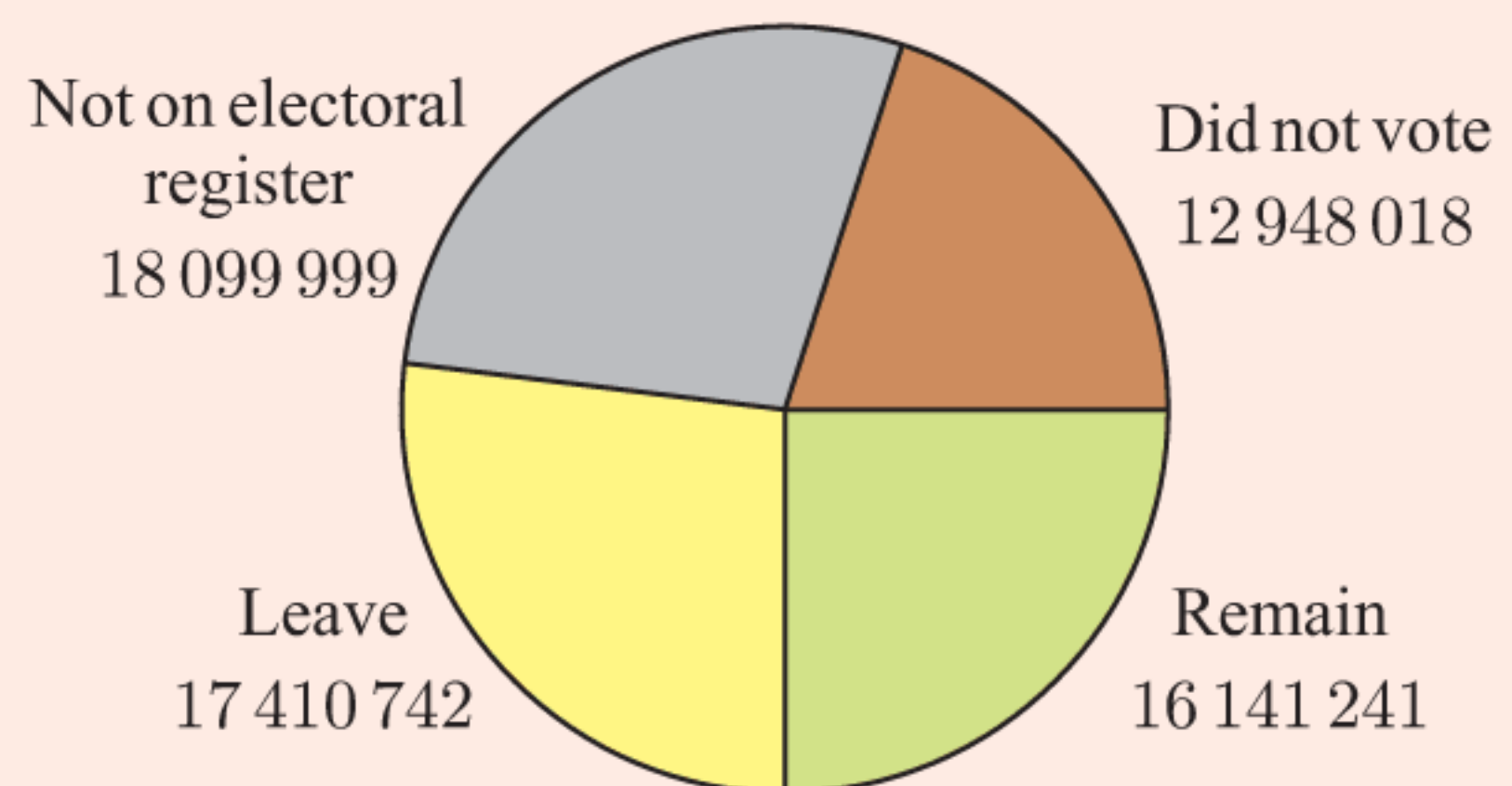
The so-called “Brexit” referendum of 2016 to determine whether the United Kingdom would remain part of the European Union is one of the most controversial democratic referendums in recent history.

- 1 Was the referendum a census or a sample?
- 2 What sampling errors may have been present? In what ways might the sample have been biased?
- 3 33 551 983 votes were counted in the referendum, and it was decided by a simple majority of 51.9% to 48.1% that the United Kingdom would leave. This is shown in the first pie chart.

“Brexit” referendum votes



Broader view of “Brexit” referendum



In the second pie chart we take a broader view to include those who did not vote and those not on the electoral register.

- a Do you think that the United Kingdom leaving the European Union can be considered “the will of the people”?
- b Do you think it is a good idea to have a non-compulsory referendum which can be carried with only a simple majority?

THEORY OF KNOWLEDGE

Clinical trials are commonly used in medical research to test the effectiveness of new treatments for conditions or diseases. They require randomly sampling patients who have the condition or disease in question.

A clinical trial usually involves dividing the sampled individuals into 2 groups:

- A **control group** that serves as the baseline for comparison. This group is usually given either a placebo or the best currently available treatment.
- A **treatment group** that receives the new treatment to be tested.

Ideally, the two groups should be as similar as possible so that differences between the results for the two groups more accurately reflect the differences between treatments, rather than the differences between individuals.

- 1 Are clinical trials necessarily practical for studying treatments of extremely rare diseases?
- 2 How should groups be sampled?

As clinical trials are experiments involving people, the consideration of **ethics** is particularly important. In 1975, the **Declaration of Helsinki**^[1] was written by the World Medical Association to provide guidelines on human experimentation.

Key points of the Declaration include:

- Patients must be made fully aware of all possible risks before they give their consent to participate.
- Patients must never be given a treatment that is known to be inferior.
- Patients are allowed to withdraw from the study at any time.

- 3 Are these ethical guidelines applicable to *any* experiment or survey that involves people?
- 4 Are there the same ethical concerns about experiments involving:
 - a animals
 - b non-living things?

In June 2014, Facebook published a study on the effects of omitting certain words from posts in a user’s “News feed” on their moods and emotions^[2]. This study was heavily criticised for its lack of ethical consideration. The only consent to use people’s data was obtained within Facebook’s general Terms and Conditions document, which must be agreed to upon making an account.

- 5 Do you think this study was an acceptable use of users’ data?
- 6 Do you think using the general Terms and Conditions document to justify “informed consent” is *fair*?
- 7 Are ethics more important than research?

- [1] www.wma.net/policy/current-policies/
- [2] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. “Experimental evidence of massive-scale emotional contagion through social networks”. In: *Proceedings of the National Academy of Sciences* 111.24 (2014), pp. 8788 - 8790. ISSN: 0027-8424. DOI: 10.1073/pnas.1320040111. eprint: <http://www.pnas.org/content/111/24/8788.full.pdf>. URL: <http://www.pnas.org/content/111/24/8788>.

C

WRITING SURVEYS

Surveys are one of the most common and simple methods for collecting data from a sample. Surveys usually consist of a series of questions which can be asked in a written **questionnaire** or an oral **interview**.

When writing survey questions we need to be very careful about how they are **worded**. Poorly worded questions can lead to misinterpretation and yield unintended or inaccurate answers. Such answers would be regarded as **measurement errors**.

GENERAL GUIDELINES FOR QUESTION WRITING

- **Keep questions simple and clear**

Simple wording will help respondents understand what the question is asking, and the context in which it is being asked. You should only ask for **one** response per question.

Questions containing double negatives, such as “Do you disagree with not vaccinating children?”, should be avoided as they often confuse the respondent.

- **Ask specific questions instead of general questions**

General questions are easier to misinterpret than specific questions.

Asking a series of specific questions gives more information and a deeper understanding of the respondent’s opinion than asking one general question.

For example, instead of asking “Do you support Party A?”, we can ask a series of questions:

- ▶ “Do you support the environmental policies that Party A introduced?”
- ▶ “Why did you choose Party A over Party B in the last election?”
- ▶ “Will you vote for Party A in the next election?”

- **Choose between structured versus unstructured questions**

Questions without any restrictions on the answer required are called **unstructured questions**. Unstructured questions are useful for exploratory surveys where the purpose of the investigation is to gauge opinions.

Questions with a set of answers for the respondent to choose from are called **structured questions**. The choices presented in structured questions can prompt respondents to remember things that they would have otherwise forgotten. An “Other” category which allows the respondent to input their own answer may be useful.

- **Keep the tone of questions neutral**

Questions containing opinions or emotive language are **biased** and can lead the respondent to answer in a certain way.

For example, the question “Do you support the dangerous practice of cycling without a helmet?” invites the respondent to answer “no”, since the question contains the judgement that riding without a helmet is dangerous.



- **Be aware of personal questions**

Questions that ask for sensitive information or opinions are called **personal questions**.

For example:

- ▶ “How much do you weigh?”
- ▶ “What is your gender?”
- ▶ “What is your religious affiliation?”
- ▶ “What is your sexual orientation?”

Such questions can make the respondent uncomfortable. They may not answer truthfully, or they may not answer at all.

If you *need* to ask a personal question, justify why the information is needed. Respondents are more likely to divulge personal information if they understand why it is needed.

EXERCISE 12C

1 Consider the question “Is your shirt red, blue, yellow, or white?”

- a State one problem with posing this as a structured question.
- b Rewrite the question as an unstructured question.
- c The colour of a person’s shirt can be *subjective* because it usually depends on individual interpretation. Explain how this might be a problem.



2 Consider the question “Do you have any allergies?”

- a List ways in which the question can be interpreted. Include any possible misinterpretations.
- b Rewrite the question so it is more specific.

3 Consider the question “Do you have any pets?”

- a List ways in which the question can be interpreted. Include any possible misinterpretations.
- b Rewrite the question so it is more specific.

4 The government has released a new proposal to move funding from education to health. A journalist wants to understand the public’s feelings about this proposal. She asks 100 people the question “Do you support the Government’s proposed cuts to education?”

- a Explain why this survey may produce a measurement error.
- b How could the question be worded so the public’s feelings about the proposal would be more accurately measured?

5 Consider the question “Where do you live?”

- a Explain why this question is likely to have a high non-response error.
- b List ways in which the question could be improved.

6 For each of the following survey questions:

- i Identify any problems with how the question is worded.
- ii Rewrite the question to fix these problems.
- a Have you or have you not been immunised against the infectious meningococcal disease?
- b Do you believe that climate change is a major issue or do you think that it is a topic thrown around by politicians to gain support?
- c Considering the fact that “fair trade cocoa” ensures that cocoa farmers are paid a minimum wage and helps prevent child labour, do you agree that fair trade certified chocolate should be more expensive than uncertified chocolate?

DISCUSSION

If a question provides choices for answers which use a rating scale, is it better to have an even or odd number of ratings?

ACTIVITY

Lily wants to survey students at her school about the school canteen.

What to do:

- 1 List various characteristics about the canteen that Lily might be interested in.
- 2 Decide which characteristics would be most relevant if:
 - a a new canteen was being constructed
 - b there was a new manager in charge of the canteen.
- 3 What sampling method(s) do you think would be most suitable for Lily? Explain your answer.
- 4 Write a complete survey for Lily, including questions about each of the characteristics listed in 1. Consider carefully your choices of unstructured and structured questions.

RESEARCH

PRECISE QUESTIONING

Precise questioning (PQ) is a question writing framework aimed at obtaining clear answers and communicating more efficiently, particularly in business settings. It was developed at Stanford University during the mid-1990s.

The seven main categories of questions in PQ are:

- | | | |
|-----------------------------|-------------------|-----------------|
| (1) Go/No go | (2) Clarification | (3) Assumptions |
| (4) Basic critical question | (5) Causes | (6) Effects |
| (7) Action | | |

What to do:

- 1 Research PQ and explain what each category of questions is meant to achieve. Give an example of each type of question in context.
- 2 Which categories of questions might be applicable when writing a questionnaire?
- 3 Do you think this framework could apply to statistical investigations? Discuss your answer.

D

TYPES OF DATA

In previous years you should have seen how variables can be described either as **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, macOS, 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 **discrete quantitative variable** or just **discrete variable** takes exact number values. It is usually a result of **counting**.

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

Some examples of continuous variables are:

- *the times taken to run a 100 m race:*
The variable would likely be between 9.5 and 25 seconds.
- *the distance of each hit in baseball:*
The variable could take values from 0 m to 100 m.

Example 3

Self Tutor

Classify each variable as categorical, discrete, or continuous:

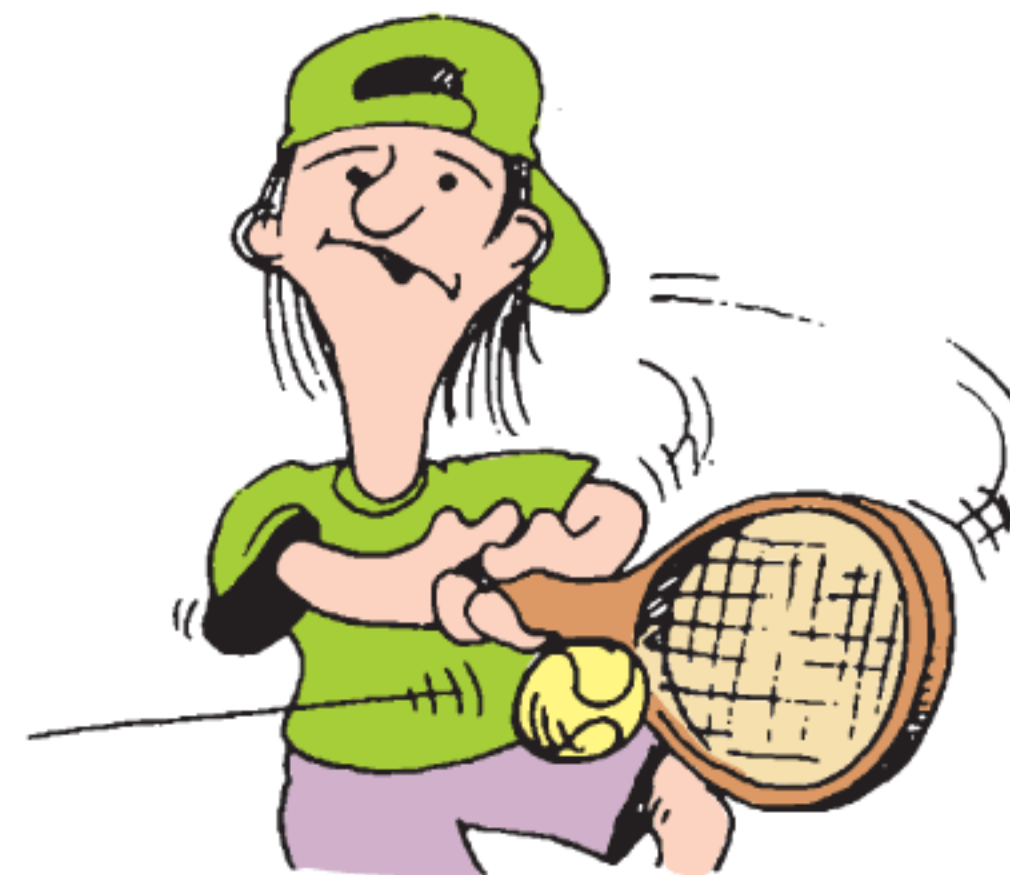
- a the number of heads when 3 coins are tossed
 - b the brand of toothpaste used by the students in a class
 - c the heights of a group of 15 year old children.
-
- a We count the number of heads. The result could be 0, 1, 2, or 3. It is a discrete variable.
 - b The variable describes the brands of toothpaste. It is a categorical variable.
 - c We measure the height of each child. 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 continuous variable.

EXERCISE 12D

- 1 Classify each variable as categorical, discrete, or continuous.
If the variable is categorical, list some possible categories.
If the variable is quantitative, suggest possible values or a range of values the variable may take.
- The number of brothers a person has.
 - The colours of lollies in a packet.
 - The time children spend brushing their teeth each day.
 - The heights of the trees in a garden.
 - The brand of car a person drives.
 - The number of petrol pumps at a service station.
 - The most popular holiday destinations.
 - The scores out of 10 in a diving competition.
 - The amount of water a person drinks each day.
 - The number of hours spent per week at work.
 - The average temperatures of various cities.
 - The items students ate for breakfast before coming to school.
 - The number of televisions in each house.

- 2 Consider the following statistics for a tennis player:

Name: Vance McFarland
Age: 28
Height: 191 cm
Country: Ireland
Tournament wins: 14
Average serving speed: 185 km h^{-1}
Ranking: 6
Career prize money: £3 720 000



Classify each variable as categorical, discrete, or continuous.

E**SIMPLE DISCRETE DATA****ORGANISING DISCRETE DATA**

One of the simplest ways to organise data is using a **tally and frequency table** or just **frequency table**.

For example, consider the data set:

1 3 1 2 4 2 4 1 5 3 1 3 2 2 4
 1 3 4 1 2 3 2 4 1 3 2 1 2 5 2

A **tally** is used to count the number of 1s, 2s, 3s, and so on. As we read the data from left to right, we place a vertical stroke in the tally column. We use |||| to represent 5 occurrences.

The **frequency** column summarises the number of occurrences of each particular data value.

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.

Value	Tally	Frequency (<i>f</i>)	Relative frequency
1		8	≈ 0.267
2		9	0.3
3		6	0.2
4		5	≈ 0.167
5		2	≈ 0.0667
	<i>Total</i>	30	

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

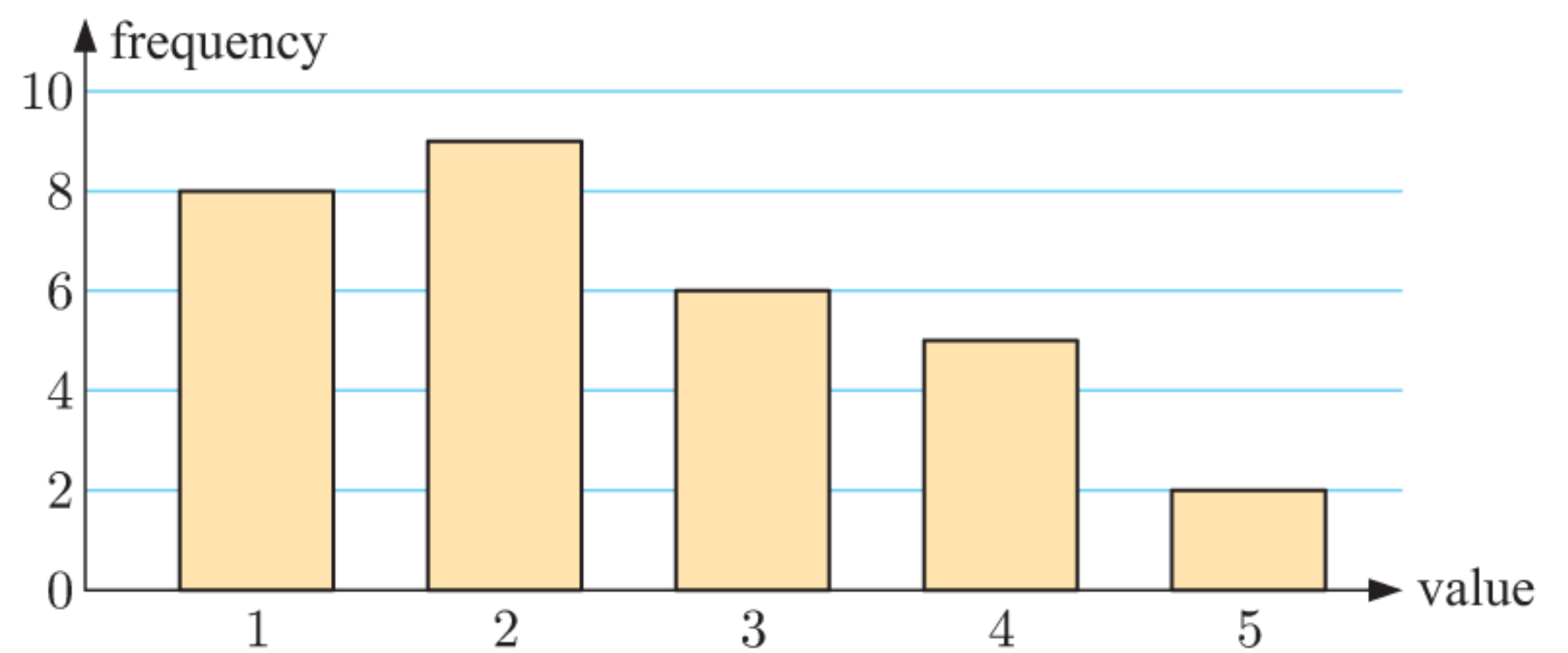


DISPLAYING DISCRETE DATA

Discrete data is displayed using a **column graph**. For this type of graph:

- The possible data values are placed on the horizontal axis.
- The frequency of data values is read from the vertical axis.
- The column widths are equal and the column height represents the frequency of the data value.
- There are gaps between columns to indicate the data is discrete.

A column graph for the data set on the previous page is shown alongside.

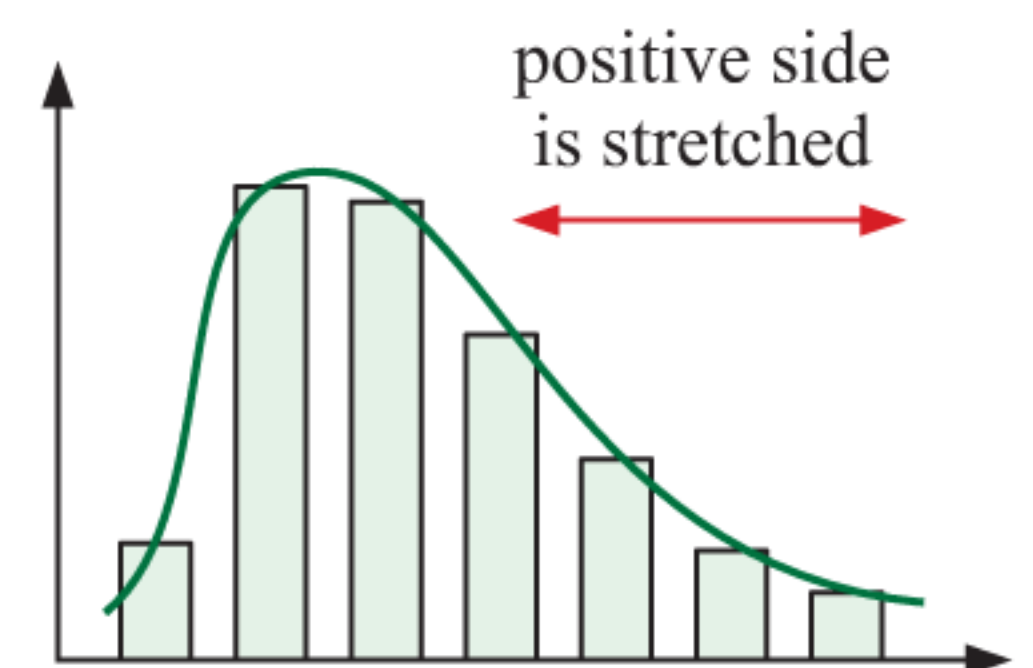
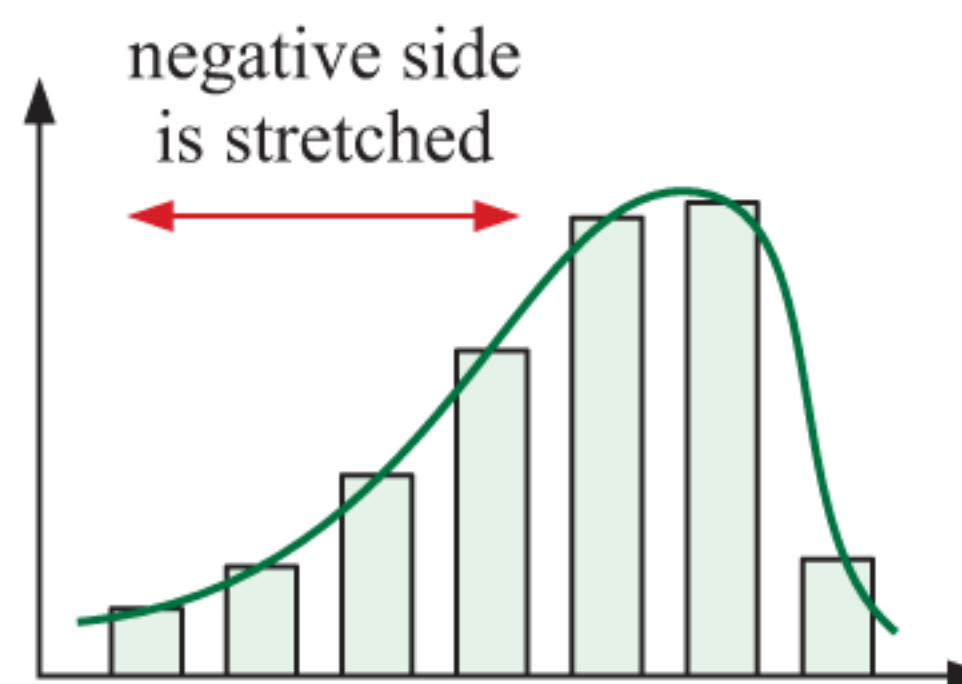
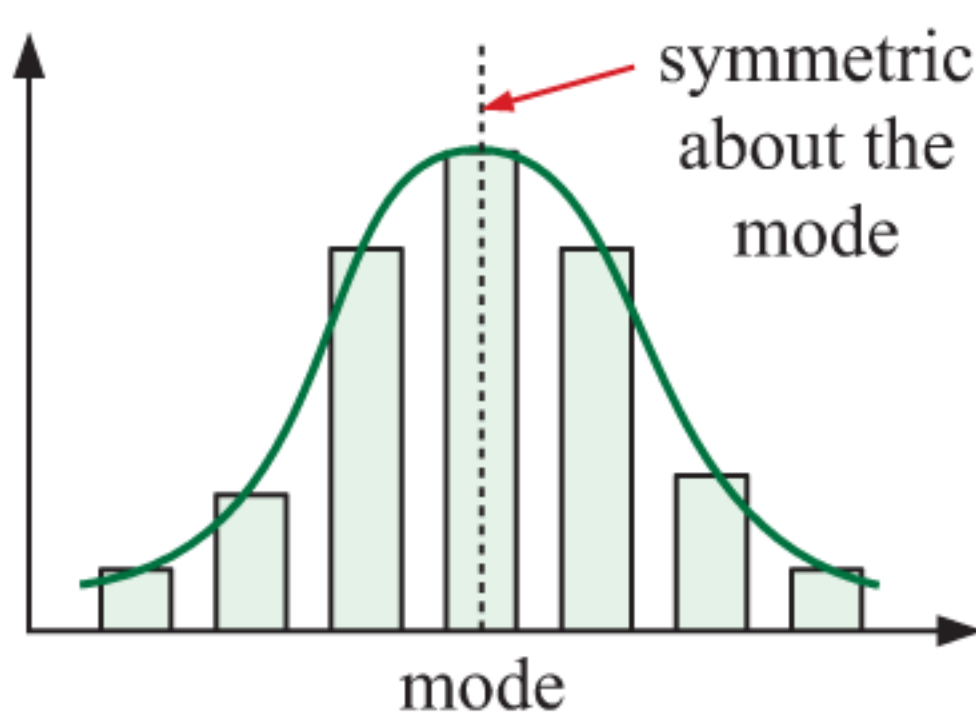


The **mode** of a data set is the most frequently occurring value. On a column graph, the mode will have the highest column. In this case the mode is 2.

DESCRIBING THE DISTRIBUTION OF A DATA SET

A column graph allows us to quickly observe the **distribution** or **shape** of the data set. We can describe the distribution as:

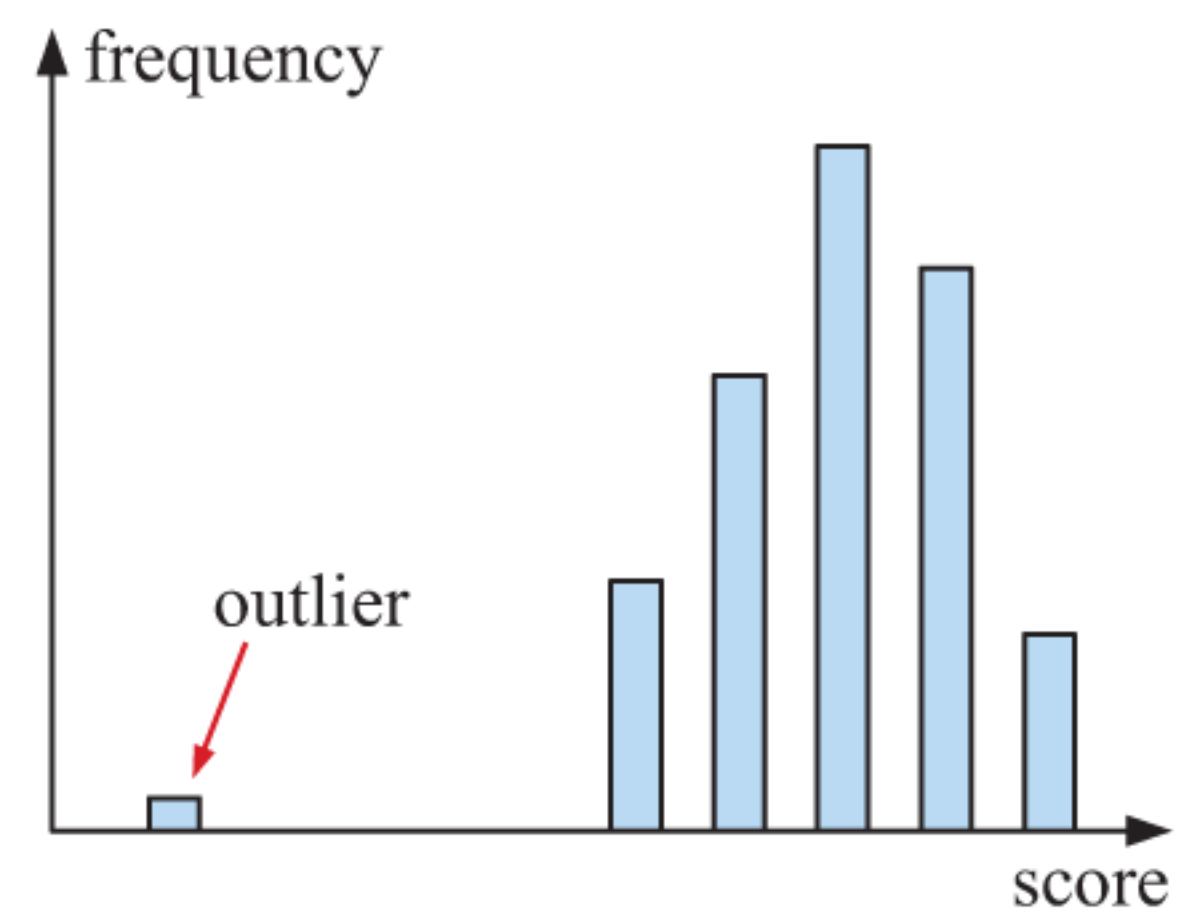
- **Symmetric**
- **Negatively skewed**
- **Positively skewed**



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.

If an outlier is a genuine piece of data, it should be retained for analysis. However, if it is found to be the result of an error in the data collection process, it should be removed from the data.

**Example 4****Self Tutor**

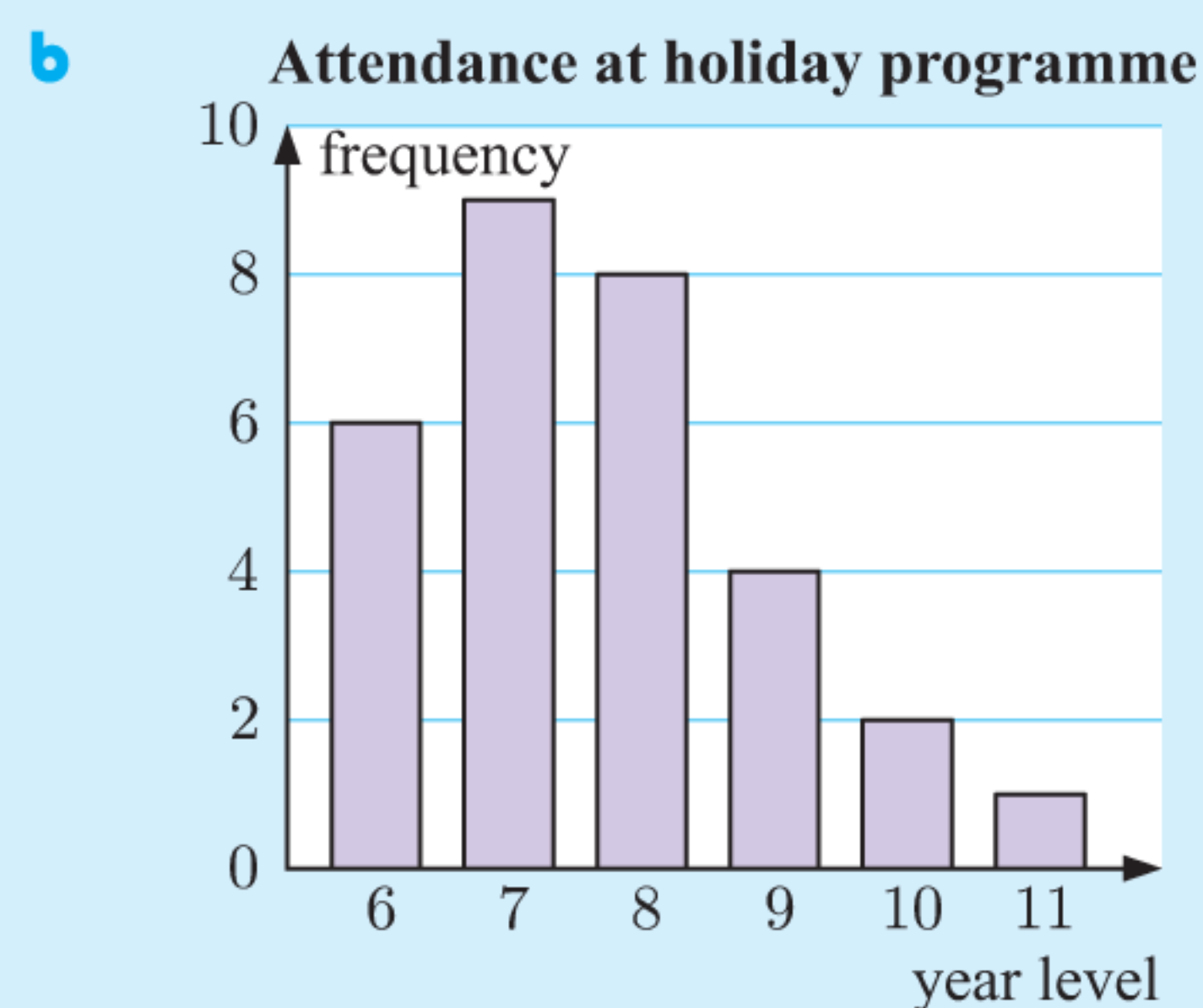
30 children attended a library holiday programme. Their year levels at school were:

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

- Record this information in a frequency table. Include a column for relative frequency.
- Construct a column graph to display the data.
- What is the modal year level of the children?
- Describe the shape of the distribution. Are there any outliers?
- What percentage of the children were in Year 8 or below?
- What percentage of the children were above Year 9?

a

Year level	Tally	Frequency	Relative frequency
6		6	0.2
7		9	0.3
8		8	≈ 0.267
9		4	≈ 0.133
10		2	≈ 0.067
11		1	≈ 0.033
<i>Total</i>		30	



- The modal year level is Year 7.
- The distribution of children's year levels is positively skewed. There are no outliers.
- $\frac{6 + 9 + 8}{30} \times 100\% \approx 76.7\%$ of the children 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.
- $\frac{2 + 1}{30} \times 100\% = 10\%$ of the children were above Year 9.
or $0.067 + 0.033 = 0.1 \quad \therefore 10\%$ were above Year 9.

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



EXERCISE 12E

- 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 misbehave and as a result are placed on detention. The studious school master records the number of students on detention each week throughout the year:

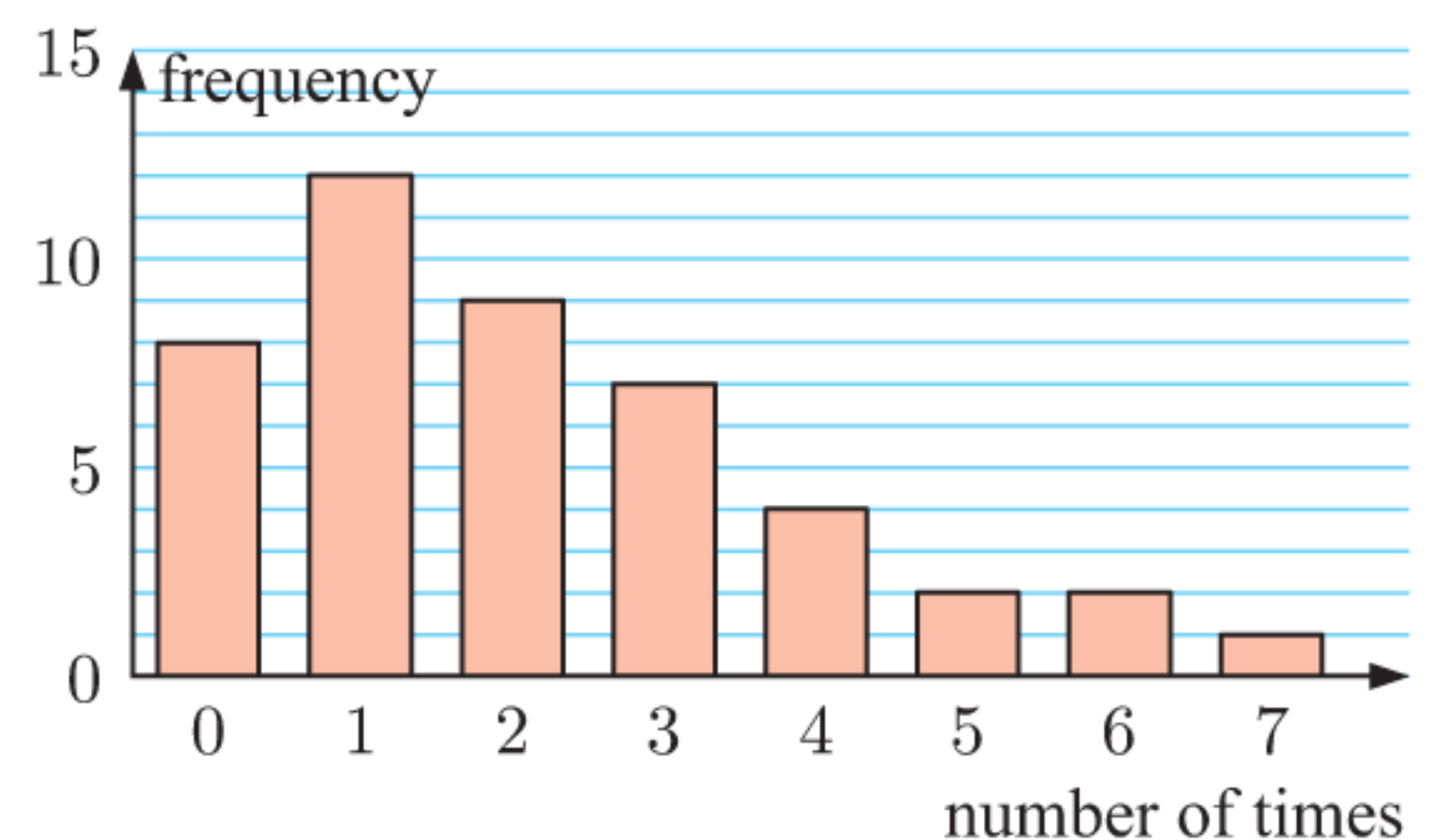
0 2 1 5 0 1 4 2 3 1 4 3 0 2 9 2 1 5 0 3
6 4 2 1 5 1 0 2 1 4 3 1 2 0 4 3 2 1 2 3

- 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 Each time Joan visits the cinema, she records the number of previews for other films which are shown before the feature. She has obtained these results:

2 4 3 1 3 2 3 4 2 5 2 3 4 3 6 5 4
3 3 6 3 4 6 4 3 1 4 2 5 4 3 5 4 5

- 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 On what percentage of occasions were at least 3 previews shown?
- 4 A random sample of people were asked “How many times did you eat out 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 out at all last week?
- d What percentage of people surveyed ate out more than three times last week?
- e Describe the distribution of the data.



F

GROUPED 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 weekdays 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 distribution of the data. It is more meaningful to **group** the data into **class intervals** and then compare the frequencies of the classes.

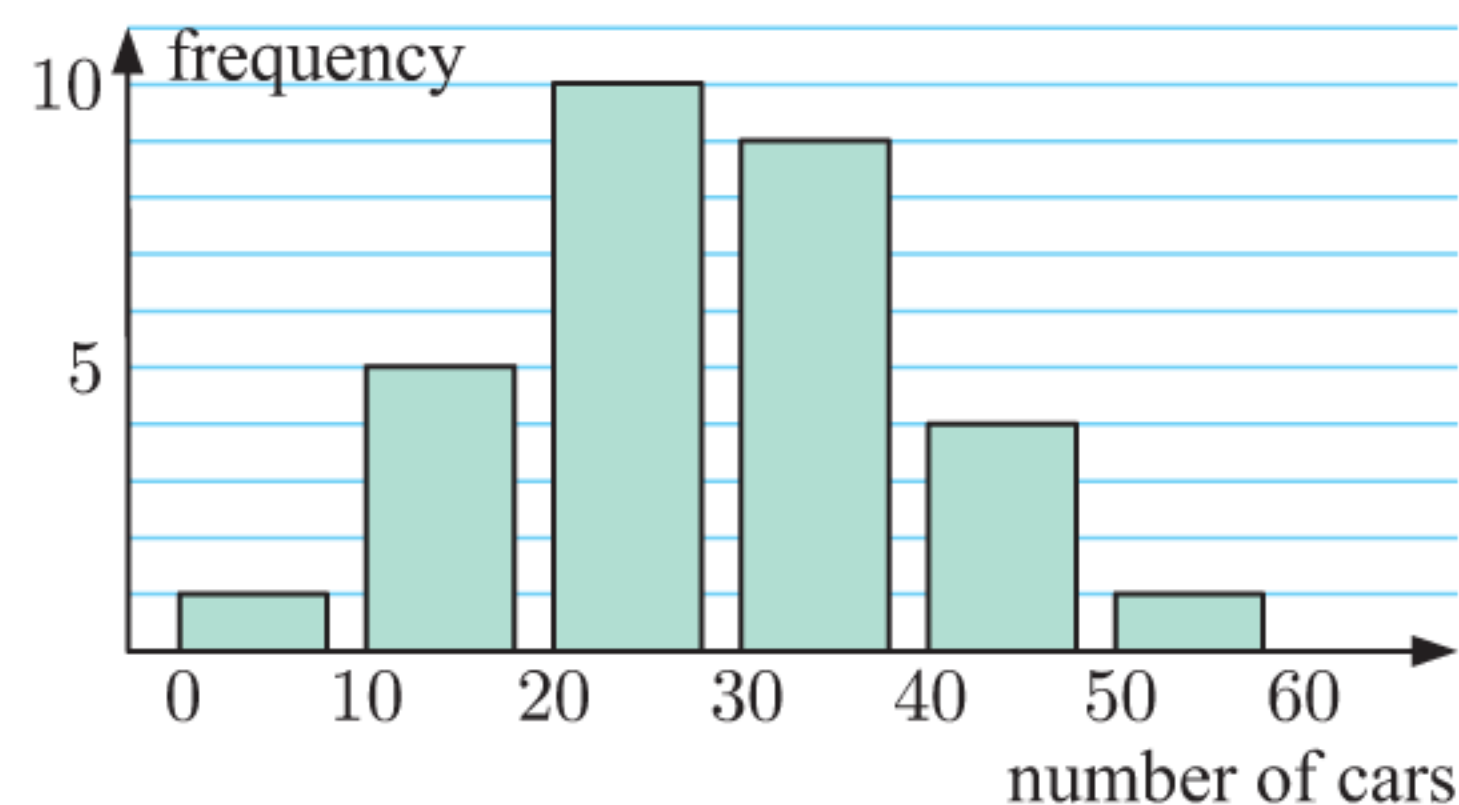
For the data given we use class intervals of width 10. The frequency table for the grouped data is shown alongside.

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		5
20 to 29		10
30 to 39		9
40 to 49		4
50 to 59		1
<i>Total</i>		30

We construct **column graphs** for grouped discrete data in the same way as for simple data.

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

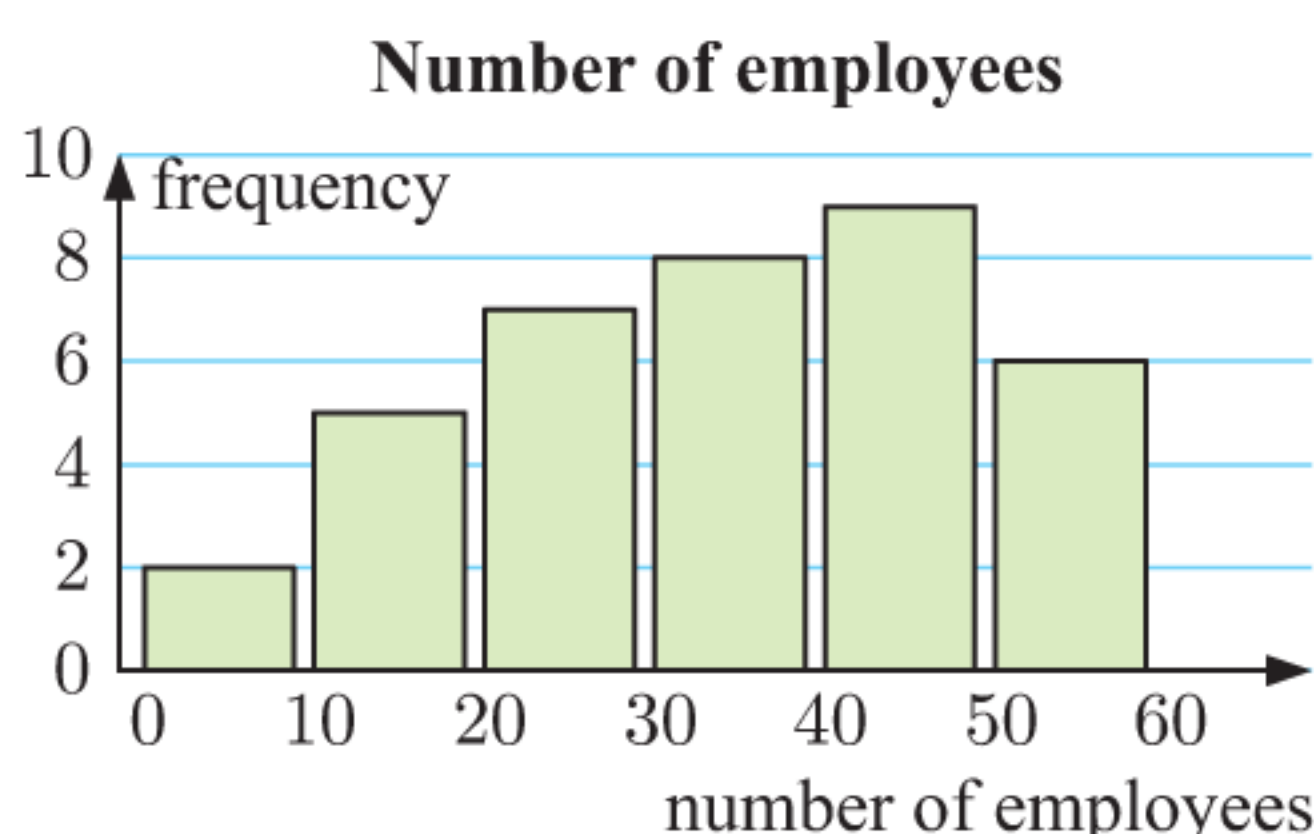


DISCUSSION

- 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 lowest and highest values are?

EXERCISE 12F

- 1 A selection of businesses were asked how many employees they had. The results are displayed on this column graph.



- How many businesses were surveyed?
- Find the modal class.
- Describe the distribution of the data.
- What percentage of businesses surveyed had less than 30 employees?
- Can you determine the highest number of employees a business had?

- 2 Arthur catches the train to school from a suburban 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.
- 3 A city council is interested in the number of houses in each street of a suburb, because it intends to place collection bins for unwanted clothing. The data they find is:

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?

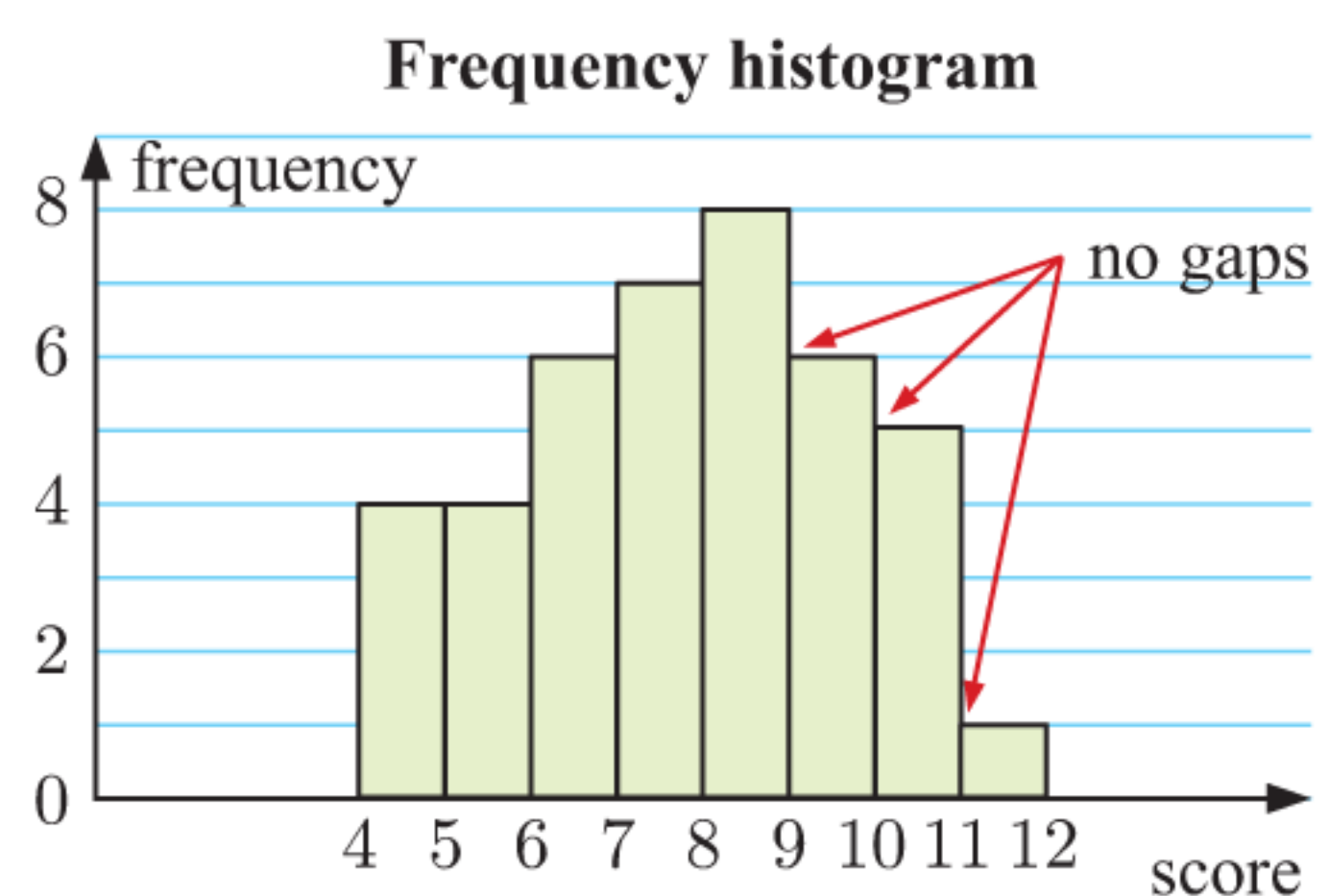
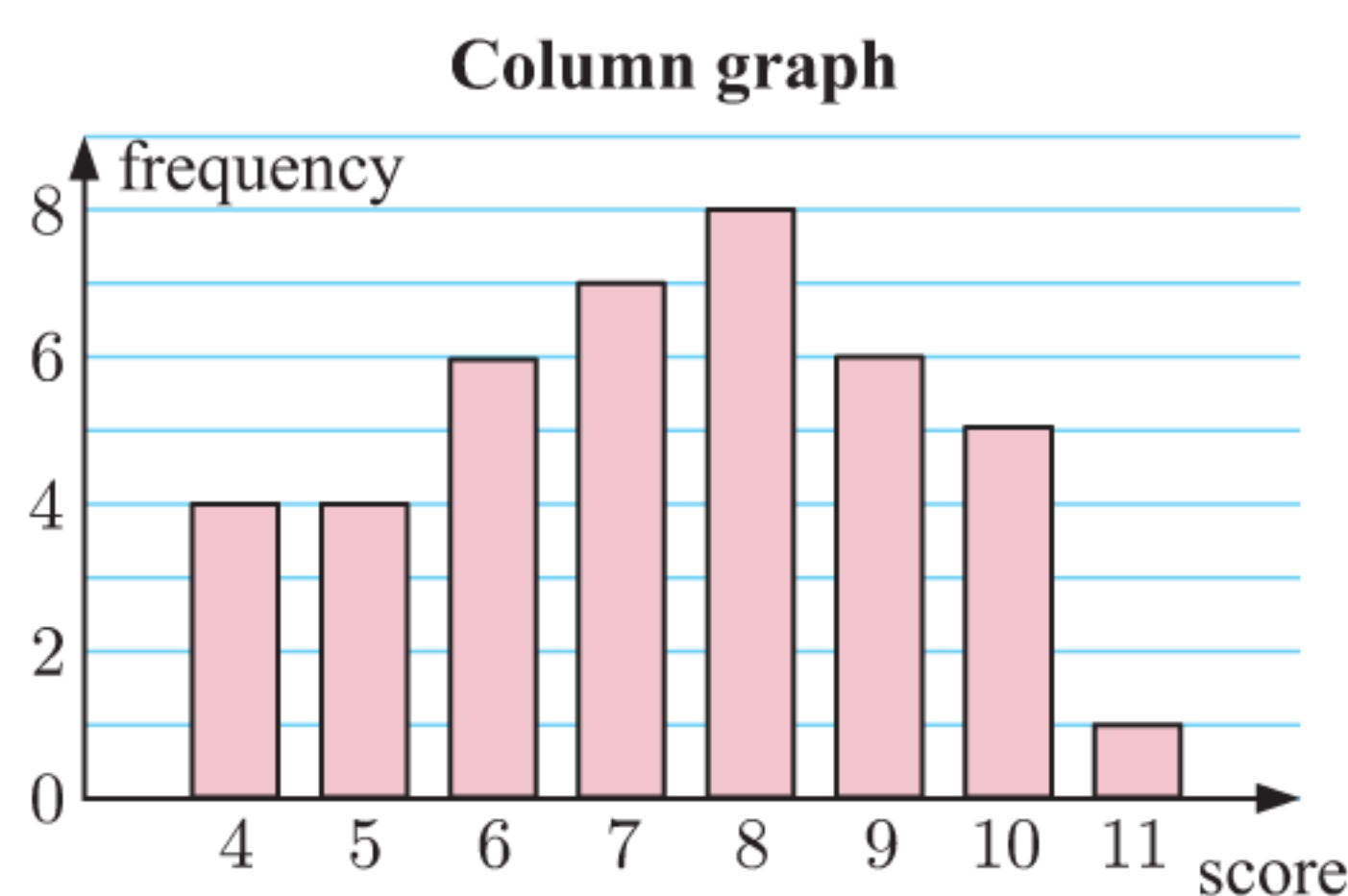
G CONTINUOUS DATA

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 a particular value. 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 continuous data. This is similar to a column graph, but the “columns” are joined together and the values at the edges of each column indicate the boundaries of that class interval.

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



INVESTIGATION

CHOOSING CLASS INTERVALS

When dividing data values into intervals, the choice of how many intervals to use is important. It affects not only the width of each class interval, but how much detail of the distribution is seen on the histogram.

What to do:

- 1 Click on the icon to access a demonstration which draws histograms of data sets with different sizes and distributions.
 - a Select the symmetrical data set with $n = 1000$ data values. Use the slider to vary the number of intervals used in the histogram.
 - i Comment on what happens to the *shape* of the histogram.
 - ii Are there features of the data that can only be seen when there are many class intervals?
 - iii When there are many class intervals, is the frequency axis necessarily useful?
 - b Repeat your investigation in **a** for other values of n . Try $n = 100$, 10 000, and 100 000. Record your observations.
 - c For each value $n = 100$, 1000, 10 000, and 100 000, try using $\approx \sqrt{n}$ class intervals. Discuss whether you think this is an appropriate number.
- 2 Experiment with the other distribution types. If the distribution is not symmetric, do you need more or fewer class intervals?

**Example 5****Self Tutor**

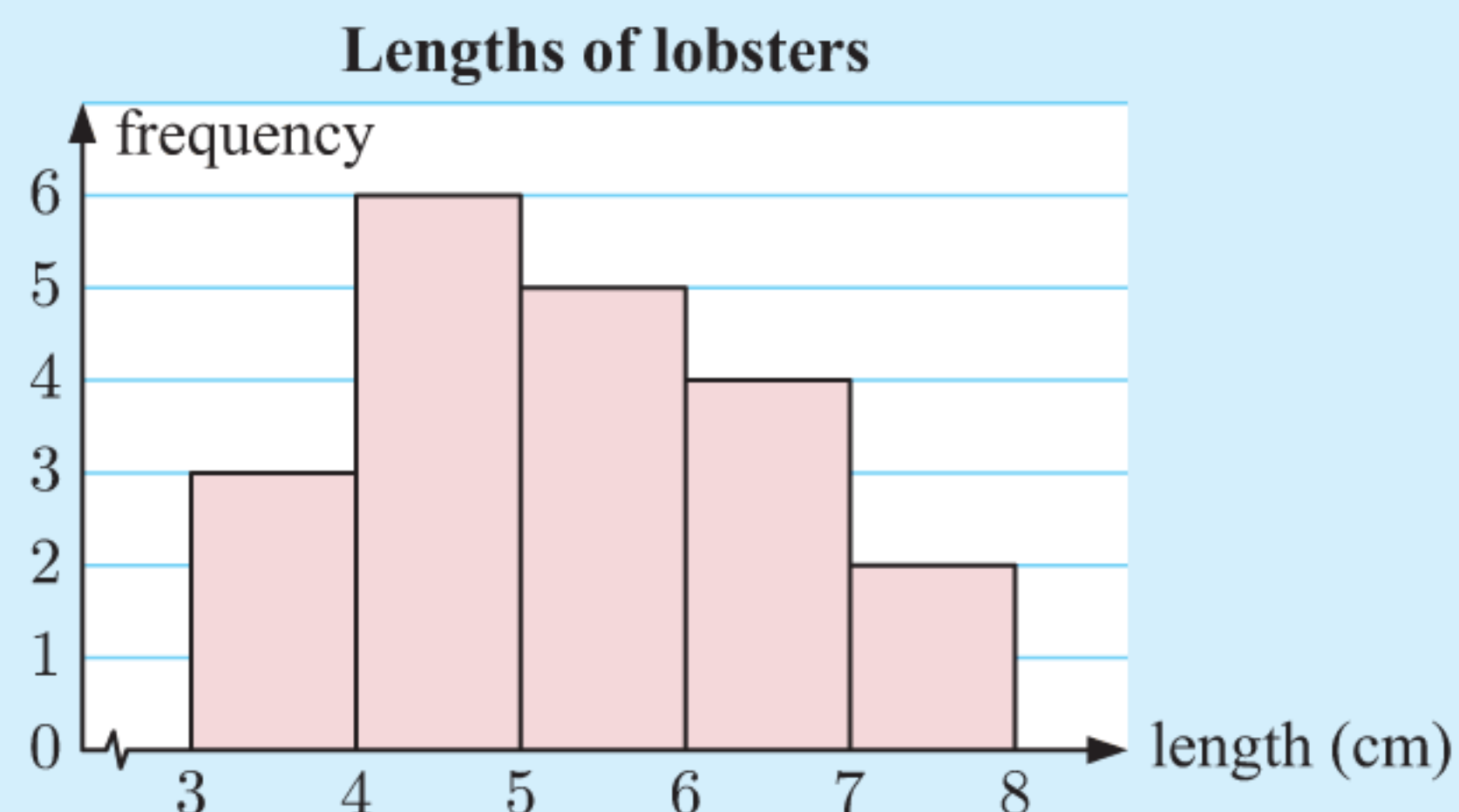
A sample of 20 juvenile lobsters was randomly selected from a tank containing several hundred. The length of each lobster is recorded in cm alongside.

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

- a Organise the data using a frequency table, and hence graph the data.
- b State the modal class and explain what this means.
- c Describe the distribution of the data.

- a The variable *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
$3 \leq l < 4$	3
$4 \leq l < 5$	6
$5 \leq l < 6$	5
$6 \leq l < 7$	4
$7 \leq l < 8$	2



- b The modal class $4 \leq l < 5$ occurs most frequently. More lobsters have lengths in this interval than in any other interval.
- c The distribution is positively skewed with no outliers.

EXERCISE 12G

1 A frequency table for the heights of a volleyball squad is given alongside.

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

- 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.

2 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 their 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 $0 \leq t < 10$, $10 \leq t < 20$, ..., $40 \leq t < 50$.
- c** Hence draw a histogram to display the data.
- d** Describe the distribution of the data.
- e** What is the modal travelling time?

3 A group of 25 junior athletes participated in a javelin 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?

4 A horticulturalist 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
$300 \leq h < 325$	12
$325 \leq h < 350$	18
$350 \leq h < 375$	42
$375 \leq h < 400$	28
$400 \leq h < 425$	14
$425 \leq h < 450$	6

- a** Display the data on a frequency histogram.
- b** How many of the seedlings are 400 mm or higher?
- c** What percentage of the seedlings are between 350 mm and 400 mm high?
- d** In total there are 1462 seedlings in the nursery. Estimate the number of seedlings which measure:
 - i** less than 400 mm
 - ii** between 375 and 425 mm.

- 5 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

- Choose suitable class intervals to group the data.
- Organise the data in a frequency table.
- Draw a frequency histogram to display the data.
- What percentage of the rats weigh less than 200 grams?

REVIEW SET 12A

- 1 Andrew is interested in the cultural background of the students at his school. He puts together a survey which he hands out to students in his Italian class.
- Explain why Andrew's sample may be biased.
 - Suggest an alternative sampling method that Andrew can use so that his results will be more representative of his population of interest.

- 2 A golf club has 1800 members with ages shown alongside. A member survey is to be undertaken to determine the proportion of members who are in favour of changes to dress regulations.

Age range	Members
under 18	257
18 - 39	421
40 - 54	632
55 - 70	356
over 70	134

- Explain why the golf club would not question all members on the proposed changes to dress regulations.
- If a sample size of 350 is used, how many of each age group will be surveyed?

- 3 Classify each variable as categorical, discrete, or continuous:

- the number of pages in a daily newspaper
- the maximum daily temperature in a city
- the manufacturer of a television
- a person's favourite flavour of ice cream
- the position taken by a player on a lacrosse field
- the time it takes to run one kilometre
- the length of a person's feet
- a person's shoe size
- the cost of a bicycle.



- 4 On a Saturday night, a team of police officers set up a drug and alcohol testing station to test drivers leaving the centre of town on a major road.
- What type of sampling method is this?
 - Do you think the sample will be biased? If so, do you think it is *sensible* for it to be biased? Explain your answer.
- 5 Consider the question "Are you healthy?"
- List ways in which the question can be interpreted. Include any possible misinterpretations.
 - Rewrite the question so it is more specific.

- 5 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

- Choose suitable class intervals to group the data.
- Organise the data in a frequency table.
- Draw a frequency histogram to display the data.
- What percentage of the rats weigh less than 200 grams?

REVIEW SET 12A

- 1 Andrew is interested in the cultural background of the students at his school. He puts together a survey which he hands out to students in his Italian class.
- Explain why Andrew's sample may be biased.
 - Suggest an alternative sampling method that Andrew can use so that his results will be more representative of his population of interest.

- 2 A golf club has 1800 members with ages shown alongside. A member survey is to be undertaken to determine the proportion of members who are in favour of changes to dress regulations.

Age range	Members
under 18	257
18 - 39	421
40 - 54	632
55 - 70	356
over 70	134

- Explain why the golf club would not question all members on the proposed changes to dress regulations.
- If a sample size of 350 is used, how many of each age group will be surveyed?

- 3 Classify each variable as categorical, discrete, or continuous:

- the number of pages in a daily newspaper
- the maximum daily temperature in a city
- the manufacturer of a television
- a person's favourite flavour of ice cream
- the position taken by a player on a lacrosse field
- the time it takes to run one kilometre
- the length of a person's feet
- a person's shoe size
- the cost of a bicycle.



- 4 On a Saturday night, a team of police officers set up a drug and alcohol testing station to test drivers leaving the centre of town on a major road.
- What type of sampling method is this?
 - Do you think the sample will be biased? If so, do you think it is *sensible* for it to be biased? Explain your answer.
- 5 Consider the question "Are you healthy?"
- List ways in which the question can be interpreted. Include any possible misinterpretations.
 - Rewrite the question so it is more specific.

- 3** Petra emailed a questionnaire to her teacher colleagues about general student behaviour in their classes.
- Explain why Petra's questionnaire may produce a high non-response error.
 - Of the 20 teachers who were emailed the questionnaire, 10 responded. Petra decides to use these 10 responses as her sample. Explain why Petra is likely to encounter a coverage error.
- 4** Rewrite the question "How did you learn about our services?" as a structured question. Include the options that respondents should choose from.
- 5** Consider the question "Were you a naughty child?"
- Identify any problems with how the question is worded.
 - Rewrite the question to address these problems.
- 6** The winning margins in 100 rugby games were recorded as follows:

<i>Margin (points)</i>	1 - 10	11 - 20	21 - 30	31 - 40	41 - 50
<i>Frequency</i>	13	35	27	18	7

Draw a column graph to present this information.

- 7** The Dzungarian or Przewalski's horse is an endangered species native to the Mongolian steppes. The adult horses in an established breeding program are weighed in kg. The results are:

274 298 302 316 296 279 325
 303 286 318 286 325 306 303
 261 315 326 293 281



- Explain why the *mass of a horse*, m kg, is a continuous quantitative variable.
 - Organise the data in a frequency table using the intervals $260 \leq m < 270$, $270 \leq m < 280$, ..., $320 \leq m < 330$.
 - Identify the modal class.
 - Draw a histogram to display the data.
 - Hence describe the distribution of the data.
- 8** 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
- Is the data discrete or continuous?
 - Organise the data using a frequency table.
 - Draw an appropriate graph to display the data.
 - Describe the distribution of the data.

Chapter

13

Statistics

Contents:

- A** Measuring the centre of data
- B** Choosing the appropriate measure
- C** Using frequency tables
- D** Grouped data
- E** Measuring the spread of data
- F** Box and whisker diagrams
- G** Outliers
- H** Parallel box and whisker diagrams
- I** Cumulative frequency graphs
- J** Variance and standard deviation

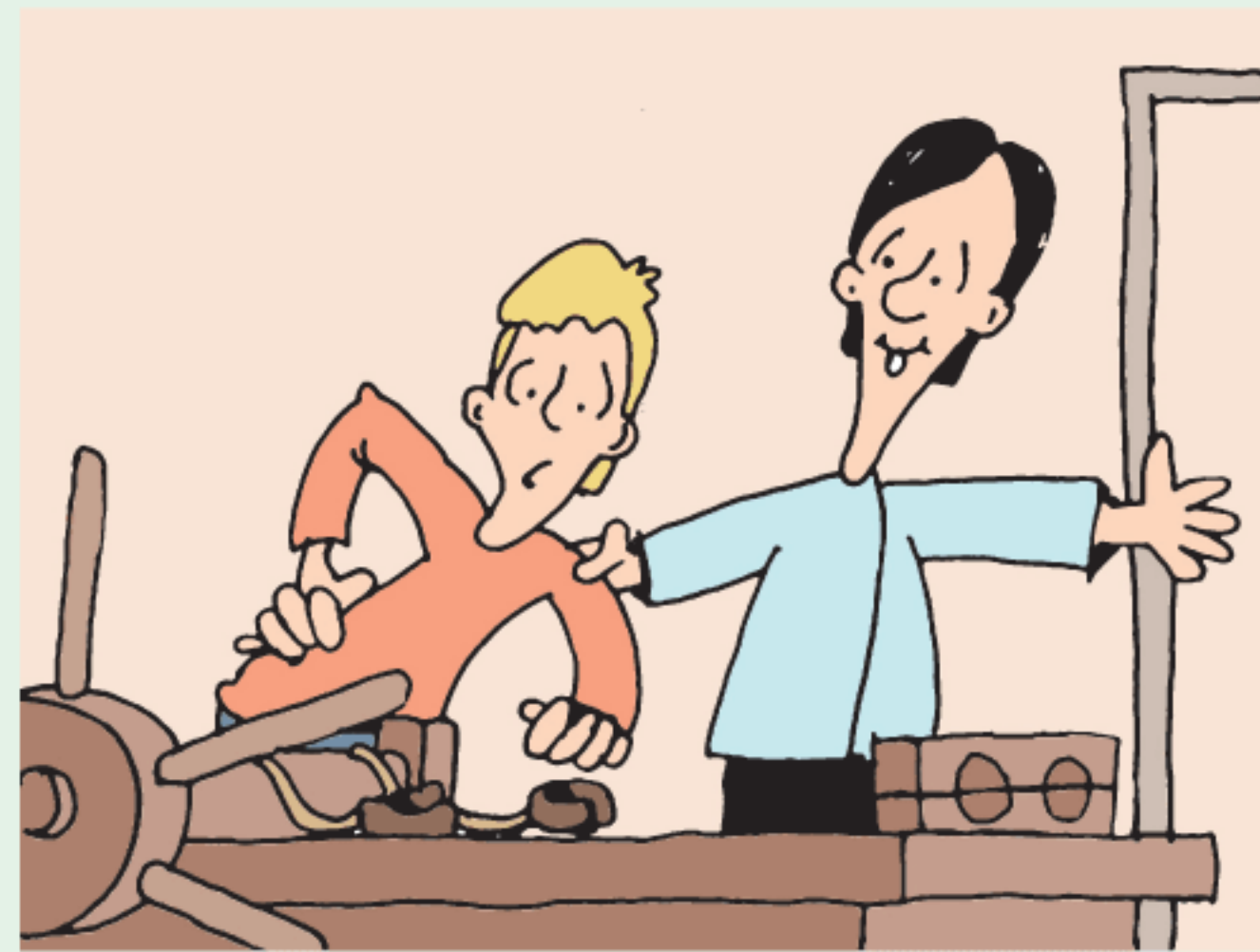


OPENING PROBLEM

Nick believes he has devised a series of stretches which can help relieve back pain. He invites people with back pain to perform the stretches for several weeks.

The participants rate their level of back pain on a scale of 1 to 10 (10 being the greatest) before and after the experiment:

<i>Before:</i>	7	9	5	6	9	7	10	6	8	9
	8	7	9	8	10	4	8	6	7	8
<i>After:</i>	4	7	6	3	8	5	9	4	5	8
	7	6	5	4	7	2	5	3	4	6



Things to think about:

- What statistics can we calculate to measure the *centre* of each data set?
- How can we use a graph to make a visual comparison between the data sets?
- Do you believe that Nick's stretching exercises reduce back pain? Explain your answer.

In the previous Chapter, we looked at how data can be collected, organised, and displayed. By looking at appropriate graphs, we can get an idea of a data set's **distribution**.

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

However, whatever statistics we calculate, it is essential to view and interpret them in the context of what we are studying.

A

MEASURING THE CENTRE OF DATA

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

In the previous Chapter we saw that:

- For discrete data, the **mode** is the most frequently occurring value in the data set.
- For continuous data, we cannot talk about a mode in this way because no two data values will be *exactly* equal. Instead we talk about a **modal class**, which is the class or group that has the highest frequency.

If a data set has two values which both occur most frequently, we say it is **bimodal**.

If a data set has three or more values which all occur most frequently, the mode is not an appropriate measure of centre to use.

THE MEAN

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

For the data set $\{x_1, x_2, x_3, \dots, x_n\}$,

$$\begin{aligned} \text{mean} &= \frac{\text{sum of all data values}}{\text{the number of data values}} \\ &= \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} \\ &= \frac{\sum_{i=1}^n x_i}{n} \end{aligned}$$

We use \bar{x} to represent the mean of a **sample**, and μ to represent the mean of a **population**.

In many cases we do not have data from all of the members of a population, so the exact value of μ is unknown. Instead we collect data from a sample of the population, and use the mean of the sample \bar{x} as an approximation for μ .

μ is the Greek letter “mu” which we pronounce as “mew”.



THE MEDIAN

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

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

The median splits the data in halves. Half of the data values 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 values, the median is one of the original data values.

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

If there are n data values listed in order from smallest to largest, the median is the $\left(\frac{n+1}{2}\right)$ th data value.

For example:

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

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



Example 1**Self Tutor**

The numbers of faulty products returned to an electrical goods store each day over a 21 day period are:

3 4 4 9 8 8 6 4 7 9 1 3 5 3 5 9 8 6 3 7 1

- a** For this data set, find:
- i** the mean
 - ii** the median
 - iii** the mode.
- b** On the 22nd day there were 9 faulty products returned. How does this affect the measures of the centre?

a i mean = $\frac{3 + 4 + 4 + \dots + 3 + 7 + 1}{21}$ ← sum of all the data values
 ← 21 data values
 $= \frac{113}{21}$
 ≈ 5.38 faulty products

ii As $n = 21$, $\frac{n+1}{2} = 11$

The ordered data set is: ~~1 1 3 3 3 3 4 4 4 5 5 6 6 7 7 8 8 8 9 9 9~~
 ↑
 11th value

∴ median = 5 faulty products

iii 3 is the data value which occurs most often, so the mode is 3 faulty products.

- b** We expect the mean to increase since the new data value is greater than the old mean.

In fact, the new mean = $\frac{113 + 9}{22} = \frac{122}{22} \approx 5.55$ faulty products.

Since $n = 22$, $\frac{n+1}{2} = 11.5$

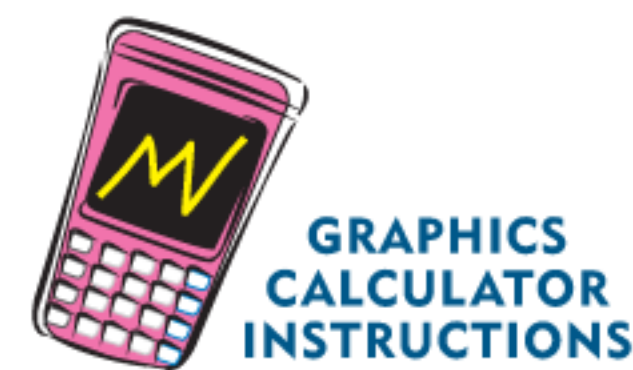
The new ordered data set is:

~~1 1 3 3 3 3 4 4 4 5 5 6 6 7 7 8 8 8 9 9 9~~
 { 5 6 }
 two middle data values

∴ the new median = $\frac{5+6}{2} = 5.5$ faulty products.

The new data set has two modes which are 3 and 9 faulty products.

You can use your **graphics calculator** or the **statistics package** to find measures of centre.

**EXERCISE 13A**

- 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** For each data set, find the: **i** mean **ii** median **iii** mode.

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

Check your answers using technology.

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

4 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?

5 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 Comment on your answers to a and b.

6 An Indian dessert shop keeps a record of how many motichoor ladoo and malai jamun they sell each day for a month:

Motichoor ladoo								Malai jamun							
62	76	55	65	49	78	71	82	37	52	71	59	63	47	56	68
79	47	60	72	58	82	76	67	43	67	38	73	54	55	61	49
50	61	70	85	77	69	48	74	50	48	53	39	45	60	46	51
63	56	81	75	63	74	54		38	57	41	72	50	44	76	

- a Find the:
 - i mean number of motichoor ladoo and malai jamun sold
 - ii median number of motichoor ladoo and malai jamun sold.
- b Which 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							Tram					
30	43	40	53	70	50	63	58	68	43	45	70	79
41	38	21	28	23	43	48	38	23	30	22	63	73
20	26	35	48	41	33		25	35	60	53		

- a Use technology to calculate the mean and median number of passengers for both the *Bus* and *Tram* data.
 - b Which method of transport do you think is more popular? Explain your answer.
- 8 A basketball team scored 43, 55, 41, and 37 points in their first four matches.
- a Find the mean number of points scored for these four matches.
 - b What score does the team need to shoot in their next match to maintain the same mean score?
 - c The team scores only 25 points in the fifth match.
 - i Will this increase or decrease their overall mean score? Explain your answer.
 - ii Find the mean number of points scored for the five matches.

Example 2**Self Tutor**

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

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

$$\therefore \text{sum of masses} = 53.7 \times 6$$

$$\therefore \text{the total mass} = 322.2 \text{ kg}$$

- 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 Given $\bar{x} = 11.6$ and $n = 10$, calculate $\sum_{i=1}^{10} x_i$.
- 11 Towards the end of a season, a netballer had played 14 matches and scored an average of 16.5 goals per game. In the final two matches of the season she scored 21 goals and 24 goals. Find the netballer's average for the whole season.
- 12 Find x if 5, 9, 11, 12, 13, 14, 17, and x have a mean of 12.
- 13 Find a if 3, 0, a , a , 4, a , 6, a , and 3 have a mean of 4.
- 14 Over the entire assessment period, Aruna averaged 35 out of a possible 40 marks for her Mathematics 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.

INVESTIGATION 1**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 centre.

What to do:

- Consider the set of data: 4, 5, 6, 6, 6, 7, 7, 8, 9, 10. Calculate:
 - the mean
 - the mode
 - the median.
- Suppose we 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:
 - the mean
 - the mode
 - the median.
- Comment on the effect that the extreme value has on:
 - the mean
 - the mode
 - the median.
- Which of the three measures of centre is most affected by the inclusion of an outlier?
- Discuss situations with your class when it would *not* be appropriate to use a particular measure of centre of a data set.

B**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.

<i>Statistic</i>	<i>Properties</i>
Mode	<ul style="list-style-type: none"> • gives the most usual value • only takes common values into account • not affected by extreme values
Mean	<ul style="list-style-type: none"> • commonly used and easy to understand • takes all values into account • affected by extreme values
Median	<ul style="list-style-type: none"> • 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 useful to know, since it probably will not be an actual shoe size. However, 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. In this case 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 13B

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

\$346 400, \$327 600, \$411 000, \$392 500, \$456 400,
\$332 400, \$348 000, \$329 500, \$331 400, \$362 500

- Calculate the mean and median selling prices. Comment on your results.
- Which measure would you use if you were:
 - a vendor wanting to sell your house
 - looking to buy a house in the district?

- 2** The annual salaries of ten office workers are:

\$33 000, \$56 000, \$33 000, \$48 000, \$34 000,
\$33 000, \$33 000, \$48 000, \$33 000, \$42 000

- Find the mode, mean, and median salaries of this group.
- Explain why the mode is an unsatisfactory measure of the centre in this case.
- Is the median a satisfactory measure of the centre 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

- Use technology to find the mean, median, and mode of the data.
- Explain why the median is not the most suitable measure of centre for this set of data.
- Explain why the mode is not the most suitable measure of centre for this set of data.
- Identify the outliers in this data set.
- The outliers are genuine pieces of data and not the result of recording errors. Should they be removed before calculating statistics?

4 Esmé runs a day-tour business in Amsterdam. She wants to offer a “family package” that includes the charges for two adults and their children. To investigate the number of children she should include in the package, she asks 30 randomly selected customers with children how many children they have. Their responses are:

2 2 2 3 4 1 1 2 1 1 1 2 2 3 4
1 4 4 2 3 1 1 1 2 1 1 2 2 3 2

- Calculate the mean, median, and modal number of children per family.
- Is the mode a useful statistic in this case?
- Suggest how many children Esmé should include in the package, giving reasons for your answer.

THEORY OF KNOWLEDGE

We have seen that the mean, the median, and mode are all statistics that give an *indication* of a data set’s centre. The actual things that they measure are quite different!

- The mode is the value with the highest frequency. It is a measure of centre in terms of *frequency*.
- The median divides the data into halves. It is a measure of centre in terms of *proportion*.
- The mean is the arithmetic average. It can be thought of as the “balancing point” of the data set’s distribution.

Other less commonly used measures for a data set $\{x_1, x_2, \dots, x_n\}$ include the:

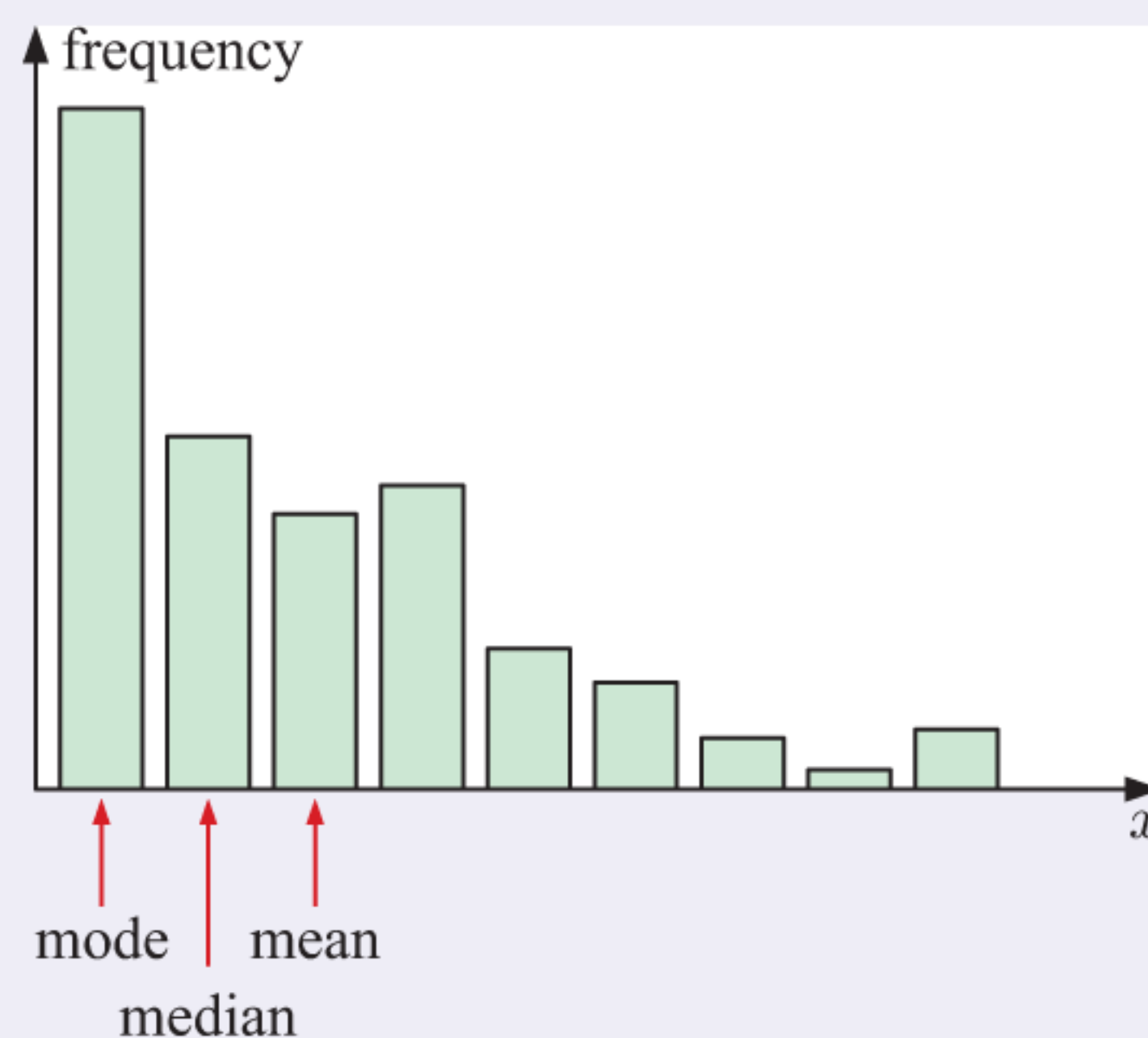
- geometric mean** = $\sqrt[n]{x_1 \times x_2 \times \dots \times x_n}$
- mid-range value** = $\frac{\text{maximum} + \text{minimum}}{2}$.

We have seen that the most appropriate measure of centre will depend on what we are investigating. In a way, we change how the “centre” of a data set is defined to suit the purpose of our investigation.

1 When we have data that is heavily skewed, the mode will be on the far left or far right on its column graph.

- Does the mode give an accurate indication of a data set’s centre in these cases?
- Is the relationship between mode and the centre of a data set purely coincidental?

2 For what kinds of data sets would the geometric mean and the mid-range value be useful?



- 3 How would *you* define the “centre” of a data set?
- 4 What makes a measure of centre objectively “better” than another measure?
- 5 Is there a *canonical* measure of centre, which means a measure of centre that is “better” than any other in all cases?

C

USING FREQUENCY TABLES

We have already seen how to organise data into a **frequency table** like the one alongside.

The mode of the data is found directly from the *Frequency* column.

Value	Frequency
3	1
4	1
5	3
6	7
7	15
8	8
9	5

mode →

THE MEAN

Adding a “Product” column to the table helps to add the data values.

For example, the value 7 occurs 15 times, and these add to $15 \times 7 = 105$.

Value (x)	Frequency (f)	Product (xf)
3	1	$3 \times 1 = 3$
4	1	$4 \times 1 = 4$
5	3	$5 \times 3 = 15$
6	7	$6 \times 7 = 42$
7	15	$7 \times 15 = 105$
8	8	$8 \times 8 = 64$
9	5	$9 \times 5 = 45$
<i>Total</i>	$\sum f = 40$	$\sum xf = 278$

Since the mean = $\frac{\text{sum of all data values}}{\text{the number of data values}}$, we find

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \dots + x_k f_k}{f_1 + f_2 + f_3 + \dots + f_k}$$

where k is the number of *different* values in the data.

$$\therefore \bar{x} = \frac{\sum_{j=1}^k x_j f_j}{\sum_{j=1}^k f_j} \quad \text{which we often abbreviate as } \frac{\sum xf}{\sum f}.$$

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

THE MEDIAN

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

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

We can see that the 20th and 21st ordered data values are both 7s.

$$\therefore \text{the median} = \frac{7+7}{2} = 7$$

Value	Frequency	Cumulative frequency
3	1	1 ← one number is 3
4	1	2 ← two numbers are 4 or less
5	3	5 ← five 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	

Example 3

Self Tutor

The table below shows the number of aces served by a sample of 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 (xf)	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	$\sum f = 55$	$\sum xf = 179$	

$$\begin{aligned} \mathbf{a} \quad \bar{x} &= \frac{\sum xf}{\sum f} \\ &= \frac{179}{55} \\ &\approx 3.25 \text{ aces} \end{aligned}$$

In this case $\frac{\sum xf}{\sum f}$ is short for $\frac{\sum_{j=1}^6 x_j f_j}{\sum_{j=1}^6 f_j}$.



- b** There are 55 data values, so $n = 55$. $\frac{n+1}{2} = 28$, so the median is the 28th ordered data value. From the cumulative frequency column, the 16th to 33rd ordered data values are 3 aces.
 \therefore the 28th ordered 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.

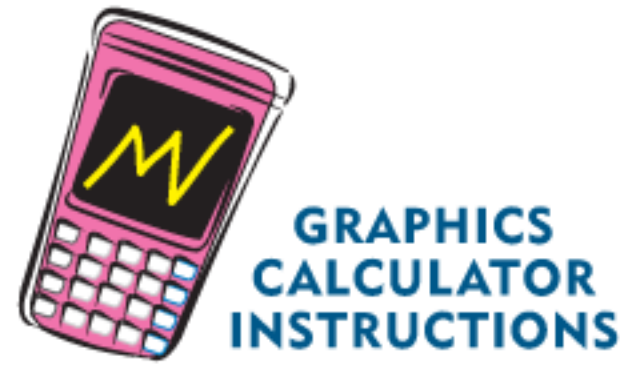
EXERCISE 13C

1 The table alongside shows the number of people in cars on a road.

Calculate the:

- a mode b median c mean.

Check your answers using your graphics calculator.



Number of people	Frequency
1	13
2	8
3	4
4	5
<i>Total</i>	30

2 The frequency table alongside shows the number of phone calls made in a day by 50 fifteen-year-olds.

- a For this data set, 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?
- e Which measure of centre would be the most suitable for this data set?

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

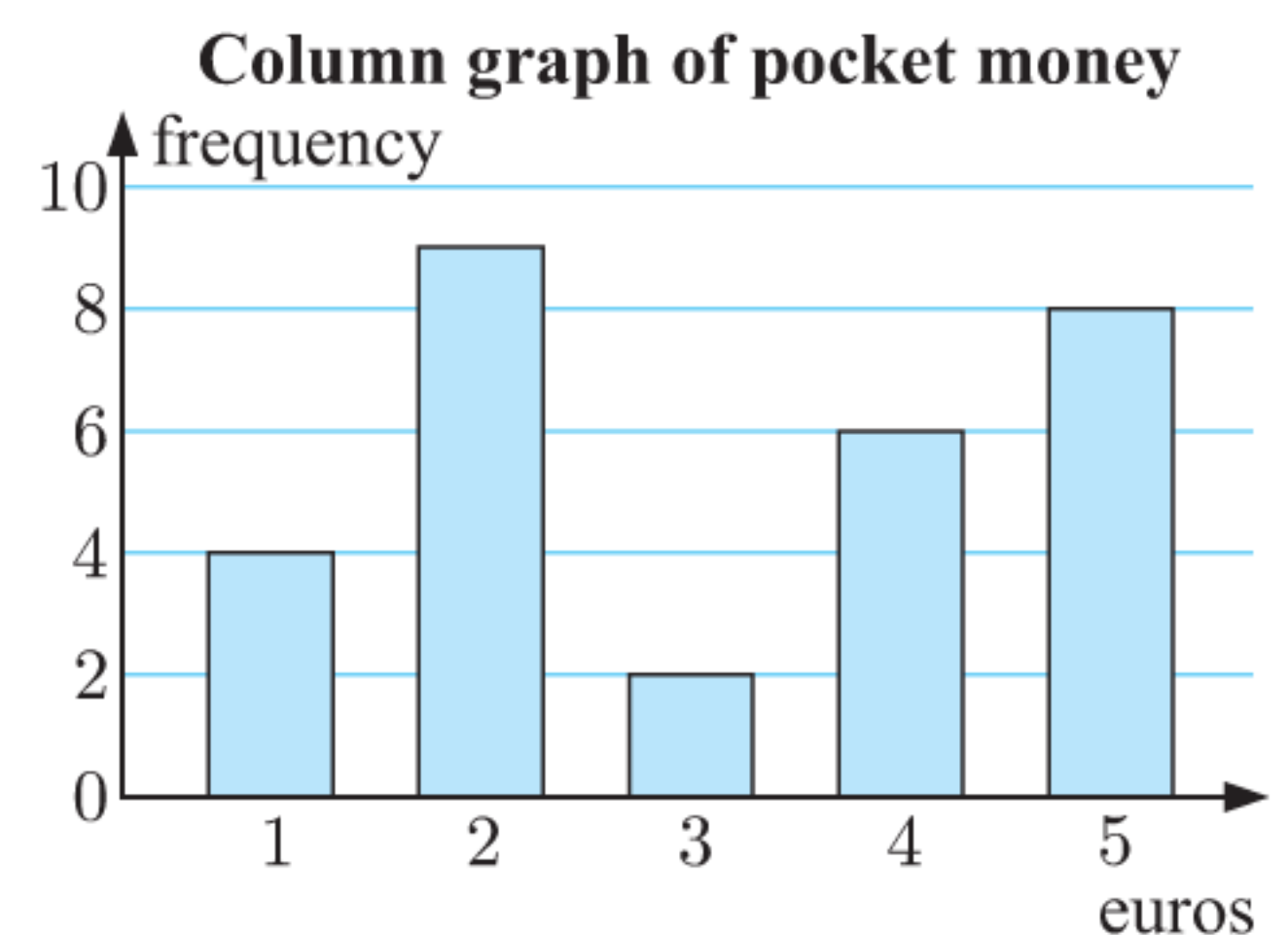
3 Families at a school in Manchester were surveyed, and the number of children in each family was recorded. The results of the survey are shown alongside.

- a Calculate the:
 i mean ii mode iii median.
- b The average British family has 2.075 children. How does this school compare to the national average?
- c Describe the skewness of the data.
- d How has the skewness of the data affected the measures of the centre of the data set?

Number of children	Frequency
1	5
2	28
3	15
4	8
5	2
6	1
<i>Total</i>	59

4 The column graph shows the weekly pocket money for a class of children.

- a Construct a frequency table from the graph.
- b Determine the total number of children in the class.
- c Find the:
 i mean ii median iii mode of the data.
- d Which of the measures of centre can be found easily using the graph only?



5 Out of 31 measurements, 15 are below 10 cm and 12 are above 11 cm. Find the median if the other 4 measurements are 10.1 cm, 10.4 cm, 10.7 cm, and 10.9 cm.

- 6 In an office of 20 people there are only 4 salary levels paid:

\$100 000 (1 person), \$84 000 (3 people),
\$70 000 (6 people), \$56 000 (10 people)

- a Calculate:
- i the median salary
 - ii the modal salary
 - iii the mean salary.
- b Which measure of central tendency might be used by the boss who is against a pay rise for the other employees?



- 7 The table shows the test scores for a class of students. A pass is a score of 5 or more.

Score	2	3	4	5	6	7	8
Frequency	0	2	3	5	x	4	1

- a Given that the mean score was 5.45, find x .
- b Find the percentage of students who passed.

D

GROUPED DATA

When information has been gathered in groups or classes, we use the **midpoint** or **mid-interval value** to represent all data values within each interval.

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

INVESTIGATION 2

MID-INTERVAL VALUES

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

This table summarises the marks out of 50 received by students in a Physics examination. 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 Physics examination mark 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 Physics examination mark 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 Physics examination mark was approximately”
- 4 Discuss with your class how accurate you think an estimate of the mean using mid-interval values will be. How is this accuracy affected by the number and width of the class intervals?

Example 4

Self Tutor

The table below shows the ages of bus drivers. Estimate the mean age, to the nearest year.

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

Age (years)	Frequency (f)	Midpoint (x)	xf
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	$\sum f = 137$		$\sum xf = 5161$

$$\begin{aligned} \bar{x} &= \frac{\sum xf}{\sum f} \\ &= \frac{5161}{137} \\ &\approx 37.7 \end{aligned}$$

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

EXERCISE 13D

1 Simone recorded the lengths of her phone calls for one week. The results are shown in the table alongside.

- a How many phone calls did she make during the week?
- b Estimate the mean length of the calls.

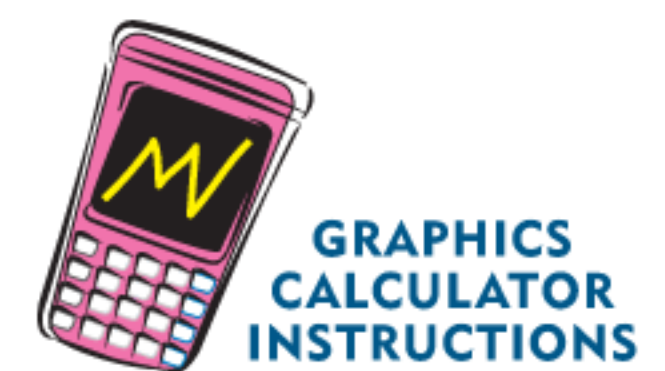
Time (t min)	Frequency
$0 \leq t < 10$	17
$10 \leq t < 20$	10
$20 \leq t < 30$	9
$30 \leq t < 40$	4

The midpoint of an interval is the average of its endpoints.



2 50 students sat a Mathematics test. Estimate the mean score given these results:

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



Check your answers using your calculator.

3 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 Estimate the mean amount of petrol sold for the day.
- d Find the modal class for this distribution. Explain your answer.

Amount of petrol (P L)	Frequency
$2000 < P \leq 3000$	4
$3000 < P \leq 4000$	4
$4000 < P \leq 5000$	9
$5000 < P \leq 6000$	14
$6000 < P \leq 7000$	23
$7000 < P \leq 8000$	16

4 The data below shows the runs scored by Jeff over an entire cricket season.

17	5	22	13	6	0	15	20
14	7	28	36	13	28	9	18
2	23	12	27	5	22	3	0
32	8	13	25	9			



- a Organise the data into the groups 0 - 9, 10 - 19, 20 - 29, 30 - 39.
- b Use your grouped data to estimate the mean number of runs scored.
- c Use the raw data to find the exact mean number of runs scored. How accurate was your estimate in b?

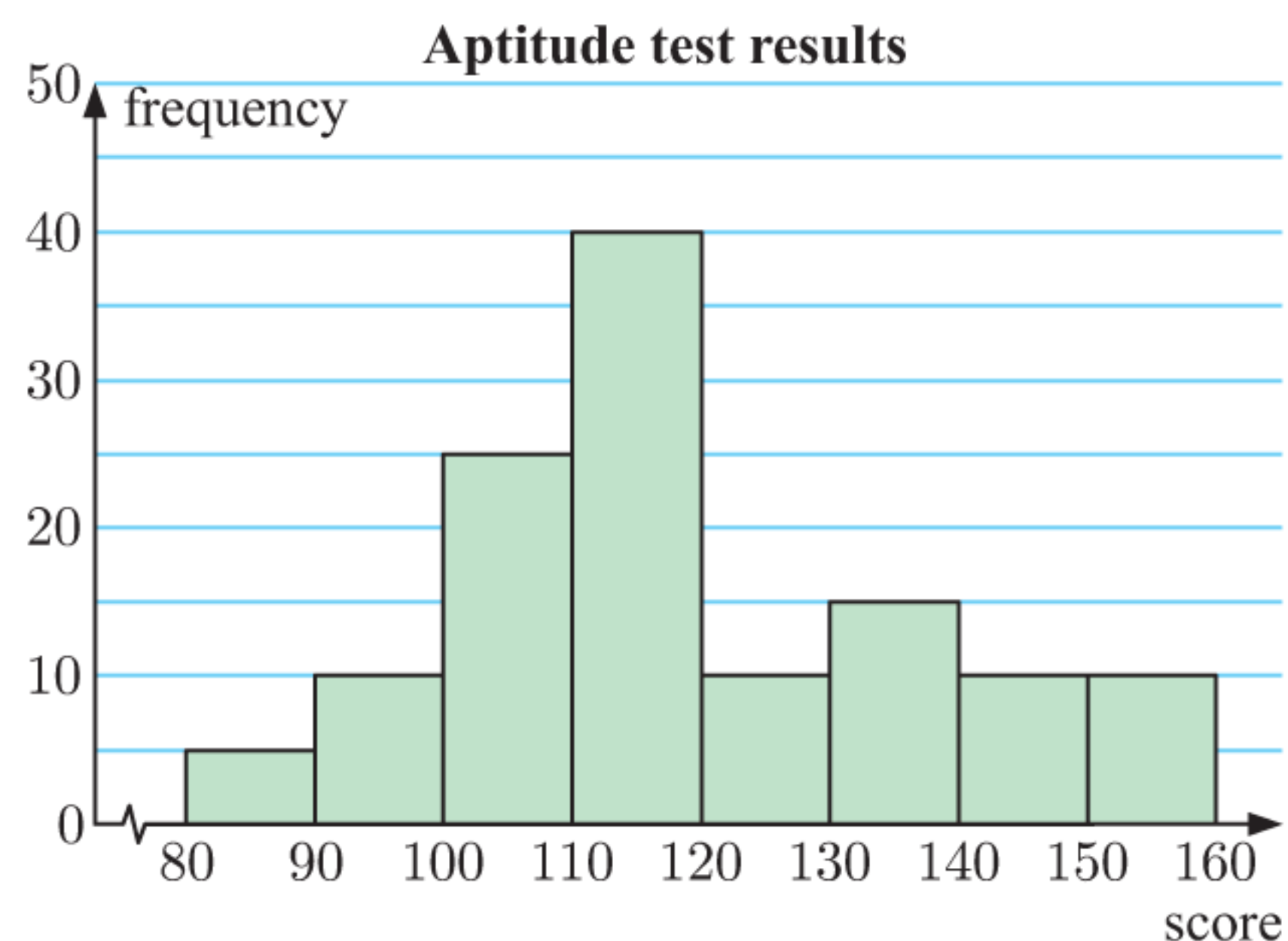
5 The manager of a bank decides to investigate the time customers wait to be served. The results for 300 customers are shown in the table alongside.

<i>Waiting time (t min)</i>	<i>Frequency</i>
$0 \leq t < 1$	p
$1 \leq t < 2$	42
$2 \leq t < 3$	50
$3 \leq t < 4$	78
$4 \leq t < 5$	60
$5 \leq t < 6$	30
$6 \leq t < 7$	16

- a Determine the value of p .
- b Estimate the mean waiting time.
- c What percentage of customers waited for at least 5 minutes?

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 took the test?
- b Estimate the mean score for the test.
- c What fraction of the people scored less than 100 for the test?
- d What percentage of the people scored more than 130 for the test?



E

MEASURING THE SPREAD OF DATA

Consider the following statements:

- The mean height of 20 boys in a Year 12 class was found to be 175 cm.
- A carpenter used a machine to cut 20 planks of length 175 cm.

Even though the means of the two data sets are the same, there will clearly be a greater *variation* in the heights of boys than in the lengths of the planks.

Commonly used statistics that measure the spread of a data set are:

- the **range**
- the **interquartile range**
- the **variance**
- the **standard deviation**.

We will look at variance and standard deviation later in the Chapter.

THE RANGE

The **range** is the difference between the **maximum** data value and the **minimum** data value.

$$\text{range} = \text{maximum} - \text{minimum}$$

As a statistic for discussing the spread of a data set, the range is not considered to be particularly reliable. This is because it only uses two data values. It may be influenced by extreme values or outliers.

However, the range is useful for purposes such as choosing class intervals.

Example 5

Self Tutor

The weight, in kilograms, of the pumpkins in Herb's crop are:
2.3, 3.1, 2.7, 4.1, 2.9, 4.0, 3.3, 3.7, 3.4, 5.1, 4.3, 2.9, 4.2
Find the range of the data.

$$\begin{aligned} \text{Range} &= \text{maximum} - \text{minimum} \\ &= 5.1 - 2.3 \\ &= 2.8 \text{ kg} \end{aligned}$$

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** (Q_1).

The middle value of the *upper* half is called the **upper quartile** (Q_3).

The **interquartile range (IQR)** is the range of the middle half of the data.

$$\begin{aligned} \text{interquartile range} &= \text{upper quartile} - \text{lower quartile} \\ \text{IQR} &= Q_3 - Q_1 \end{aligned}$$

The median is sometimes referred to as Q_2 because it is the 2nd quartile.



Example 6

Self Tutor

For the data set 5 5 7 3 8 2 3 4 6 5 7 6 4, find:

- a the median
- b Q_1 and Q_3
- c the interquartile range.

The ordered data set is: 2 3 3 4 4 5 5 5 6 6 7 7 8 (13 data values)

- a Since $n = 13$, $\frac{n+1}{2} = 7$ \therefore the median is the 7th data value.

~~2 3 3 4 4 5 5 5 6 6 7 7 8~~

\therefore median = 5

- b Since the median is a data value we now ignore it and split the remaining data into two:

lower half
upper half

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

$$Q_1 = \text{median of lower half} = \frac{3+4}{2} = 3.5$$

$$Q_3 = \text{median of upper half} = \frac{6+7}{2} = 6.5$$

- c $\text{IQR} = Q_3 - Q_1 = 6.5 - 3.5 = 3$

Example 7

Self Tutor

For the data set 12 24 17 10 16 29 22 18 32 20, find:

- a** the median **b** Q_1 and Q_3 **c** the interquartile range.

The ordered data set is: 10 12 16 17 18 20 22 24 29 32 (10 data values)

- a** Since $n = 10$, $\frac{n+1}{2} = 5.5$ \therefore the median is the average of the 5th and 6th data values.

~~10 12 16 17 18 20 22 24 29 32~~

$$\therefore \text{median} = \frac{\text{5th value} + \text{6th value}}{2} = \frac{18 + 20}{2} = 19$$

- b** We have an even number of data values, so we include all data values when we split the data set into two:

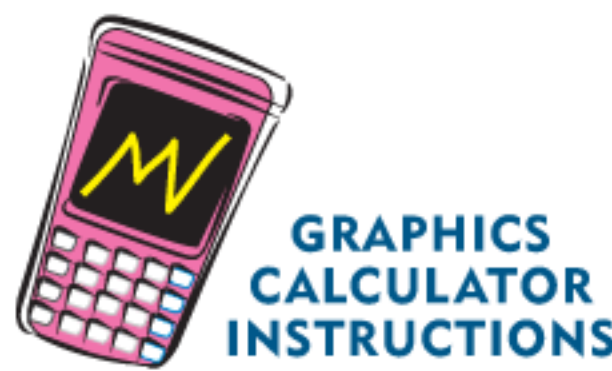
 lower half upper half
 { 10 12 16 17 18 } { 20 22 24 29 32 }

$$Q_1 = \text{median of lower half} = 16$$

$$Q_3 = \text{median of upper half} = 24$$

- c** $\text{IQR} = Q_3 - Q_1 = 24 - 16 = 8$

Technology can be used to help calculate the interquartile range. Your graphics calculator gives the values of Q_1 and Q_3 , from which the interquartile range is found using $\text{IQR} = Q_3 - Q_1$.



	Rad	Norm1	ab/c	Real
1-Variable				
minX	=	10		↑
Q1	=	16		
Med	=	19		
Q3	=	24		
maxX	=	32		
Mod	=	10		↓

EXERCISE 13E

- 1** For each of the following data sets, make sure the data is ordered and then find:
- i** the median
 - ii** the lower and upper quartiles
 - iii** the range
 - iv** the interquartile range.

- a** 5, 6, 9, 10, 11, 13, 15, 16, 18, 20, 21
b 7, 7, 10, 13, 14, 15, 18, 19, 21, 21, 23, 24, 24, 26
c 21, 24, 19, 32, 15, 43, 38, 29
d 32, 45, 26, 28, 52, 57, 41, 69, 33, 20

Check your answers using your graphics calculator.

- 2** Jane and Ashley's monthly telephone bills are shown below:

Jane: \$35, \$47, \$29, \$38, \$29, \$34, \$42, \$29, \$36, \$40, \$36, \$31

Ashley: \$19, \$24, \$26, \$19, \$23, \$40, \$35, \$59, \$32, \$42, \$26, \$24

- a** Find the mean and median for each data set.
- b** Find the range and interquartile range for each data set.
- c** Which person generally pays more for their telephone bills?
- d** Which person has the greater variability in their telephone bills?

- 3** **a** Find the range and interquartile range for the data set:
- | | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|
| 9 | 12 | 7 | 15 | 14 | 22 | 18 | 11 | 20 | 15 |
| 20 | 10 | 13 | 67 | 25 | 18 | 11 | 7 | 14 | 19 |
- b** Identify the outlier in the data set.
c Recalculate the range and interquartile range with the outlier removed.
d Which measure of spread is more affected by the outlier?
- 4** Derrick and Gareth recorded the number of minutes they slept each night for 15 nights:
- Derrick:* 420, 435, 440, 415, 380, 400, 430, 450, 210, 445, 425, 445, 450, 420, 425
Gareth: 360, 420, 460, 430, 480, 340, 450, 490, 500, 460, 330, 470, 340, 480, 370
- a** Calculate the range and interquartile range for each data set.
b Which person’s data has the lower:
 i range **ii** interquartile range?
c Which measure of spread is more appropriate for determining who is generally the more consistent sleeper? Explain your answer.
- 5** $a, b, c, d, e, f, g, h, i, j, k, l,$ and m are 13 data values which have been arranged in *ascending* order.
a Which variable represents the median?
b Write down an expression for:
 i the range **ii** the interquartile range.
- 6** A data set has the following known measures of centre and spread:
- | | | | | |
|----------------|--------|------|-------|---------------------|
| <i>Measure</i> | median | mode | range | interquartile range |
| <i>Value</i> | 9 | 7 | 13 | 6 |
- Find the new value of each of these measures if every member of the data set is:
a increased by 2 **b** doubled.

DISCUSSION

Consider the data set:

5, 7, 7, 8, 9, 11, 11, 12, 14, 14, 15

- 1** Calculate Q_1 , Q_3 , and the IQR:

- by hand
- using your graphics calculator
- using a spreadsheet.

Do you get the same answers in all 3 cases?

- 2** If there is an odd number of data values, some statistical packages calculate quartiles by *including* the median in each half of the data.
- a** Check to see whether your spreadsheet calculates quartiles this way.
b Does this method necessarily change the *interpretation* of the calculated values?
c Are statistical packages that do this necessarily “wrong”?

F

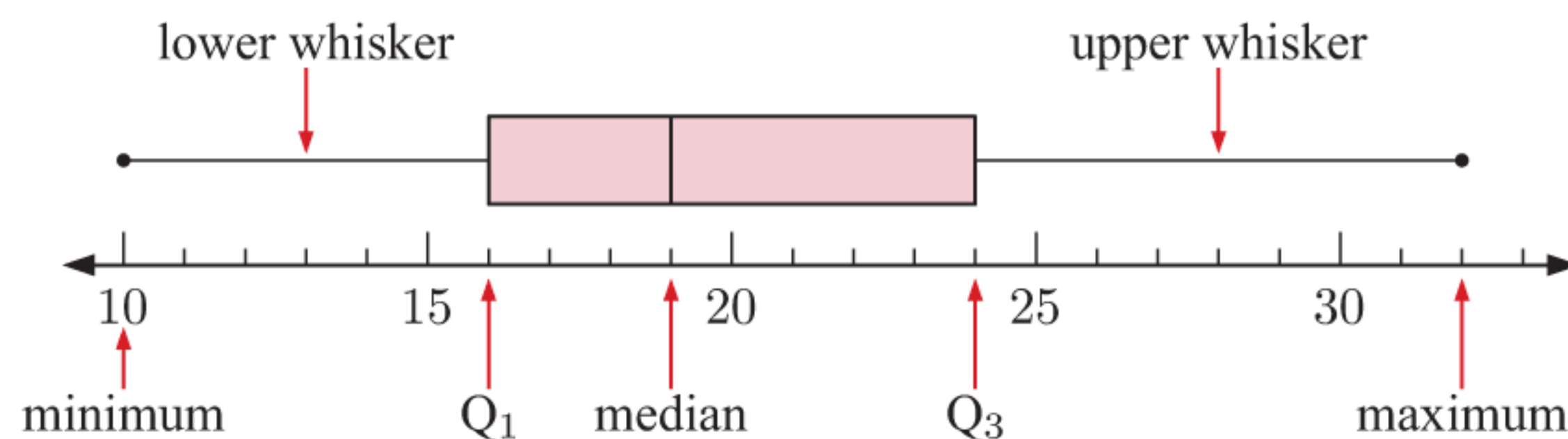
BOX AND WHISKER DIAGRAMS

A **box and whisker diagram** or simply **box plot** 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 7** on page 330, the five-number summary and box plot are:

minimum = 10
 $Q_1 = 16$
 median = 19
 $Q_3 = 24$
 maximum = 32

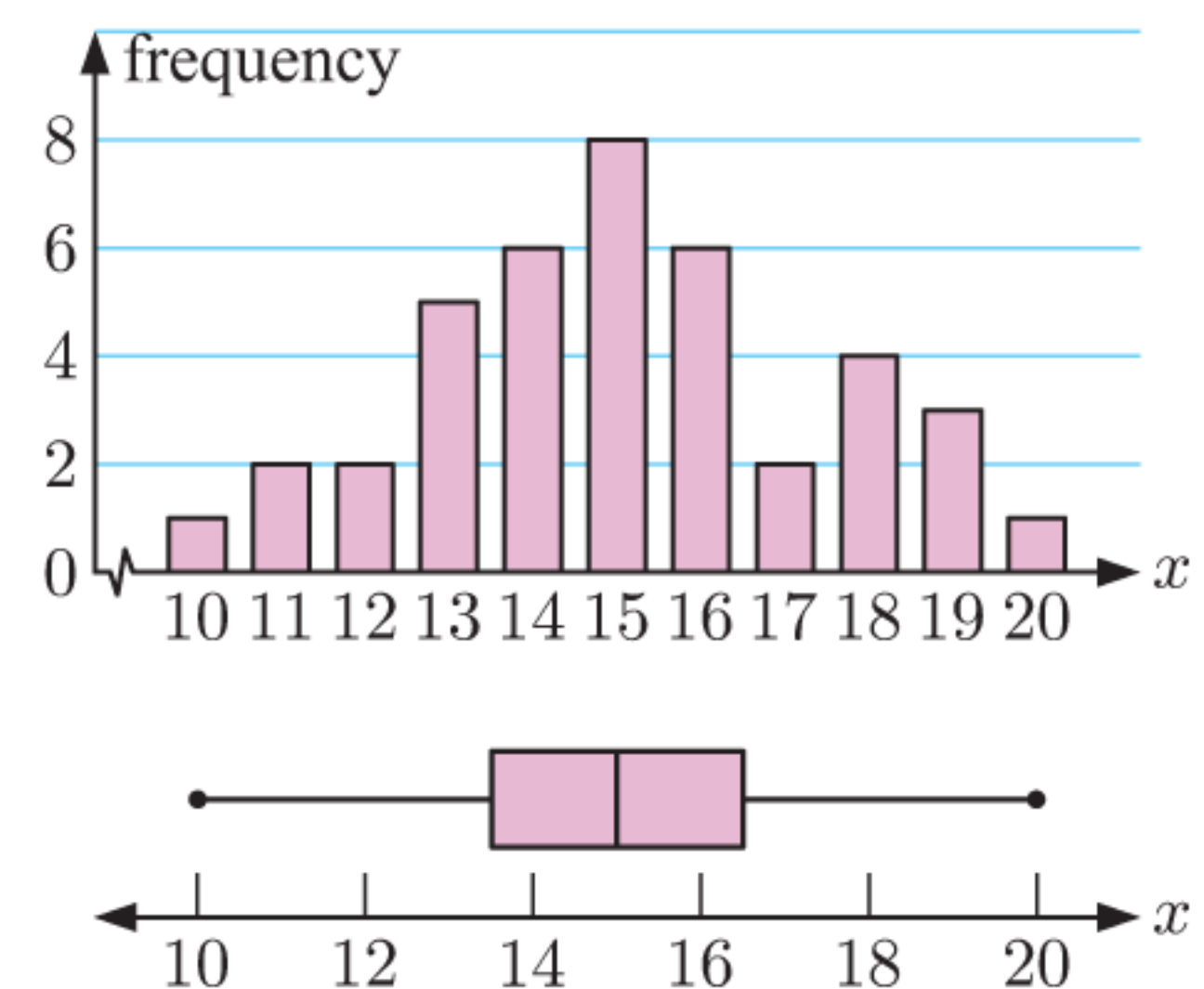


You should notice that:

- 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.

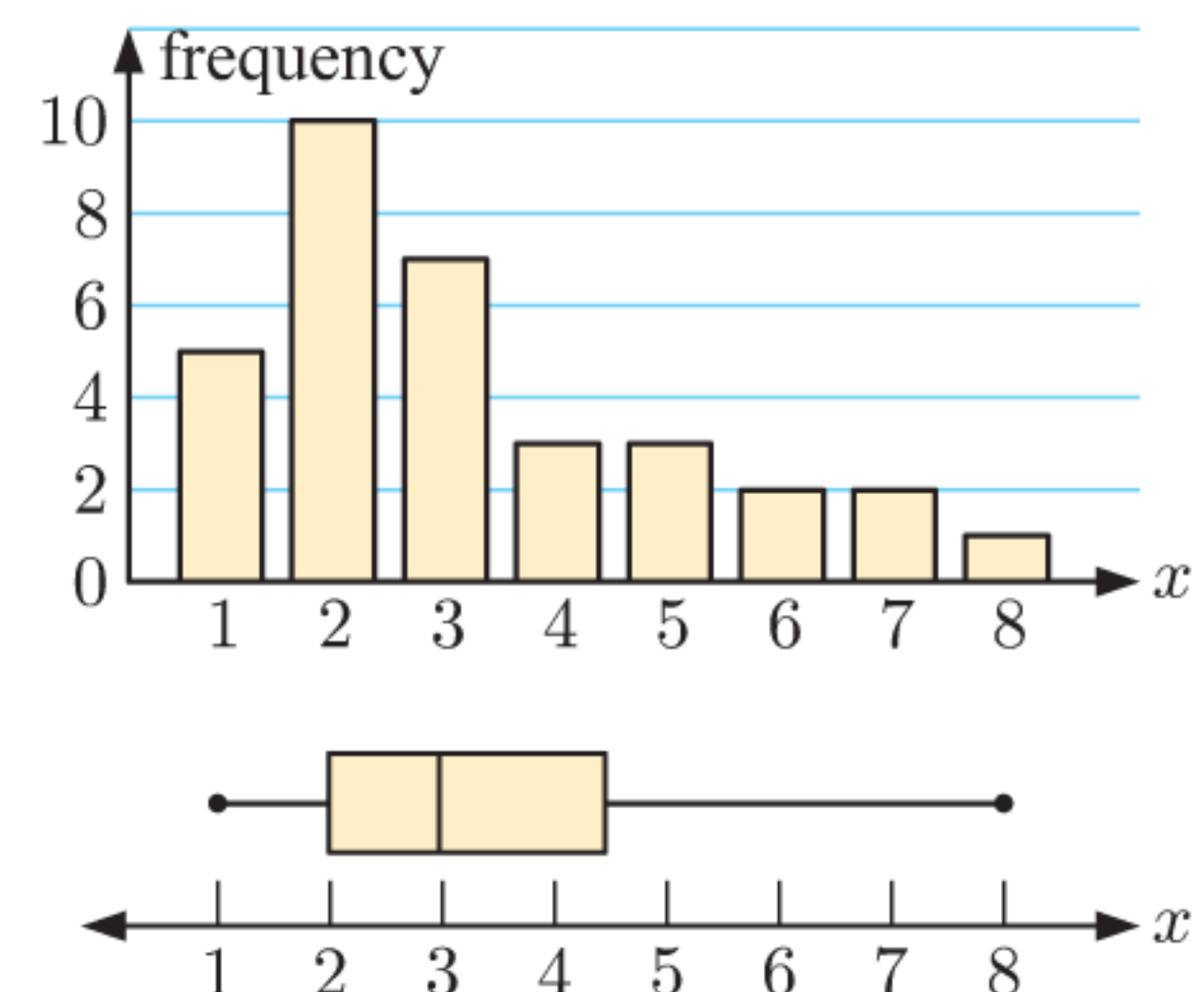
INTERPRETING A BOX PLOT

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



The whiskers of the box plot 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 box plot.



The upper whisker is longer than the lower whisker and the median line is closer to the left hand side of the box.

F

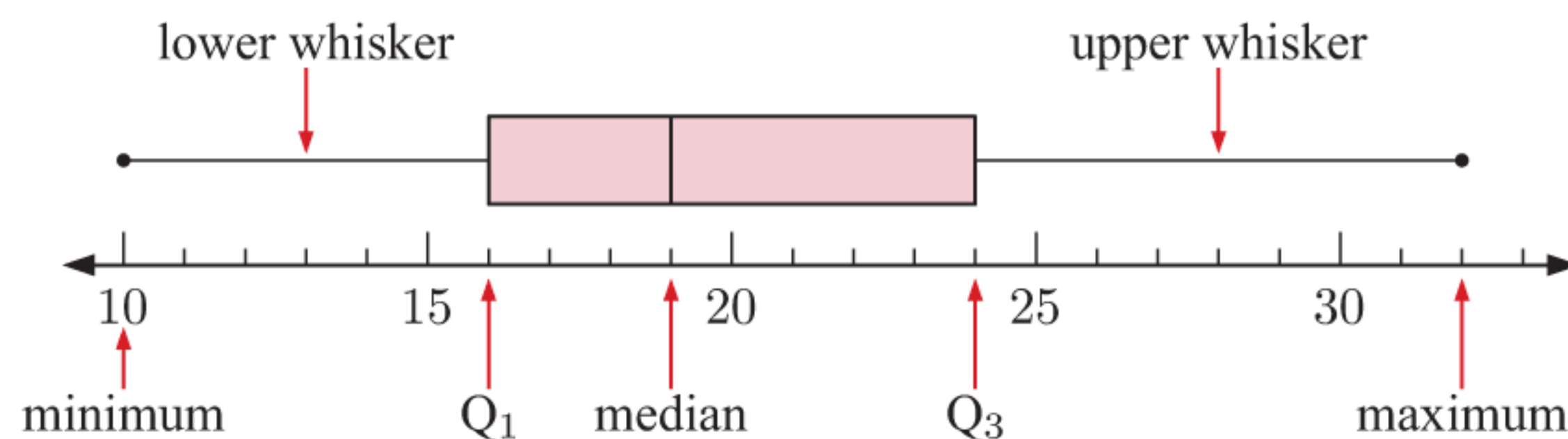
BOX AND WHISKER DIAGRAMS

A **box and whisker diagram** or simply **box plot** 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 7** on page 330, the five-number summary and box plot are:

minimum = 10
 $Q_1 = 16$
 median = 19
 $Q_3 = 24$
 maximum = 32

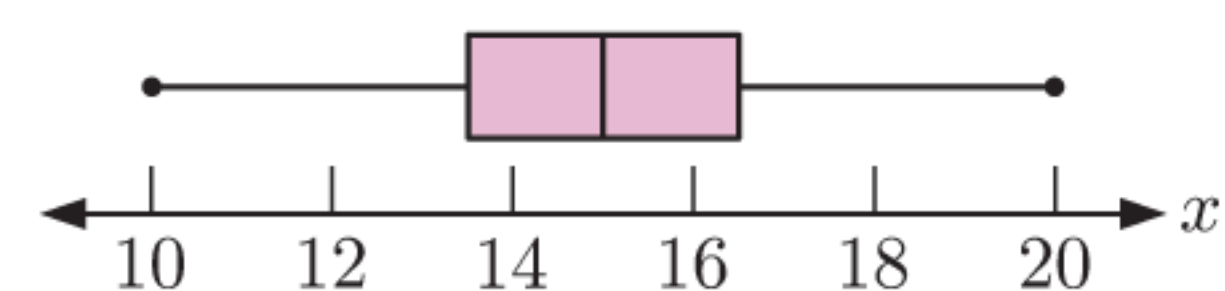
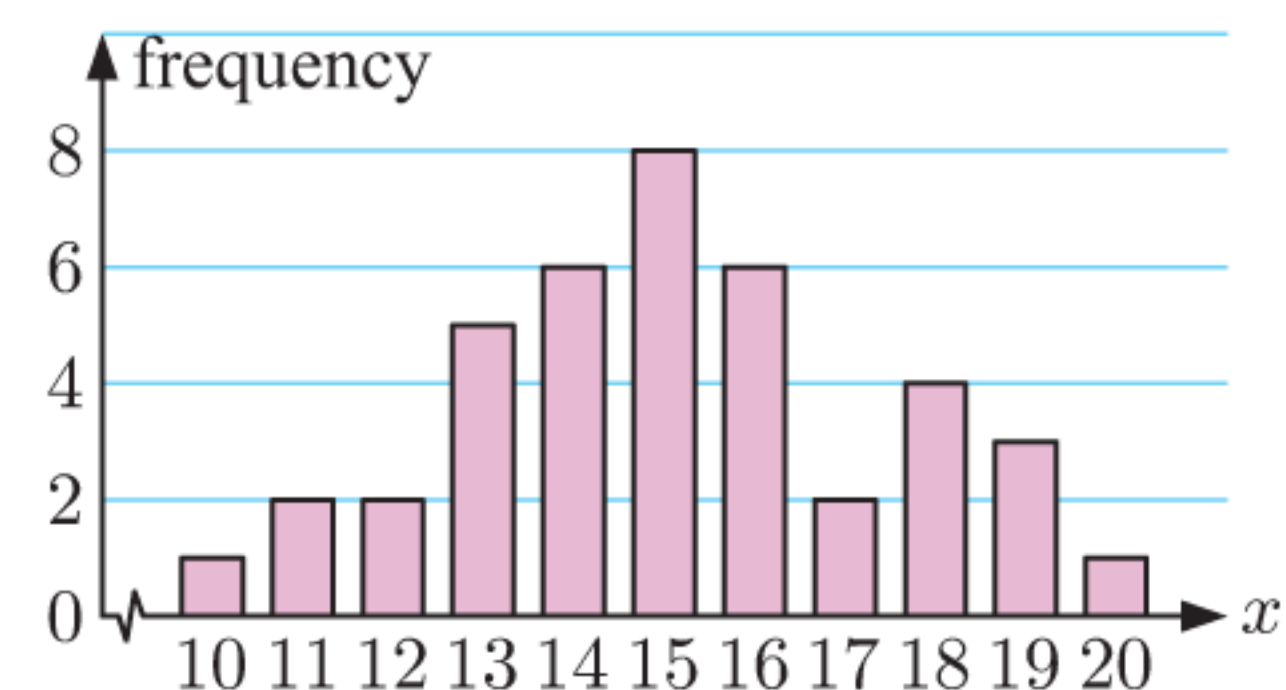


You should notice that:

- 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.

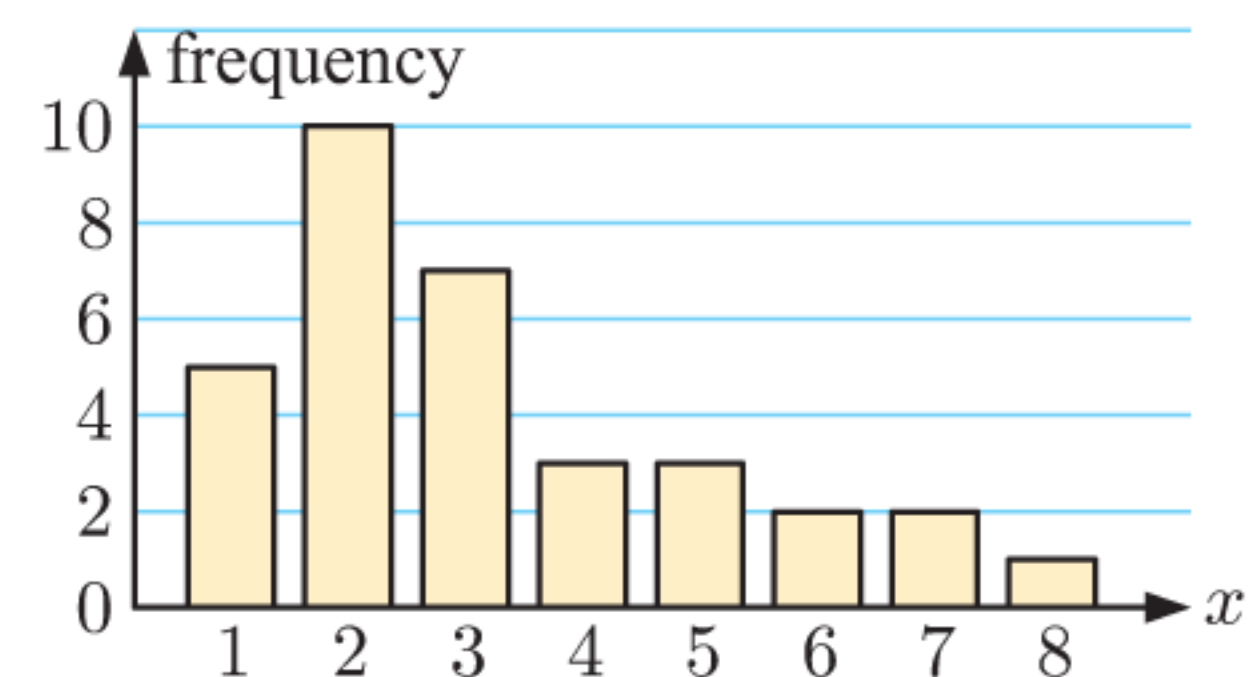
INTERPRETING A BOX PLOT

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



The whiskers of the box plot 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 box plot.



The upper whisker is longer than the lower whisker and the median line is closer to the left hand side of the box.

G

OUTLIERS

We have seen that **outliers** are extraordinary data that are separated from the main body of the data. However, we have so far identified outliers rather informally by looking at the data directly, or at a column graph of the data.

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

- **upper boundary = upper quartile + 1.5 × IQR**
Any data larger than the upper boundary is an outlier.
- **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 box plot. There may be more than one outlier at either end. Each whisker extends to the last value that is not an outlier.

Example 9
Self Tutor

Test the following data for outliers. Hence construct a box plot for the data.

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

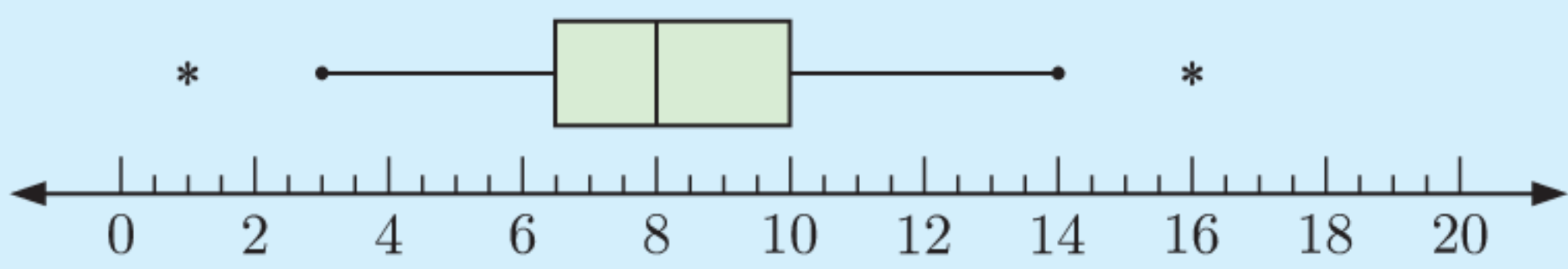
The ordered data set is:

1	3	3	5	6	7	7	7	8	8	8	8	9	9	10	10	12	13	14	16	{n = 20}
↓				↓				↓	↓					↓					↓	
min = 1				Q ₁ = 6.5				median = 8						Q ₃ = 10					max = 16	


IQR = Q₃ – Q₁ = 3.5

<i>Test for outliers:</i>	upper boundary = upper quartile + 1.5 × IQR = 10 + 1.5 × 3.5 = 15.25	and	lower boundary = lower quartile – 1.5 × IQR = 6.5 – 1.5 × 3.5 = 1.25
---------------------------	---	-----	---

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



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



EXERCISE 13G

- 1 A data set has lower quartile = 31.5, median = 37, and upper quartile = 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 are outliers?
 - d Draw a box plot of the data set.

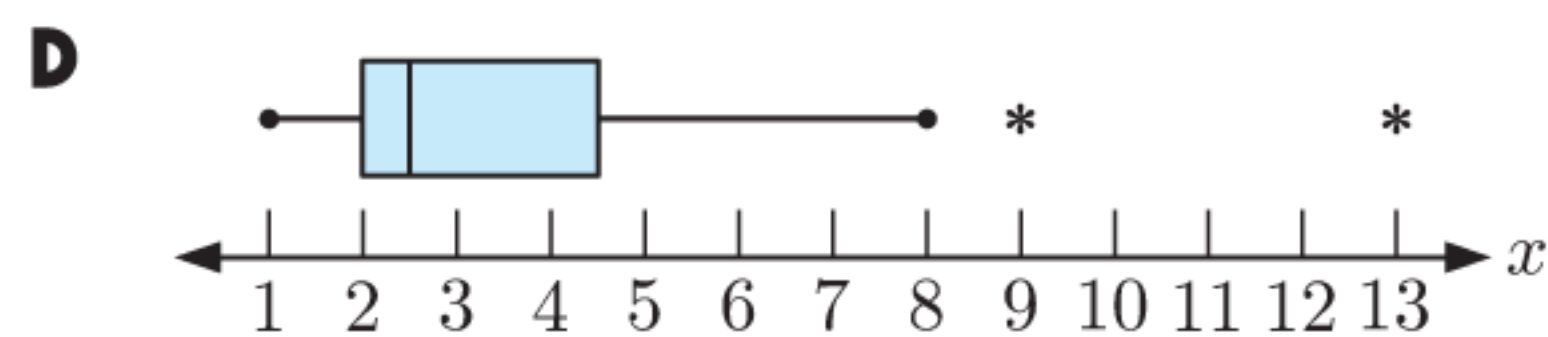
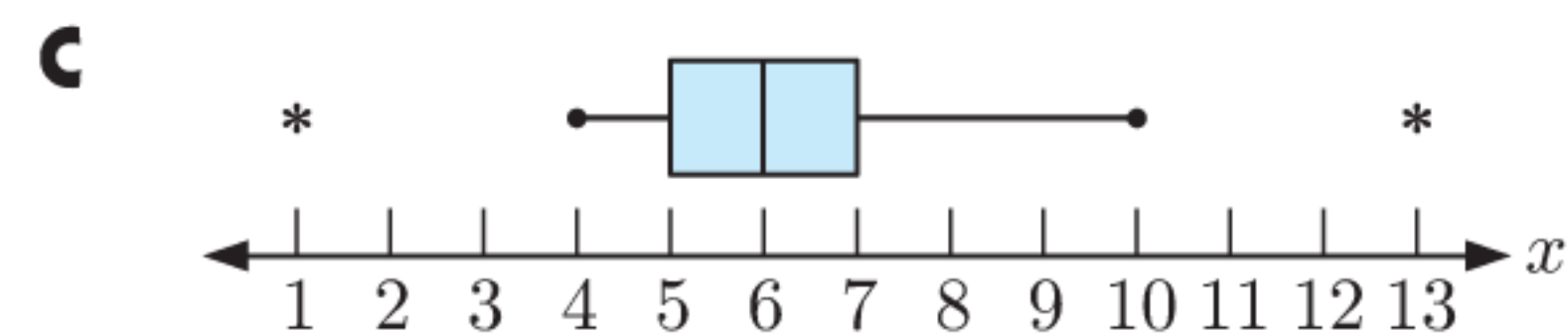
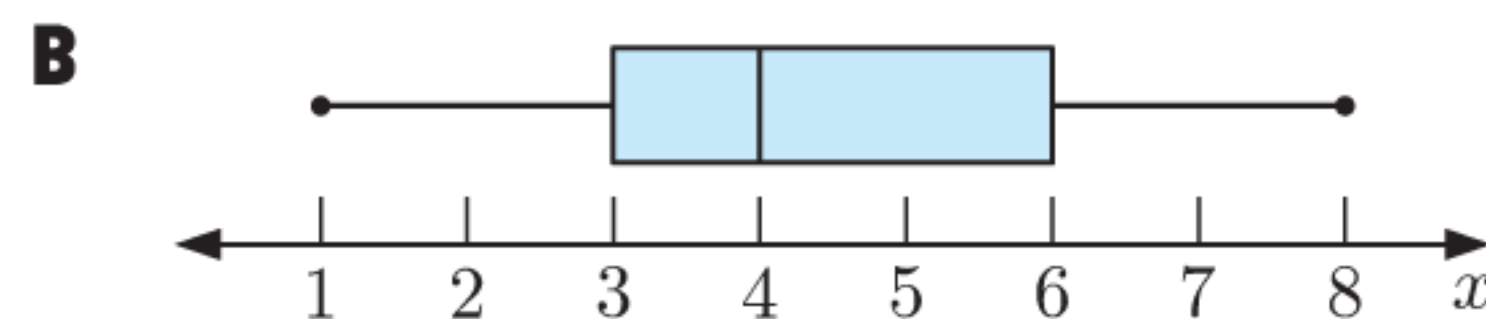
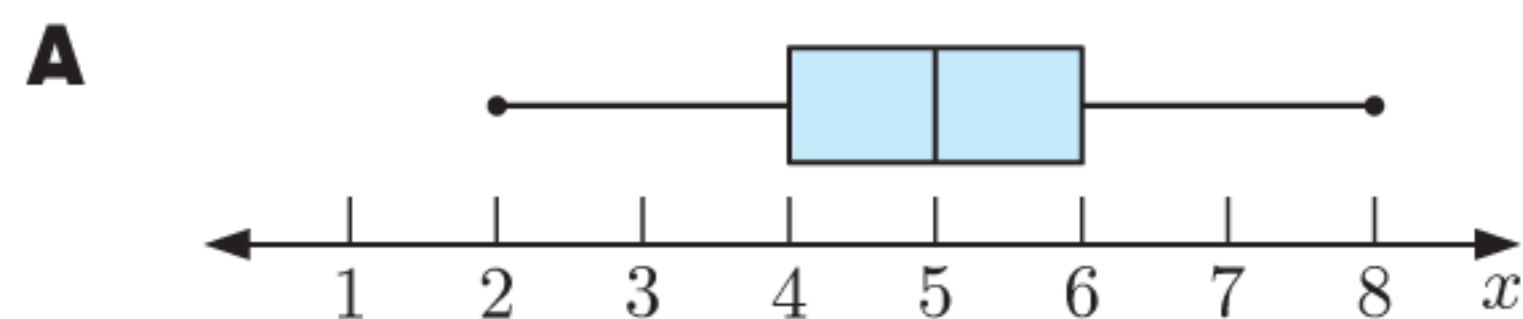
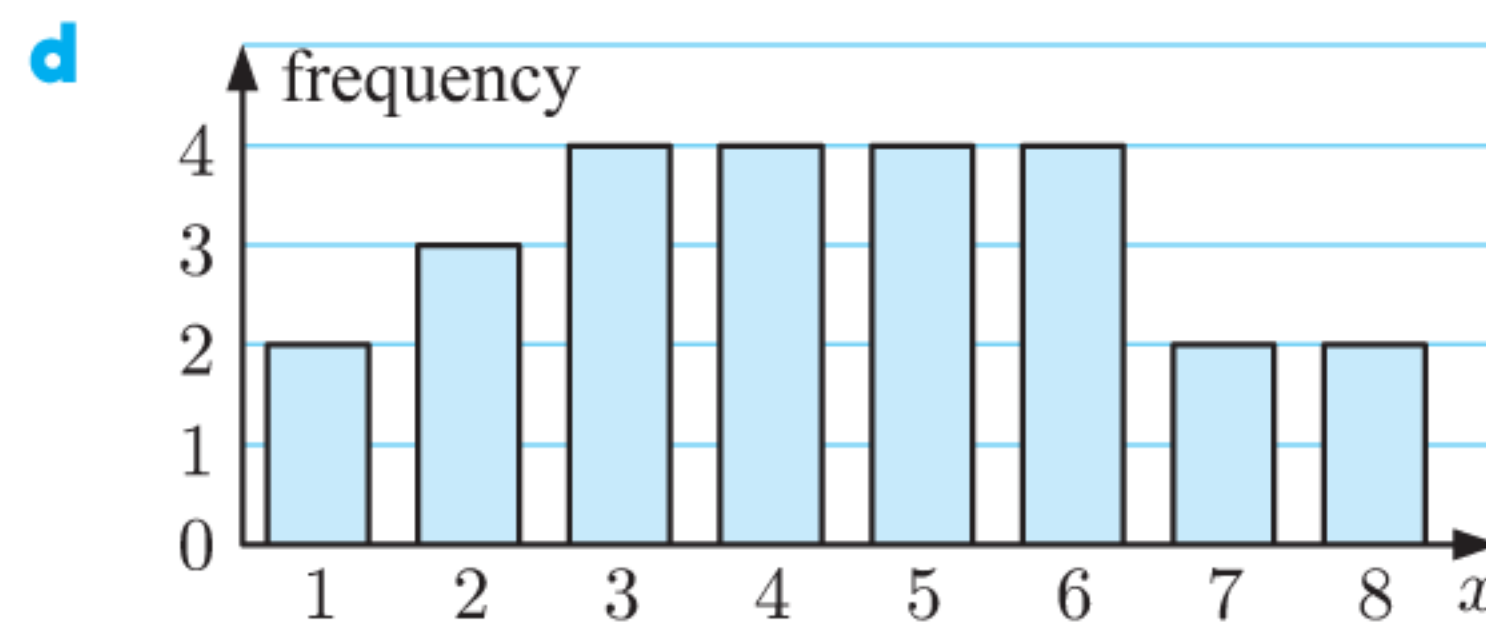
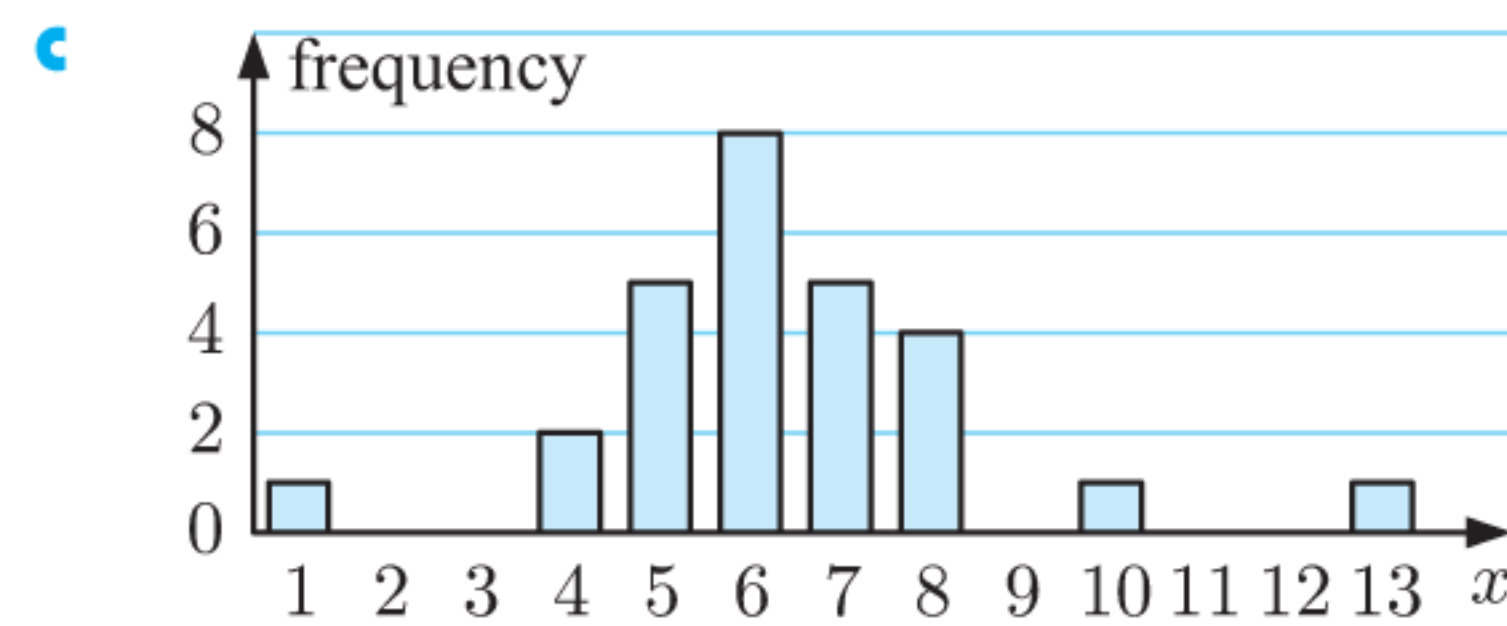
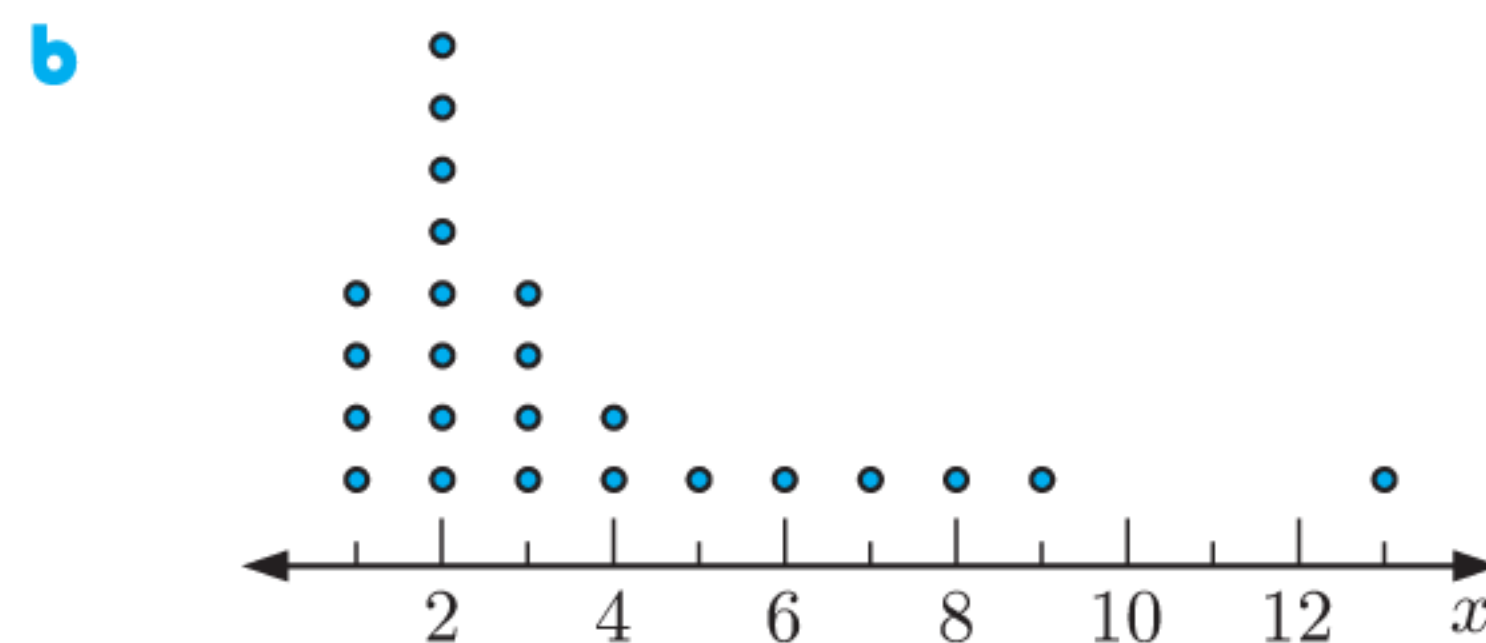
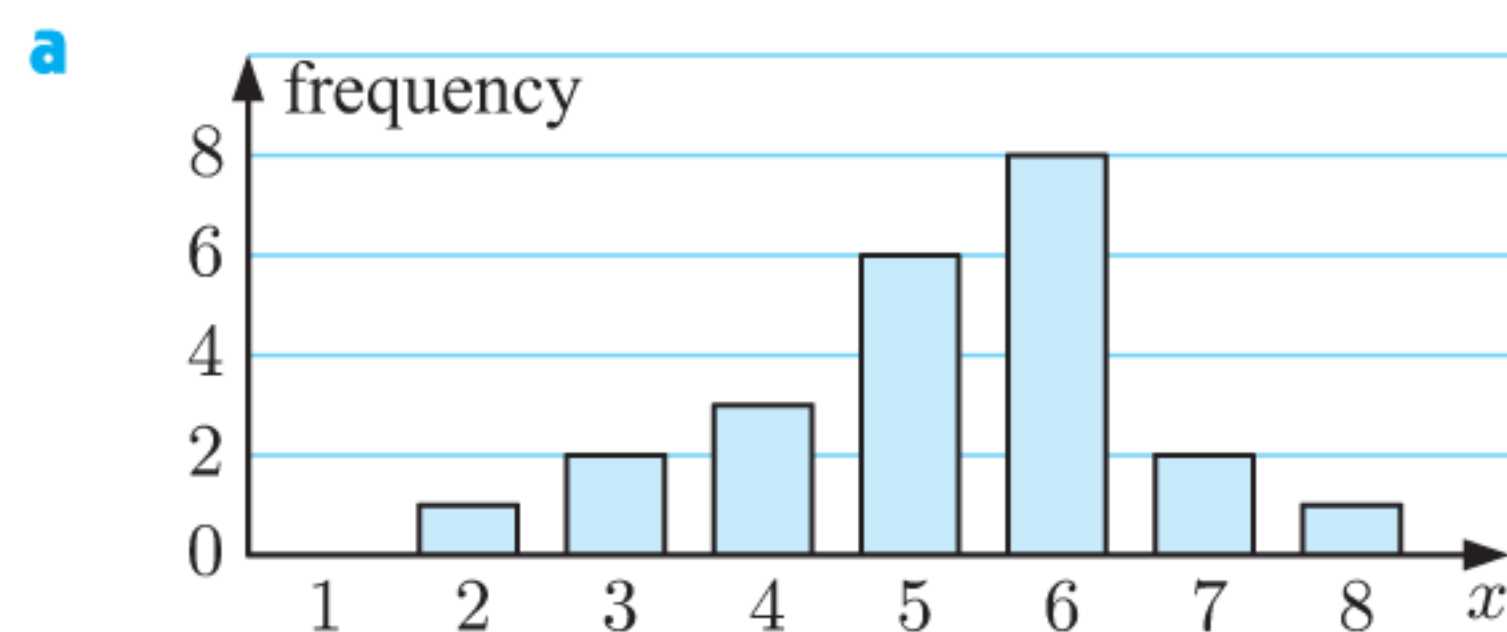
2 James goes bird watching for 25 days. The number of birds he sees each day are:

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 Find the lower and upper boundaries, and hence identify any outliers.
- d Draw a box plot of the data set.



3 Match each graph with its box plot:



4 The data below shows the number of properties sold by a real estate agent each week in 2018:

2	2	1	3	2	2	2	2	1	4	1	5	1
1	1	2	2	2	3	1	7	2	2	2	0	2
2	4	4	3	3	1	0	2	4	1	2	1	3
0	2	3	1	2	1	3	4	2	2	2	1	3

- a Draw a column graph to display the data.
- b From the column graph, does the data appear to have any outliers?
- c Calculate the upper and lower boundaries to test for outliers and hence check your answer to b.
- d Construct a box plot for the data.

H

PARALLEL BOX AND WHISKER DIAGRAMS

A **parallel box and whisker diagram** or **parallel box plot** 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 10

Self Tutor

A hospital trialling a new anaesthetic 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 predictable.

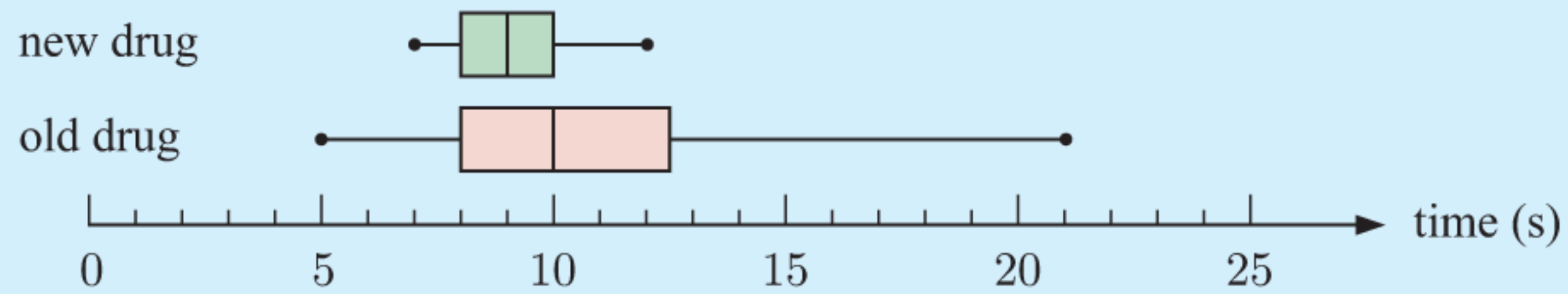
Old drug times (s): 8, 12, 9, 8, 16, 10, 14, 7, 5, 21,
13, 10, 8, 10, 11, 8, 11, 9, 11, 14

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

Draw a parallel box plot for the data sets and use it to compare the two drugs.

The five-number summaries are:

For the old drug:	min = 5	For the new drug:	min = 7
	Q ₁ = 8		Q ₁ = 8
	median = 10		median = 9
	Q ₃ = 12.5		Q ₃ = 10
	max = 21		max = 12



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

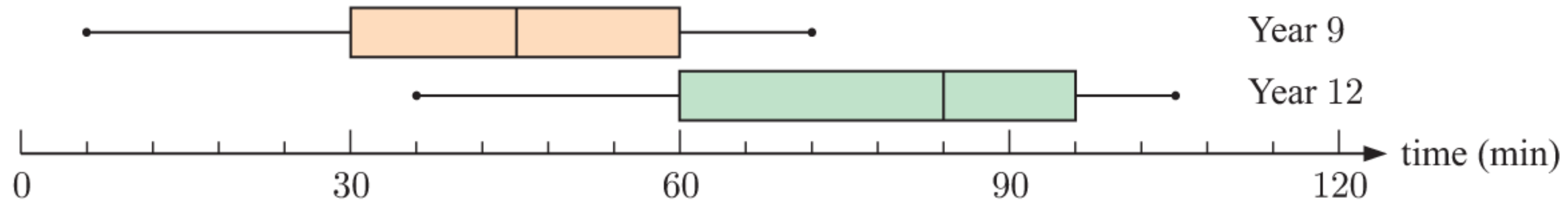
Comparing the spreads:

range for old drug = 21 – 5	range for new drug = 12 – 7
= 16	= 5
IQR for old drug = Q ₃ – Q ₁	IQR for new drug = Q ₃ – Q ₁
= 12.5 – 8	= 10 – 8
= 4.5	= 2

The new drug times are less “spread out” than the old drug times, so the new drug is more predictable.

EXERCISE 13H

1 The following parallel box plots 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		

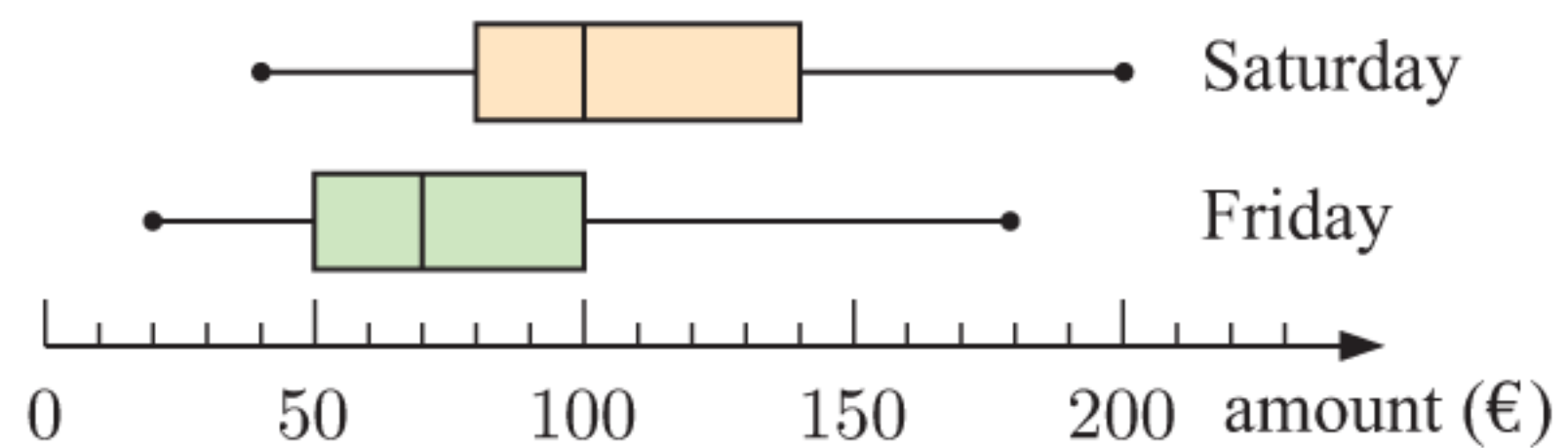
b For each group, determine the:

- i range
- ii interquartile range.

c Determine whether the following statements are true or false, or if there is 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 on a Saturday. The results are displayed on the parallel box plot shown.



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

b For each data set, determine the

- i range
- ii interquartile range.

3 After the final examination, the results of two classes studying the same subject were compiled in this parallel box plot.

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 Students who scored at least 70% received an achievement award. Find the percentage of students who received an award in:

- i class 1
- ii class 2.

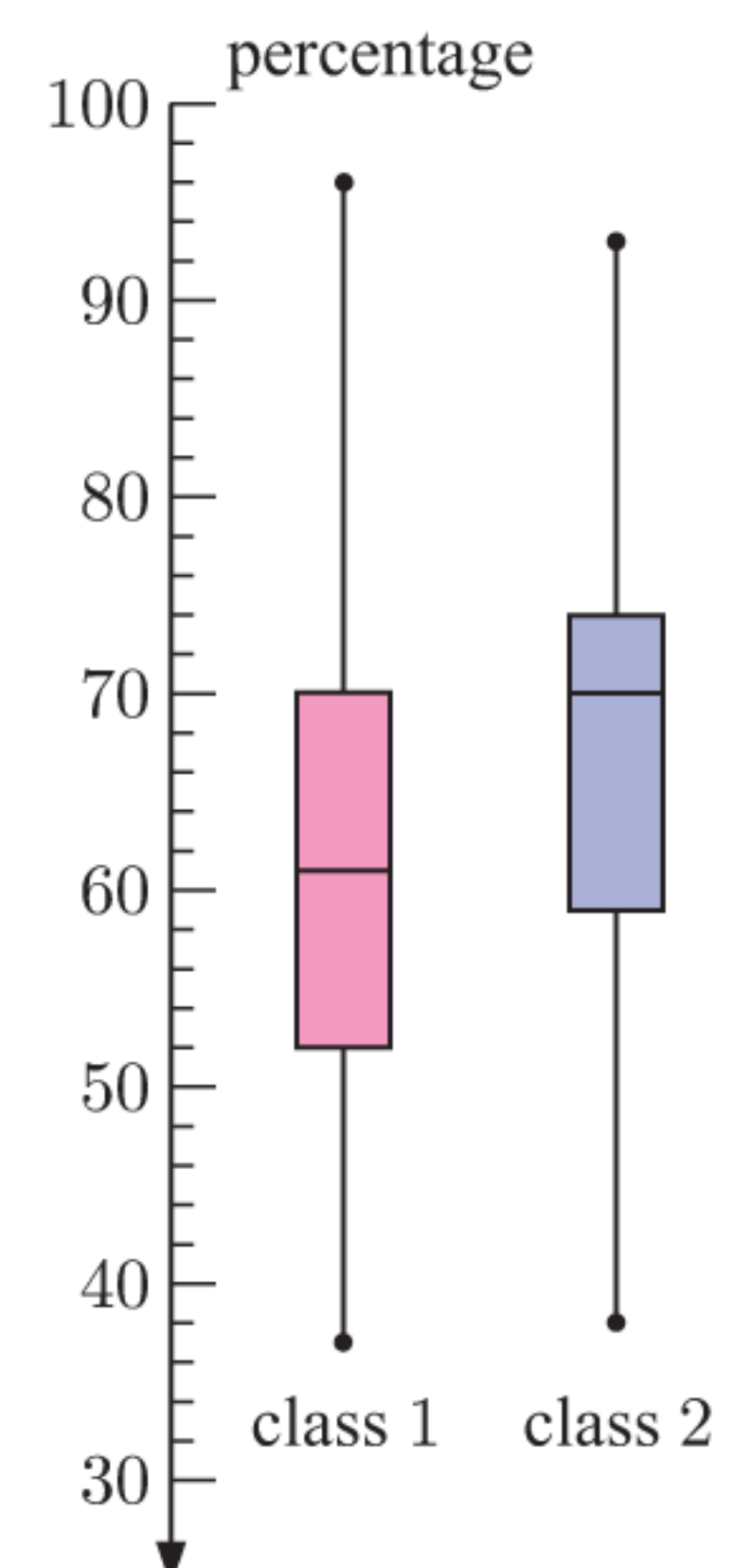
e Describe the distribution of marks in:

- 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 The data below are the durations, in minutes, of Kirsten and Erika's last 25 phone calls.

Kirsten: 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

Erika: 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

- Find the five-number summary for each data set.
 - Display the data in a parallel box plot.
 - Compare and comment on the distributions of the data.
- 5 Emil and Aaron play in the same handball team and are fierce but friendly rivals when it comes to scoring. During a season, the numbers of goals they scored in each match were:

Emil: 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

Aaron: 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

- Is the variable discrete or continuous?
- Enter the data into a graphics calculator or statistics package.
- Produce a column graph for each data set.
- Describe the shape of each distribution.
- Compare the measures of the centre of each distribution.
- Compare the spreads of each distribution.
- Draw a parallel box plot for the data.
- 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:

Old type: 103 96 113 111 126 100 122 110 84 117 103 113 104 104
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

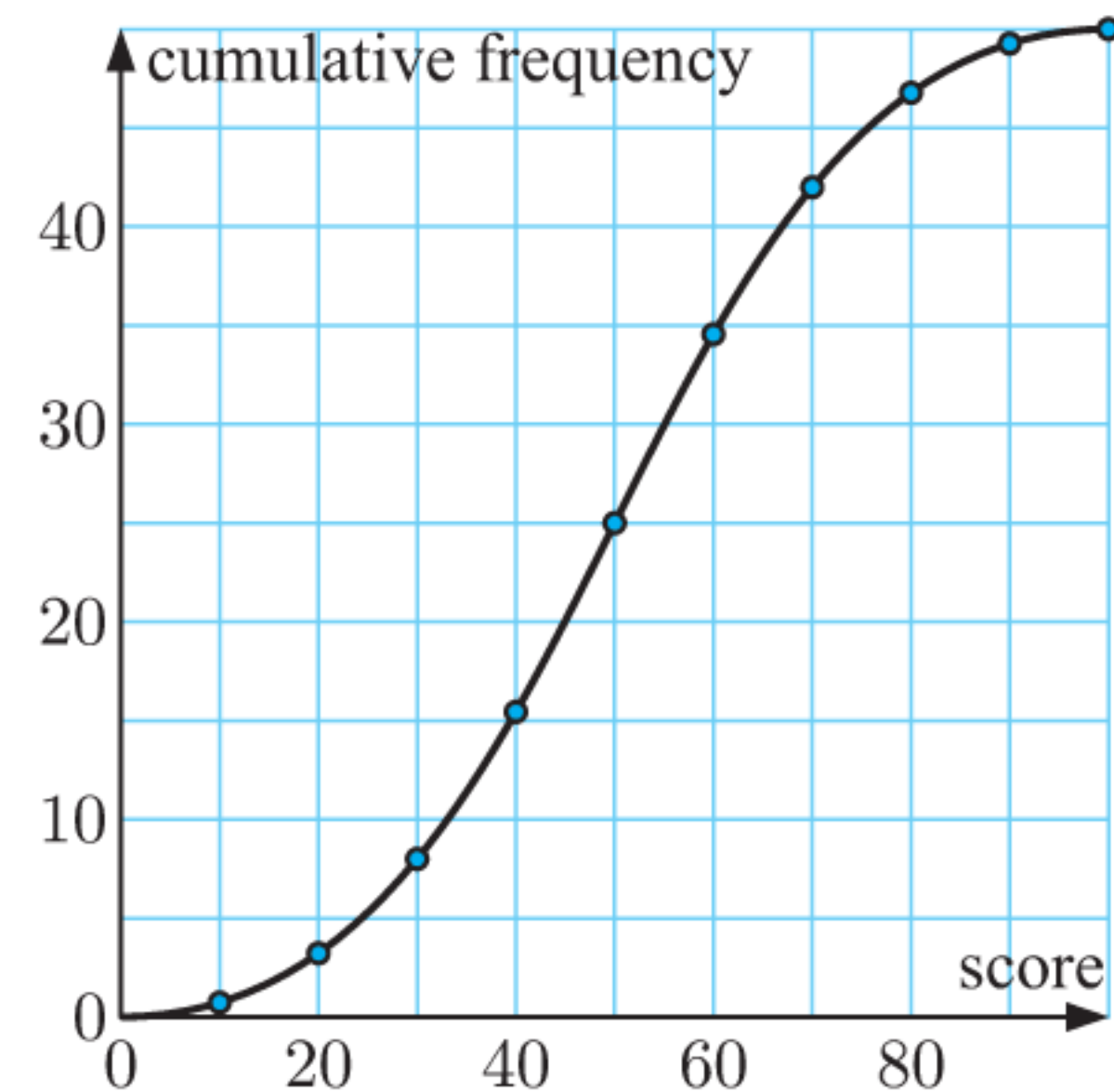
New type: 146 131 132 160 128 119 133 117 139 123 109 129 109 131
191 117 132 107 141 136 146 142 123 144 145 125 164 125
133 124 153 129 118 130 134 151 145 131 133 135

- Is the variable discrete or continuous?
- Enter the data into a graphics calculator or statistics package. Compare the measures of centre and spread.
- Draw a parallel box plot.
- Describe the shape of each distribution.
- What conclusions, if any, can be drawn from the data?

I CUMULATIVE FREQUENCY GRAPHS

If we want to know the number or proportion of scores that lie above or below a particular value, we add a **cumulative frequency** column to a **frequency 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 neighbouring 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** (Q_1) is the 25th percentile
- the **median** (Q_2) is the 50th percentile
- the **upper quartile** (Q_3) is the 75th percentile.

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

Example 11

Self Tutor

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

- Add a cumulative frequency column to the table.
- Represent the data on a cumulative frequency graph.
- Use your graph to estimate the:
 - median finishing time
 - number of competitors who finished in less than 155 minutes
 - percentage of competitors who took more than 159 minutes to finish
 - time taken by a competitor who finished in the top 20% of runners completing the marathon.

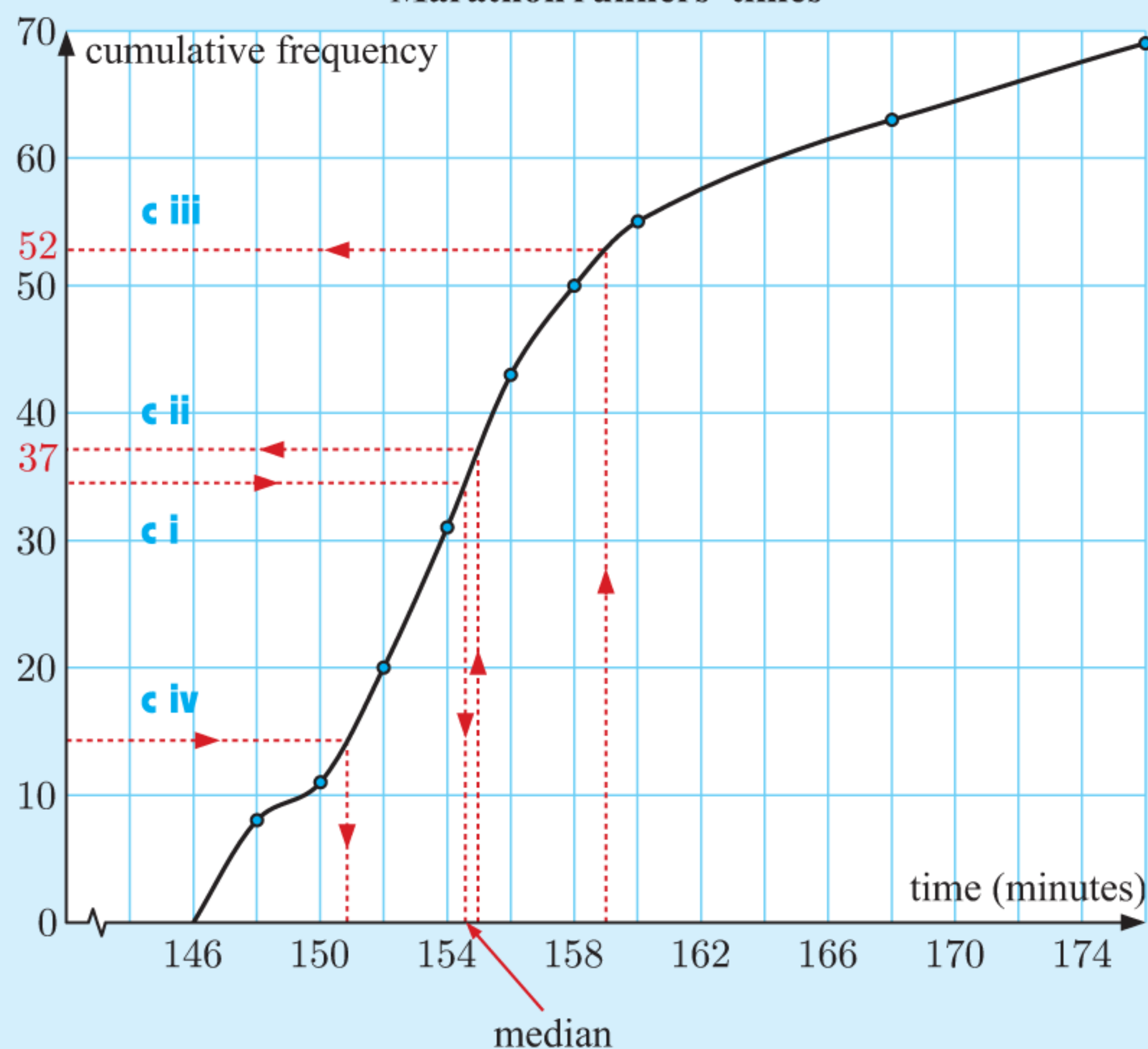
Time (t min)	Frequency
$146 \leq t < 148$	8
$148 \leq t < 150$	3
$150 \leq t < 152$	9
$152 \leq t < 154$	11
$154 \leq t < 156$	12
$156 \leq t < 158$	7
$158 \leq t < 160$	5
$160 \leq t < 168$	8
$168 \leq t < 176$	6

a

Time (t min)	Frequency	Cumulative frequency
$146 \leq t < 148$	8	8
$148 \leq t < 150$	3	11
$150 \leq t < 152$	9	20
$152 \leq t < 154$	11	31
$154 \leq t < 156$	12	43
$156 \leq t < 158$	7	50
$158 \leq t < 160$	5	55
$160 \leq t < 168$	8	63
$168 \leq t < 176$	6	69

$8 + 3 = 11$ competitors completed the marathon in less than 150 minutes.

50 competitors completed the marathon in less than 158 minutes.

b
Marathon runners' times


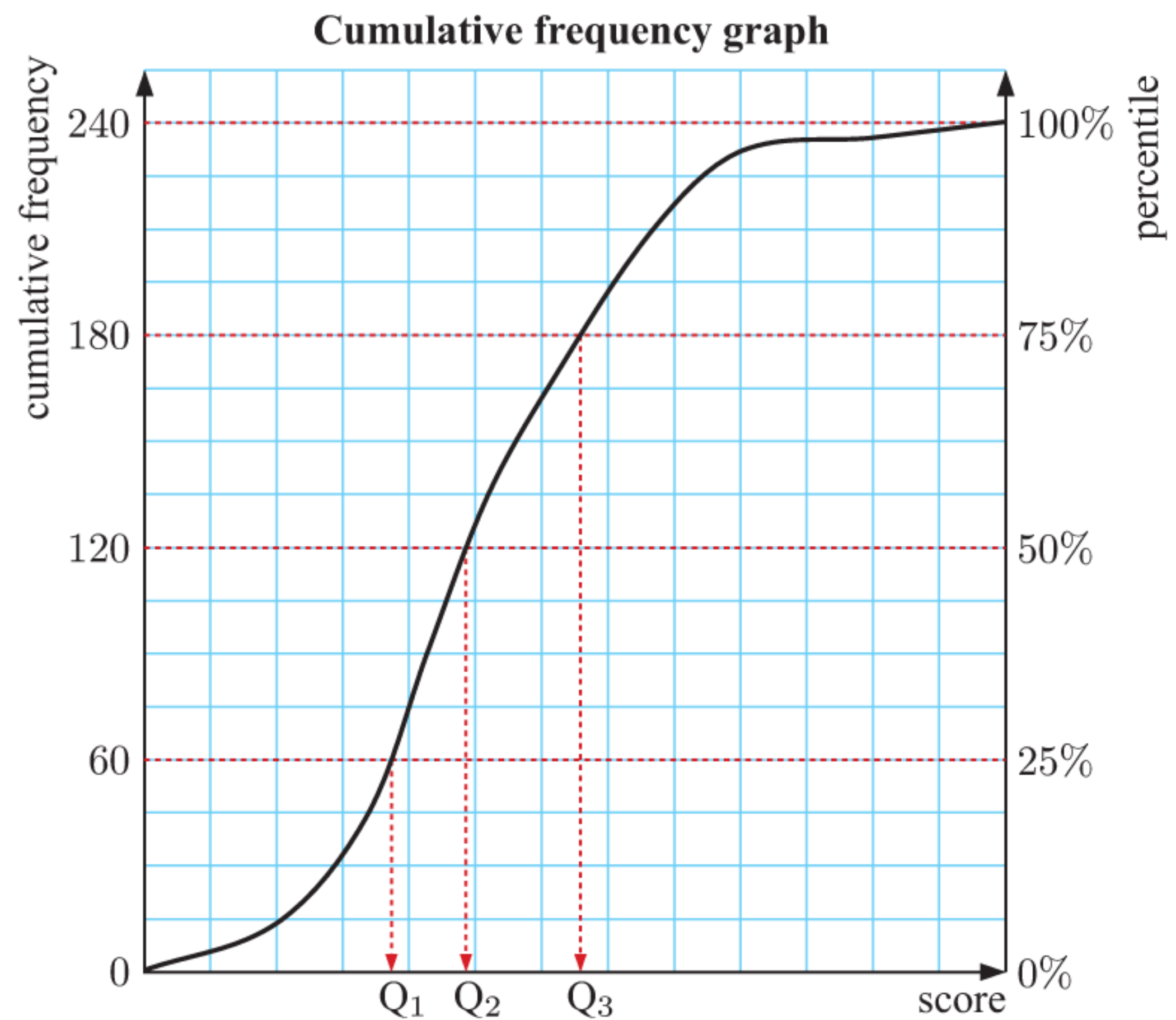
The cumulative frequency gives a *running total* of the number of runners finishing by a given time.



- c**
- i** The median is the 50th percentile. As 50% of 69 is 34.5, we start with the cumulative frequency 34.5 and find the corresponding time.
The median ≈ 154.5 min.
 - ii** Approximately 37 competitors took less than 155 min to complete the race.
 - iii** $69 - 52 = 17$ competitors took more than 159 min.
 $\therefore \frac{17}{69} \approx 24.6\%$ took more than 159 min.
 - iv** As 20% of 69 is 13.8, we start with the cumulative frequency 14 and find the corresponding time.
The top 20% of competitors took less than 151 min.

Another way to calculate percentiles is to add a separate scale to the cumulative frequency graph.

For example, 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 13I

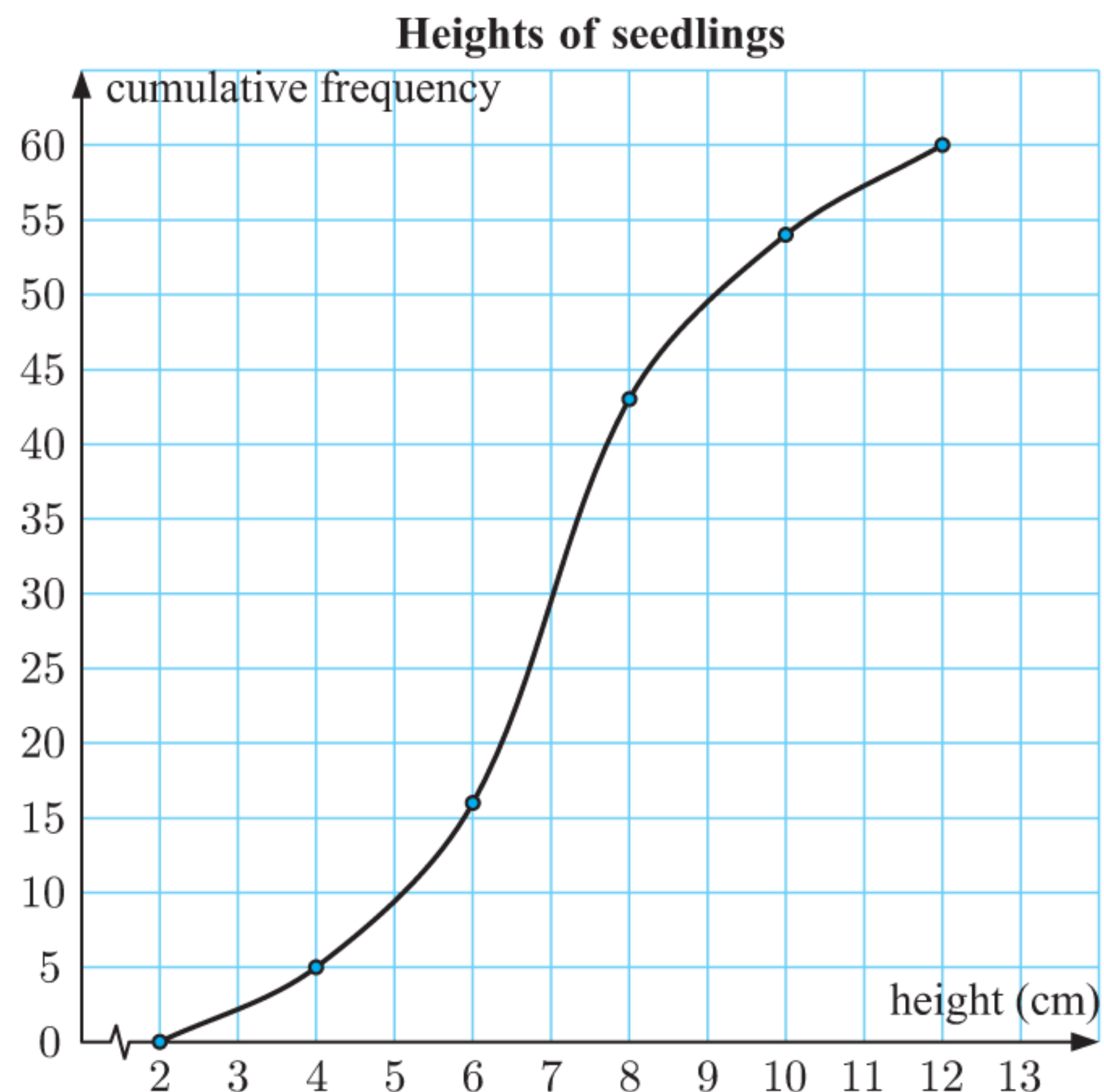
- 1 The examination scores of a group of students are shown in the table.

- a Draw a cumulative frequency graph for the data.
- b Find the median examination mark.
- c How many students scored 65 marks or less?
- d How many students scored at least 50 but less than 70 marks?
- e If the pass mark was 45, how many students failed?
- f If the top 16% of students were awarded credits, what was the credit mark?

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

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

- 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 Find the 90th percentile for the data and explain what this value represents.



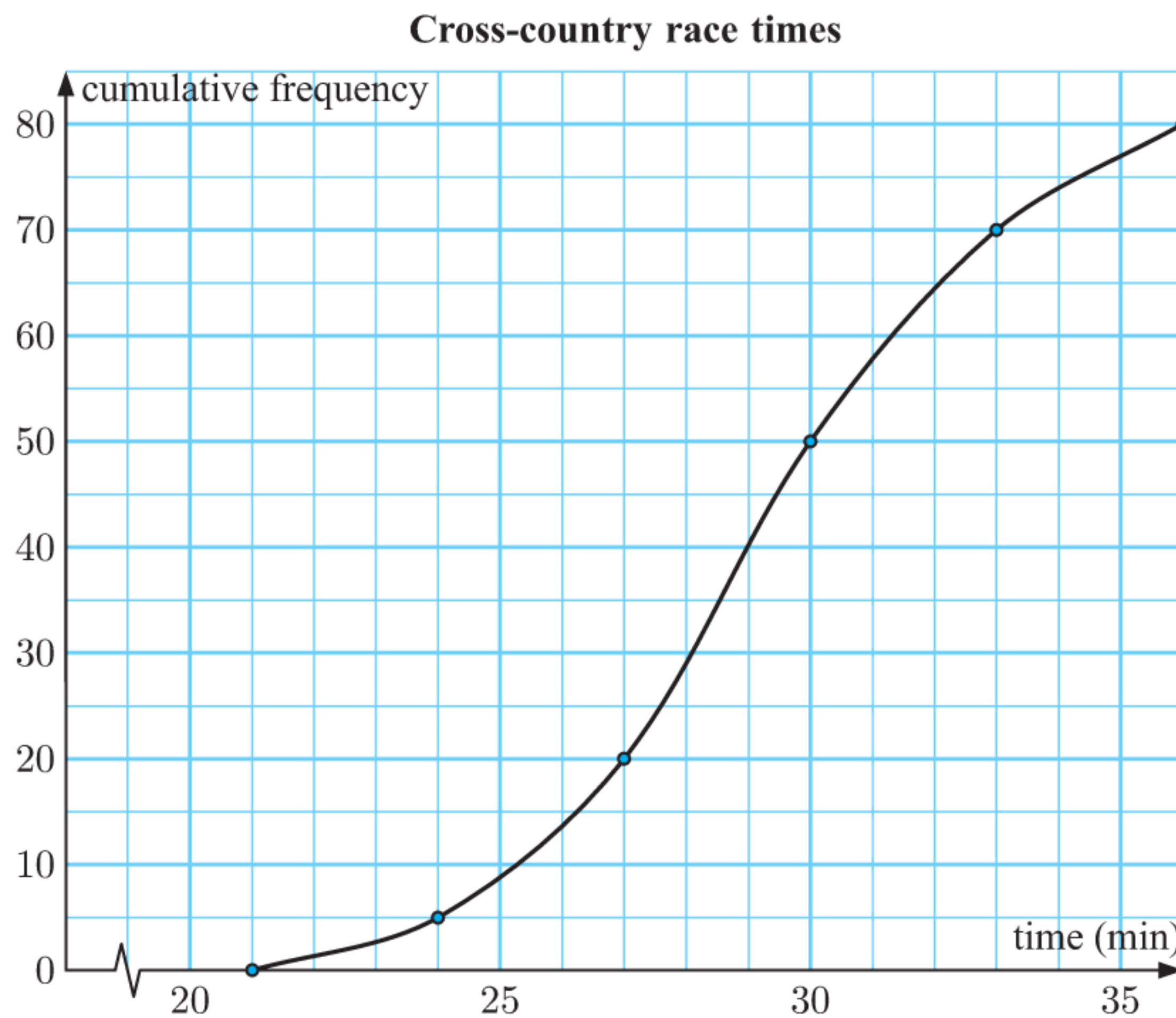
- 3** The following table summarises the age groups of car drivers involved in accidents in a city for a given year.
- a** Draw a cumulative frequency graph for the data.
 - b** Estimate the median age of the drivers involved in accidents.
 - c** Estimate the percentage of drivers involved in accidents who had an age of 23 or less.
 - d** Estimate the probability that a driver involved in an accident is aged:
 - i** 27 years or less
 - ii** 27 years.

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

- 4** The following data are 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 $24 \leq x < 27$, $27 \leq x < 30$, 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**.
- 5** The following cumulative frequency graph displays the performances of 80 competitors in a cross-country race.



- a** Find the lower quartile.
- b** Find the median.
- c** Find the upper quartile.
- d** Find the IQR.
- e** Estimate the 40th percentile.

- f** Use the cumulative frequency curve to complete the following table:

Time (t min)	$21 \leq t < 24$	$24 \leq t < 27$	$27 \leq t < 30$	$30 \leq t < 33$	$33 \leq t < 36$
Number of competitors					

- 6 The table shows the lifetimes of a sample of electric light globes.
- Draw a cumulative frequency graph for the data.
 - Estimate the median life of a globe.
 - Estimate the percentage of globes which had a life of 2700 hours or less.
 - Estimate the number of globes which had a life between 1500 and 2500 hours.

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

- 7 The following frequency distribution was obtained by asking 50 randomly selected people to measure the length of their feet. Their answers are 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

- Between what limits are lengths rounded to 20 cm?
- Rewrite the frequency table to show the data in the class intervals you have just described.
- Hence draw a cumulative frequency graph for the data.
- Estimate:
 - the median foot length
 - the number of people with foot length 26 cm or more.

J

VARIANCE AND STANDARD DEVIATION

The problem with using the range and the IQR as measures of spread or dispersion is that both of them only use two values in their calculation. As a result, some data sets can have their spread characteristics hidden when only the range or IQR are quoted.

So we need to consider alternative measures of spread which take into account all data values of a data set. We therefore turn to the **variance** and **standard deviation**.

POPULATION VARIANCE AND STANDARD DEVIATION

The **population variance** of a data set $\{x_1, x_2, x_3, \dots, x_n\}$ is

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$$

where μ is the population mean
and n is the number of data values.

The variance is the average of the squares of the distances from the mean.



We observe that if the data values x_i are situated close together around the mean μ , then the values $(x_i - \mu)^2$ will be small, and so the variance will be small.

The **standard deviation** is the square root of the variance.

The **population standard deviation** of a data set $\{x_1, x_2, x_3, \dots, x_n\}$ is

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n}}$$

The standard deviation measures the degree to which the data *deviates* from the mean.



The square root in the standard deviation is used to correct the units. For example, if x_i is the weight of a student in kg, the variance σ^2 would be in kg^2 , and σ would be in kg.

The standard deviation is a **non-resistant** measure of spread. This is due to its dependence on the mean and because extreme data values will give large values for $(x_i - \mu)^2$. It is only a useful measure if the distribution is approximately symmetrical.

The IQR and percentiles are more appropriate tools for measuring spread if the distribution is considerably skewed.

Example 12

Self Tutor

Find the population variance and standard deviation for the data set: 3 12 8 15 7

The mean $\mu = \frac{3 + 12 + 8 + 15 + 7}{5} = 9$

The population variance $\sigma^2 = \frac{\sum (x - \mu)^2}{n}$
 $= \frac{86}{5}$
 $= 17.2$

The population standard deviation $\sigma = \sqrt{17.2}$
 ≈ 4.15

x	$x - \mu$	$(x - \mu)^2$
3	-6	36
12	3	9
8	-1	1
15	6	36
7	-2	4
<i>Total</i>		86

SAMPLE VARIANCE AND STANDARD DEVIATION

If we are only given a *sample* of data from a larger population, we calculate a statistic called the **sample variance**. This statistic is used to *estimate* the variance of the population.

For a sample of n data values $\{x_1, x_2, x_3, \dots, x_n\}$ with sample mean \bar{x} :

- The **sample variance** is $s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$.

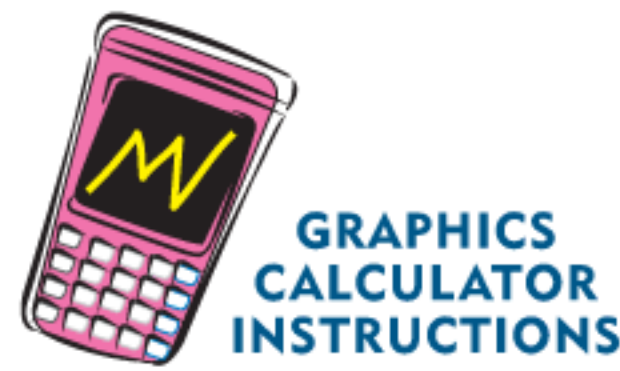
- The **sample standard deviation** is $s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$.

The sample variance s^2 is an **unbiased estimate** of the population variance σ^2 .



CALCULATING VARIANCE AND STANDARD DEVIATION

We commonly use technology to find the standard deviation. If we are given the whole population we use σ_x , and if we are given a sample we use s_x . We can then square these values to find the population and sample variance, σ_x^2 and s_x^2 respectively.



	Deg	Norm1	d/c	Real
1-Variable				
\bar{x}	=	30.0333333		
Σx	=	901		
Σx^2	=	33211		
σx	=	14.3189462		
$s x$	=	14.5637323		
n	=	30		

Example 13

Self Tutor

Kylie is interested in the ages of spectators at a rugby match. She selects a sample of 30 spectators. Their ages are shown below:

17 24 30 10 42 48 37 19 28 53 29 40 11 21 9
43 22 59 46 52 31 13 7 26 32 47 22 15 26 42

Use technology to find the sample standard deviation.

Casio fx-CG50

	Deg	Norm1	d/c	Real
1-Variable				
\bar{x}	=	30.0333333		
Σx	=	901		
Σx^2	=	33211		
σx	=	14.3189462		
$s x$	=	14.5637323		
n	=	30		

TI-84 Plus CE

NORMAL FLOAT AUTO REAL RADIAN MP	
1-Var Stats	
\bar{x}	=30.03333333
Σx	=901
Σx^2	=33211
$s x$	=14.56373231
σx	=14.31894627
n	=30
minX	=7
↓Q1	=19

HP Prime

Statistics 1Var Numeric View	
H1	
Min	7
Q1	19
Med	28.5
Q3	42
Max	59
Σx	901
Σx^2	33,211
\bar{x}	30.03333333
$s x$	14.5637323118
σx	14.3189462679
Sample standard deviation of X	
More	OK

The sample standard deviation $s \approx 14.6$ years.

The reason there are two formulae for standard deviation is that if we have data which is sampled from a large population, the sample standard deviation s provides a better estimate for the actual population standard deviation σ than if we use the formula for σ on the sample.

However, in the **Mathematics: Analysis and Approaches HL** course you are expected to calculate all standard deviations as though they were populations. For this reason, two answers are given for some questions in the following Exercise.

EXERCISE 13J

- 1 Consider the following data sets:

Data set A: 10 7 5 8 10

Data set B: 4 12 11 14 1 6

- Show that each data set has mean 8.
- Which data set appears to have the greater spread? Explain your answer.
- Find the population variance and standard deviation of each data set. Use technology to check your answers.

- 2 Skye recorded the number of pets owned by each student in her class.

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

- a Describe the population in this case.
 - b Use technology to find the population standard deviation of the data.
 - c Find the population variance of the data.
- 3 The ages of members of an Olympic water polo team are: 22, 25, 23, 28, 29, 21, 20, 26.
- a Calculate the mean and population standard deviation for this group.
 - b The same team members are chosen to play in the next Olympic Games 4 years later. Calculate the mean and population standard deviation of their ages at the next Olympic Games.
 - c Comment on your results in general terms.

- 4 A hospital selected a sample of 20 patients and asked them how many glasses of water they had consumed that day. The results were:

5 2 1 0 4 1 0 2 7 4
8 2 7 6 1 2 3 8 0 2

Find the standard deviation of the data.

Mathematics: Analysis and Approaches students should use the population standard deviation.



- 5 Danny and Jennifer recorded how many hours they spent on homework each day for 14 days.

Danny: $3\frac{1}{2}$, $3\frac{1}{2}$, 4, $2\frac{1}{2}$, 3, $3\frac{1}{2}$, 3, $1\frac{1}{2}$, 3, 4, $2\frac{1}{2}$, 4, 4, 3

Jennifer: $2\frac{1}{2}$, 1, $2\frac{1}{2}$, 2, 2, $2\frac{1}{2}$, $1\frac{1}{2}$, 2, 2, $2\frac{1}{2}$, 2, 2, 2, $1\frac{1}{2}$

- a Calculate the mean number of hours each person spent on homework.
 - b Which person generally studies for longer?
 - c Calculate the standard deviation for each data set.
 - d Which person studies more consistently?
- 6 Tyson wants to compare the swimming speeds of boys and girls at his school. He randomly selects 10 boys and 10 girls, and records the time, in seconds, each person takes to swim two laps of the 25 m school pool.

Boys: 32.2, 26.4, 35.6, 30.8, 28.5, 40.2, 27.3, 38.9, 29.0, 31.3

Girls: 36.2, 33.5, 28.1, 39.8, 31.6, 35.7, 37.3, 36.0, 39.7, 29.8

- a Copy and complete the table:

	Boys	Girls
<i>Mean \bar{x}</i>		
<i>Median</i>		
<i>Standard deviation</i>		
<i>Range</i>		



- b Which group:
 - i generally swims faster
 - ii has the greater spread of swimming speeds?
- c How could Tyson improve the reliability of his findings?

7 Two baseball coaches compare the number of runs scored by their teams in their last ten games:

<i>Rockets</i>	0	10	1	9	11	0	8	5	6	7
<i>Bullets</i>	4	3	4	1	4	11	7	6	12	5

- a Show that the two teams have the same mean and range of runs scored.
- b Which team's performance do you suspect is more variable? Check your answer by finding the standard deviation for each data set.
- c Does the range or the standard deviation give a better indication of variability?

8 The number of visitors to a museum and an art gallery each day during December are shown.

<i>Museum:</i>	1108	1019	850	1243	1100	923	964	847	918	820	781
	963	814	881	742	911	1101	952	864	943	1087	1132
	906	1050	0	826	986	1040	1127	1084	981		
<i>Art gallery:</i>	1258	1107	1179	1302	1236	1386	1287	1313	1269	1332	1094
	1153	1275	1168	1086	1276	1342	1153	1227	1305	1187	1249
	1300	1156	1074	1168	1299	1257	1134	1259	1366		

- a For each data set, calculate the:
 - i mean
 - ii standard deviation.
- b Which place had the greater spread of visitor numbers?
- c
 - i Identify the outlier in the *Museum* data.
 - ii Give a reason why this outlier may have occurred.
 - iii Do you think it is reasonable to remove the outlier when comparing the numbers of visitors to these places? Explain your answer.
 - iv Recalculate the mean and standard deviation with the outlier removed.
 - v Discuss the effect of the outlier on the standard deviation.

9 Anna and Apple want to compare the distributions of weights from two different populations A and B, so they take samples from each population. Apple uses the sample standard deviation formula, and finds that $s_A > s_B$. Anna uses the formula for the population standard deviation on the samples. Will Anna find that $\sigma_A > \sigma_B$ for the samples? Explain your answer.

10 A set of 8 integers $\{1, 3, 5, 7, 4, 5, p, q\}$ has mean 5 and population variance 5.25. Find p and q given that $p < q$.

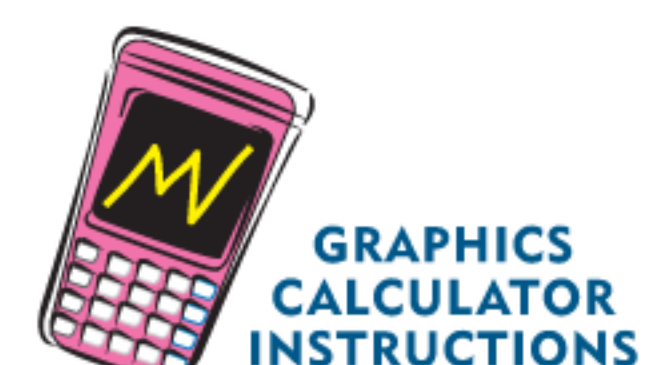
11 A set of 10 integers $\{3, 9, 5, 5, 6, 4, a, 6, b, 8\}$ has mean 6 and population variance 3.2. Find a and b given that $a > b$.

12 a Prove that $\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n (x_i^2) - n\mu^2$.

b The data set $\{x_1, x_2, \dots, x_{25}\}$ has $\sum_{i=1}^{25} x_i^2 = 2568.25$ and population standard deviation 5.2. Find the mean of the data set.

13 Find the population standard deviation of this data set.

Value	Frequency
3	1
4	3
5	11
6	5

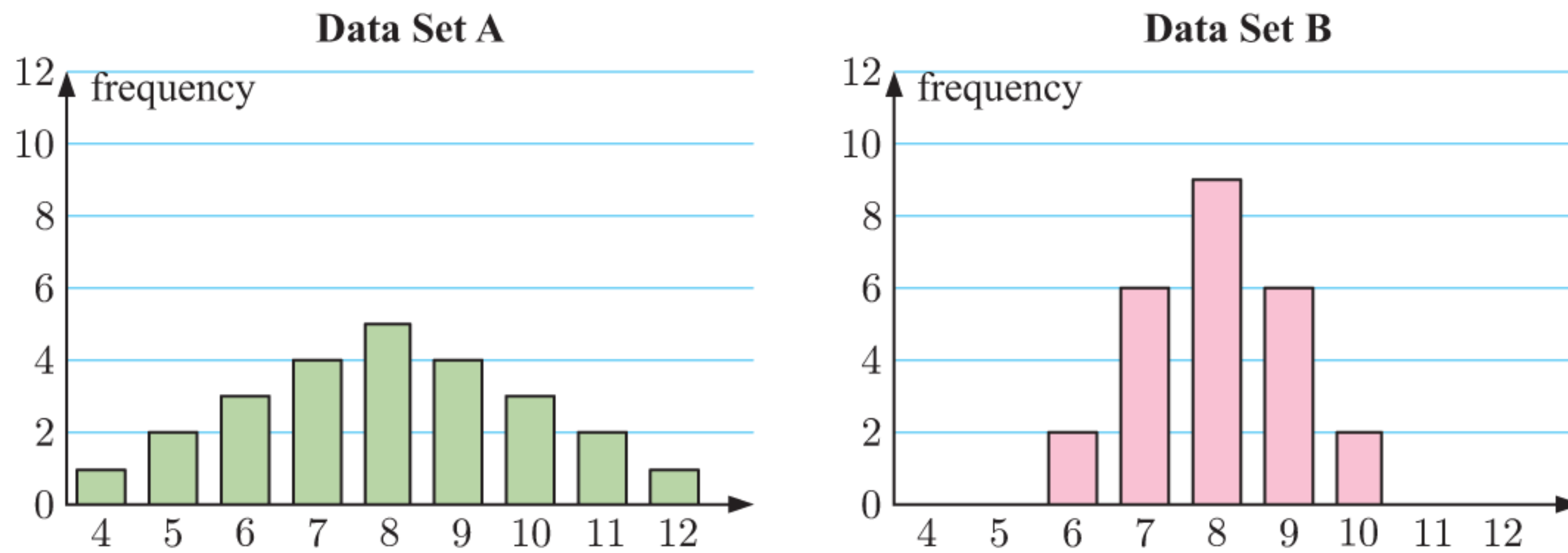


14 The table shows the ages of squash players at the Junior National Squash Championship.

<i>Age</i>	11	12	13	14	15	16	17	18
<i>Frequency</i>	2	1	4	5	6	4	2	1

Find the mean and population standard deviation of the ages.

15 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 data set.
- c Find the population standard deviation for each data set. Comment on your answers.
- d The other measures of spread for the two data sets are given in the table.

<i>Data set</i>	<i>Range</i>	<i>IQR</i>
A	8	3
B	4	2

In what way does the standard deviation give a better description of how the data is distributed?

16 The table alongside shows the results obtained by female and male students in a test out of 20 marks.

- a Looking at the table:
 - i Which group appears to have scored better in the test?
 - ii Which group appears to have a greater spread of scores?
 Justify your answers.
- b Calculate the mean and population standard deviation for each group.

<i>Score</i>	<i>Females</i>	<i>Males</i>
12	0	1
13	0	0
14	0	2
15	0	3
16	2	4
17	6	2
18	5	0
19	1	1
20	1	0

17 Brianna and Jess are conducting a survey of their class. Brianna asked every student (including Jess and herself) how many children there are in their family. Jess asked every student (including Brianna and herself) how many siblings or step-siblings they have. How will their results compare in terms of mean and standard deviation? Explain your answer.



Example 14**Self Tutor**

The table alongside summarises the examination scores for 80 randomly selected students. Estimate the standard deviation for the data.

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

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

For continuous data or data grouped in classes, use the mid-interval value to represent all data in that interval.

**Casio fx-CG50**

1-Variable	
\bar{x}	=59.75
Σx	=4780
Σx^2	=308200
σx	=16.8058769
sx	=16.9119087
n	=80

TI-84 Plus CE

1-Var Stats	
\bar{x}	=59.75
Σx	=4780
Σx^2	=308200
Sx	=16.91190877
σx	=16.80587695
n	=80
minX	=4.5
$\downarrow Q_1$	=54.5

TI-nspire

OneVar data:freq: stat.results	
"Title"	"One-Variable Statistics"
" \bar{x} "	59.75
" Σx "	4780.
" Σx^2 "	308200.
" $Sx := S_{n-1}X$ "	16.9119
" $\sigma x := \sigma_n X$ "	16.8059
" n "	80.
"MinX"	4.5

The sample standard deviation ≈ 16.9 .

The population standard deviation ≈ 16.8 .

- 18** The lengths of 30 randomly selected 12-day old babies were measured and the following data obtained:

For the given data, estimate the:

- a** mean **b** standard deviation.

Length (L cm)	Frequency
$40 \leq L < 42$	1
$42 \leq L < 44$	1
$44 \leq L < 46$	3
$46 \leq L < 48$	7
$48 \leq L < 50$	11
$50 \leq L < 52$	5
$52 \leq L < 54$	2

c Check your answer by:

i multiplying each value by 9

ii dividing each value by 4.

4 Suppose a data set $\{x_i\}$ has mean μ and standard deviation σ . Write down the mean and standard deviation for the data set:

a $\{ax_i\}$

b $\{x_i + k\}$

c $\{ax_i + k\}$

INVESTIGATION 4

ESTIMATING THE VARIANCE AND STANDARD DEVIATION OF A POPULATION

In this Investigation we consider the accuracy of using a sample to make inferences about a whole population. This will help you to see why statisticians have a subtly different formula for the standard deviation of a sample.

The Year 12 students at a school were asked to record how many minutes they spent travelling to school. The results were collected in a survey the following morning.

There are a total of 150 Year 12 students at the school, and these are split into 6 classes.

What to do:

1 Click on the icon to obtain a spreadsheet containing all of the responses to the survey.

SPREADSHEET



a Use the frequency table in the spreadsheet to draw a histogram for the data. Describe this distribution.

b The summary statistics in the spreadsheet are calculated using all of the survey responses, and hence are the *true* population values. Find the true population variance.

2 10 students were randomly selected from each class to form 6 samples. Their responses to the survey are shown below:

<i>Sample 1:</i>	10	14	16	9	16	15	15	21	9	21
<i>Sample 2:</i>	11	9	11	16	16	13	10	12	21	16
<i>Sample 3:</i>	12	10	14	7	13	11	21	20	15	9
<i>Sample 4:</i>	20	19	19	19	13	19	22	15	10	19
<i>Sample 5:</i>	19	13	23	11	17	4	14	21	13	11
<i>Sample 6:</i>	19	11	16	6	8	13	10	22	20	11

a Calculate the *sample* statistics s and s^2 for each sample.

b Calculate the *population* statistics σ and σ^2 for each sample.

c Which set of estimates from **a** and **b** are generally closer to the true population variance and standard deviation?

d Does your answer to **c** explain why we have different variance and standard deviation formulae for a sample as opposed to a population?

3 To see which set of estimators (population or sample) are better at estimating the true population variance and standard deviation, we will consider a simulation based on the survey responses from the school.

Click on the icon to obtain a spreadsheet with 1000 simulations of the survey results. The values s , s^2 , σ , and σ^2 are calculated for each simulated sample. The average values for each estimator are shown in the table on the sheet labelled “Summary”.

SPREADSHEET



	A	B	C	D	E	F
1	Actual values				Estimator	Average estimate
2	μ	15		Variance	σ^2	23.794
3	σ	5			s^2	25.046
4	σ^2	25		Standard deviation	σ	4.814
5	n	20			s	4.939

Based on the calculations in the spreadsheet, which set of estimates (population or sample) are generally closer to the true values? Does your conclusion agree with your answer to **2 c**?

- 4 Change the values for μ and σ in the spreadsheet. This will now effectively simulate the results for a different distribution, perhaps the travel times for the students at a different school. Does your choice of μ or σ affect your conclusion regarding the choice of estimators?
- 5 Why is it important to have accurate estimates of the variance and standard deviation of a population?

REVIEW SET 13A

- 1 For each of the following data sets, find the: **i** mean **ii** median.
 - a 0, 2, 3, 3, 4, 5, 5, 6, 6, 7, 7, 8
 - b 2.9, 3.1, 3.7, 3.8, 3.9, 3.9, 4.0, 4.5, 4.7, 5.4
- 2 Katie loves cats. She visits every house in her street to find out how many cats live there. The responses are given below:

Number of cats	0	1	2	3	4	5
Frequency	36	9	11	5	1	1

- a Draw a graph to display this data.
- b Describe the distribution.
- c Find the:
 - i** mode
 - ii** mean
 - iii** median.
- d Which of the measures of centre is most appropriate for this data? Explain your answer.

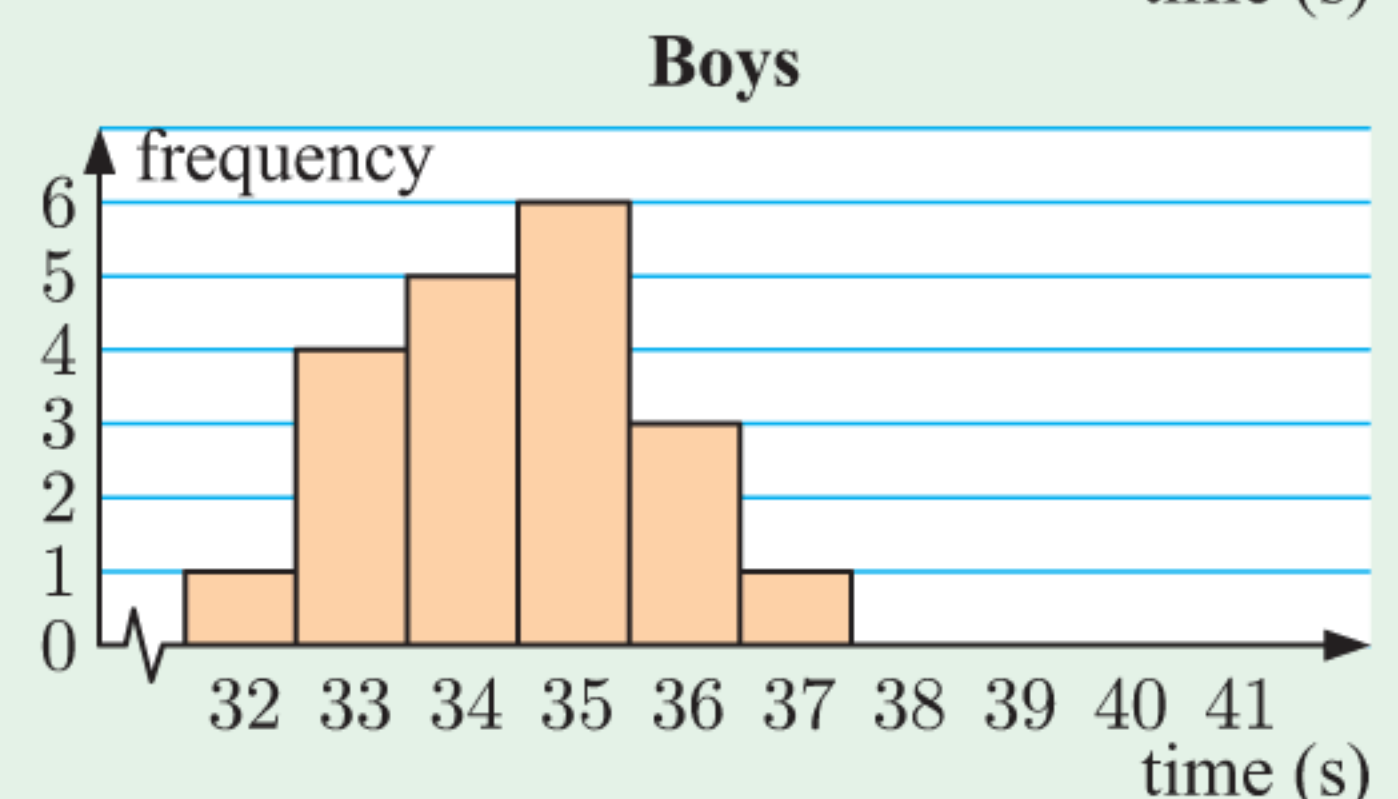
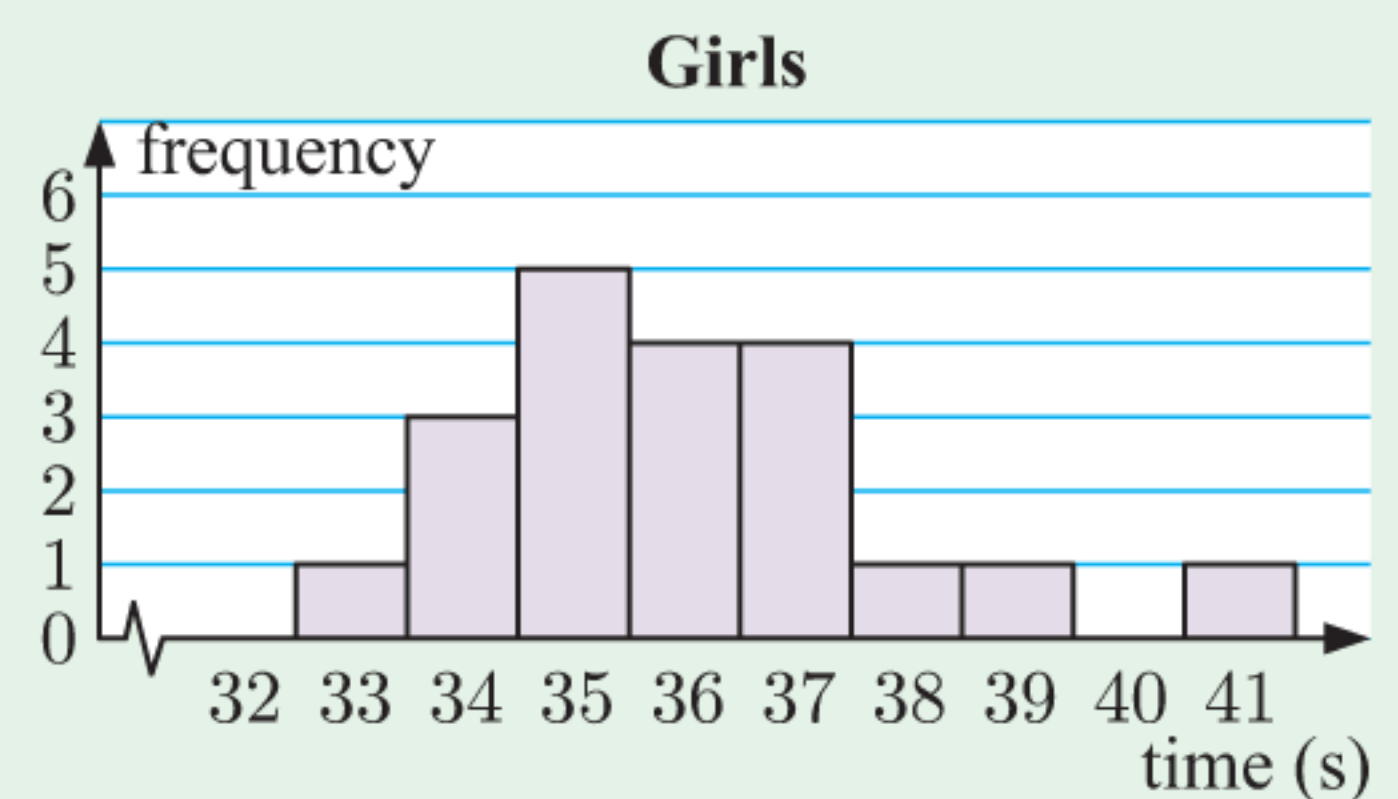


- 3 The histograms alongside show the times for the 50 metre freestyle recorded by members of a swimming squad.

a Copy and complete:

Distribution	Girls	Boys
median		
mean		
modal class		

- b Discuss the distributions of times for the boys and girls. What conclusion can you make?



- 4 The data set $4, 6, 9, a, 3, b$ has a mean and mode of 6. Find the values of a and b given that $a > b$.
- 5 Consider the data set: $k - 2, k, k + 3, k + 3$.
- Show that the mean of the data set is equal to $k + 1$.
 - Suppose each number in the data set is increased by 2. Find the new mean of the data set in terms of k .

- 6 The winning margins in 100 basketball games were recorded. The results are summarised alongside.

Margin (points)	Frequency
1 - 10	13
11 - 20	35
21 - 30	27
31 - 40	18
41 - 50	7

- Explain why you cannot calculate the mean winning margin from the table exactly.
- Estimate the mean winning margin.

- 7 Consider this data set:

19, 7, 22, 15, 14, 10, 8, 28, 14, 18, 31, 13, 18, 19, 11, 3, 15, 16, 19, 14

- Find the five-number summary for the data.
- Find the range and IQR.
- Draw a box plot of the data set.

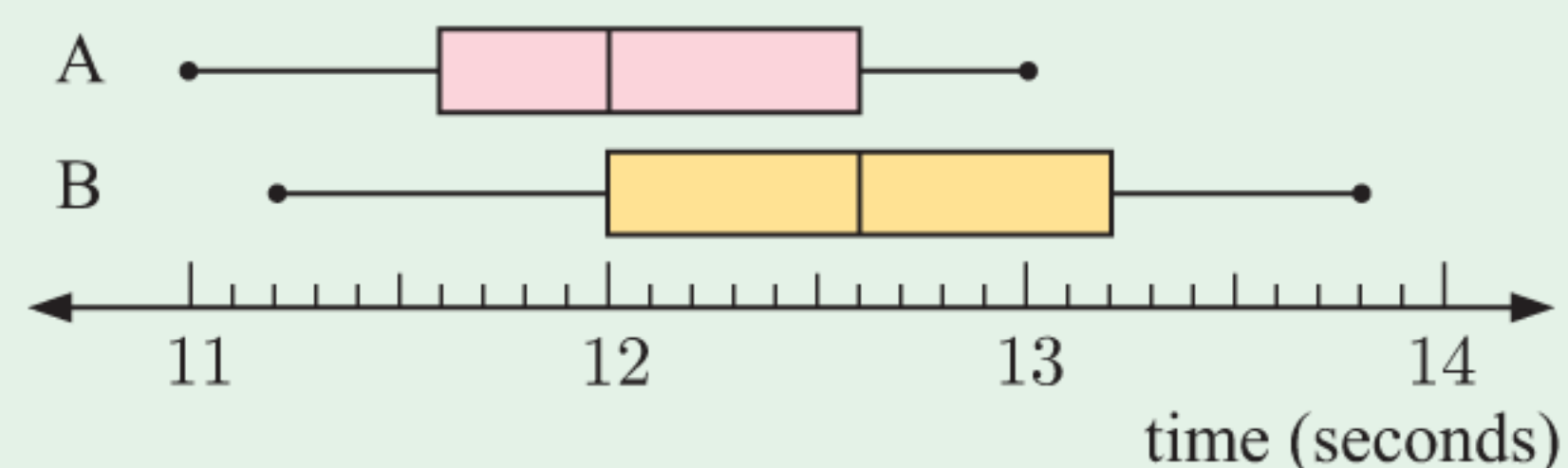
- 8 Katja's golf scores for her last 20 rounds were:

90 106 84 103 112 100 105 81 104 98
107 95 104 108 99 101 106 102 98 101

For this data set, find the:

- median
- interquartile range
- mean
- standard deviation.

- 9 The parallel box plot alongside shows the 100 metre sprint times for the members of two athletics squads.

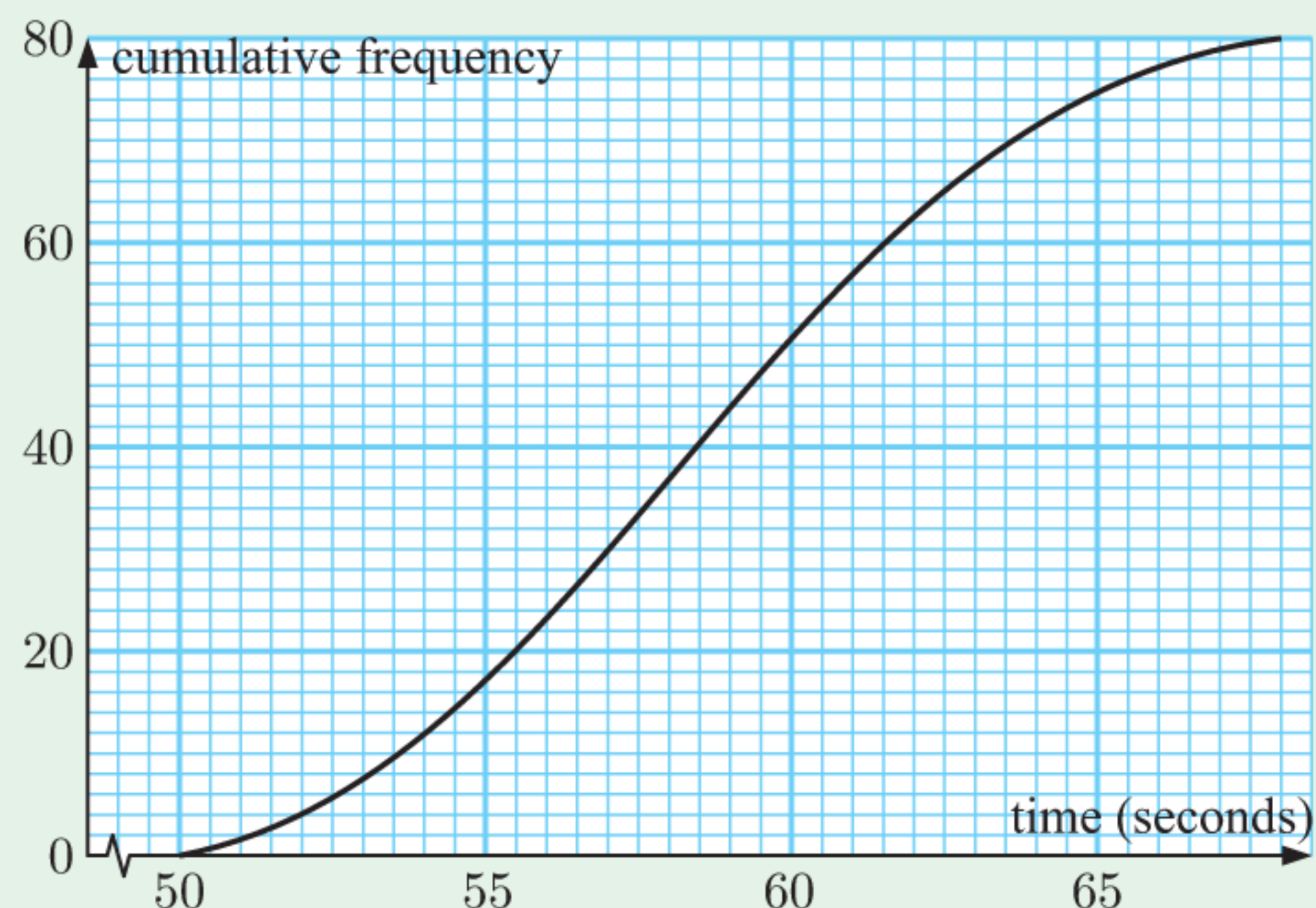


- Determine the five-number summaries for both A and B.
- For each group, calculate the range and interquartile range.
- Copy and complete:
 - The members of squad generally ran faster because
 - The times in squad are more varied because

- 10 80 senior students ran 400 metres in a Physical Education program. Their times were recorded and the results were used to produce the cumulative frequency graph shown.

Estimate:

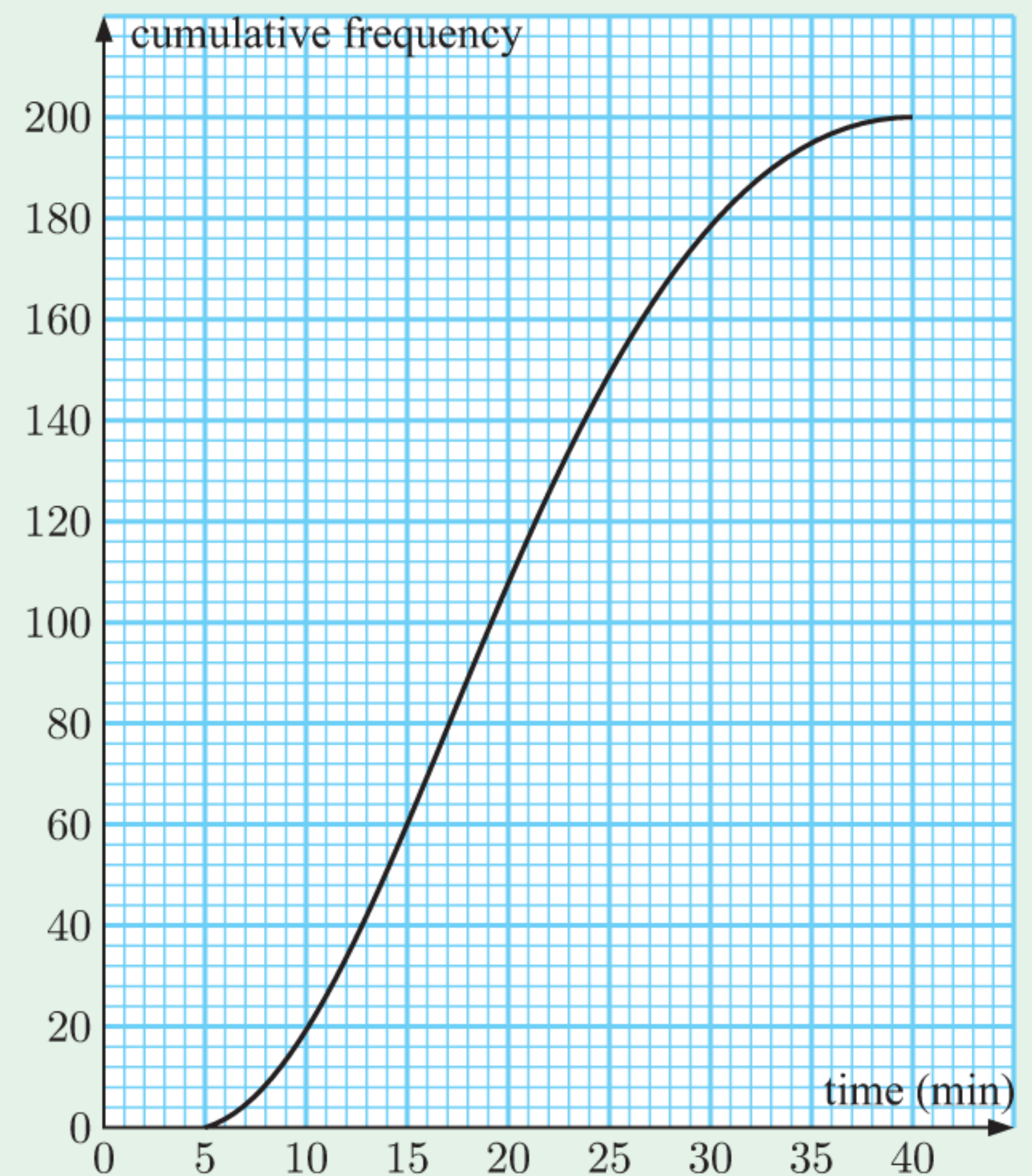
- the median
- the interquartile range
- the time corresponding to the top 10% of runners.



11 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 .
- c** Use the cumulative frequency curve to complete the following table:

Time (t min)	Frequency
$5 \leq t < 10$	
$10 \leq t < 15$	
⋮	
$35 \leq t < 40$	



12 Find the population variance and standard deviation for each data set:

- 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

13 The table alongside shows the number of matches in a sample of boxes.

<i>Number</i>	47	48	49	50	51	52
<i>Frequency</i>	21	29	35	42	18	31

- a** Find the mean and standard deviation for this data.
- b** Does this result justify a claim that the average number of matches per box is 50?

14 The number of litres of petrol purchased by a random sample of motor vehicle drivers is shown alongside. For the given data, estimate the:

Litres (L)	Number of vehicles
$15 \leq L < 20$	5
$20 \leq L < 25$	13
$25 \leq L < 30$	17
$30 \leq L < 35$	29
$35 \leq L < 40$	27
$40 \leq L < 45$	18
$45 \leq L < 50$	7

- a** mean
- b** standard deviation.

15 Pratik is a quality control officer for a biscuit company. He needs to check that 250 g of biscuits go into each packet, but realises that the weight in each packet will vary slightly.

- a** Would you expect the standard deviation for the whole population to be the same for one day as it is for one week? Explain your answer.
- b** If a sample of 100 packets is measured each day, what measure would be used to check:
 - i** that an average of 250 g of biscuits goes into each packet
 - ii** the variability of the mass going into each packet?
- c** Explain the significance of a low standard deviation in this case.

REVIEW SET 13B

- 1 Heike is preparing for an athletics carnival. She records her times in seconds for the 100 m sprint each day for 4 weeks.

Week 1: 16.4 15.2 16.3 16.3 17.1 15.5 14.9

Week 2: 14.9 15.7 15.1 15.1 14.7 14.7 15.3

Week 3: 14.3 14.2 14.6 14.6 14.3 14.3 14.4

Week 4: 14.0 14.0 13.9 14.0 14.1 13.8 14.2

- a Calculate Heike's mean and median time for each week.
b Do you think Heike's times have improved over the 4 week period? Explain your answer.

- 2 A die was rolled 50 times.
The results are shown in the table alongside.
Find the:

Number	Frequency
1	10
2	7
3	8
4	5
5	12
6	8

- a mode b mean c median.

- 3 The data in the table alongside has mean 5.7.

<i>Value</i>	2	5	x	$x + 6$
<i>Frequency</i>	3	2	4	1

- a Find the value of x .
b Find the median of the distribution.

- 4 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 .

- 5 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

- 6 Draw a box and whisker diagram for the following data:
11, 12, 12, 13, 14, 14, 15, 15, 15, 16, 17, 17, 18.
- 7 Consider the data set: 120, 118, 132, 127, 135, 116, 122, 93, 128.
- a Find the standard deviation for the data.
b Find the upper and lower quartiles of the data set.
c Are there any outliers in the data set?
d Draw a box plot to display the data.

- 8** 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.

<i>Brand X</i>						<i>Brand Y</i>					
871	885	878	882	889	885	909	906	913	891	898	901
916	913	886	905	907	898	894	894	928	893	924	892
874	904	901	894	897	899	927	907	901	900	907	913
908	901	898	894	895	895	921	904	903	896	901	895
910	904	896	893	903	888	917	903	910	903	909	904

- a** Copy and complete the table alongside.
b Display the data on a parallel box plot.
c Comment on which brand:
i has more peanuts per jar
ii has a more consistent number of peanuts per jar.

	<i>Brand X</i>	<i>Brand Y</i>
min		
Q ₁		
median		
Q ₃		
max		
IQR		

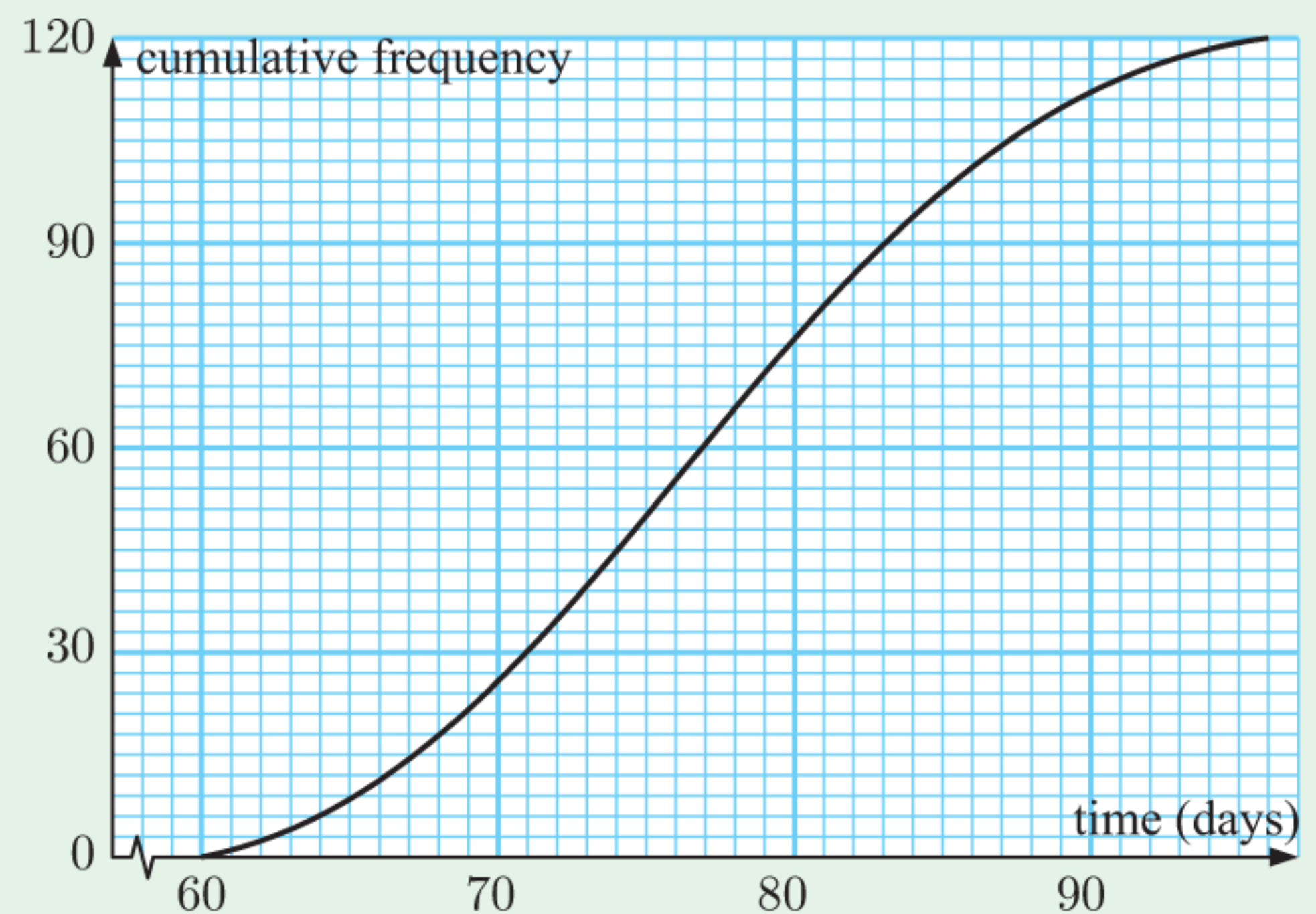
- 9** Consider the frequency table alongside:

- a** Find the values of p and m .
b Hence find the mode, median, and range of the data.
c Given that $\sum_{i=1}^5 x_i f_i = 254$, write the mean \bar{x} as a fraction.

<i>Score</i>	<i>Frequency</i>	<i>Cumulative frequency</i>
6	2	2
7	4	m
8	7	13
9	p	25
10	5	30

- 10** 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.



- 11** Consider the data in the table below:

<i>Scores (x)</i>	$0 \leq x < 10$	$10 \leq x < 20$	$20 \leq x < 30$	$30 \leq x < 40$	$40 \leq x < 50$
<i>Frequency</i>	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.

12 To test the difficulty level of a new computer game, a company measures the time taken for a group of players to complete the game. Their results are displayed in the table alongside.

Completion time (t min)	Number of players
$0 \leq t < 30$	1
$30 \leq t < 60$	4
$60 \leq t < 90$	12
$90 \leq t < 120$	18
$120 \leq t < 150$	7
$150 \leq t < 180$	2

- How many players were surveyed?
- Write down the modal class.
- Draw a cumulative frequency graph for the data.
- The game is considered too easy if either the mean or median completion time is below 90 minutes.
 - Estimate the median completion time using your cumulative frequency graph.
 - Estimate the mean completion time.
 - Hence comment on whether the game is too easy.
- Complete this sentence:
The middle 50% of players completed the game in times between and minutes.

13 A random sample of weekly supermarket bills was recorded in the table alongside.

For the given data, estimate the:

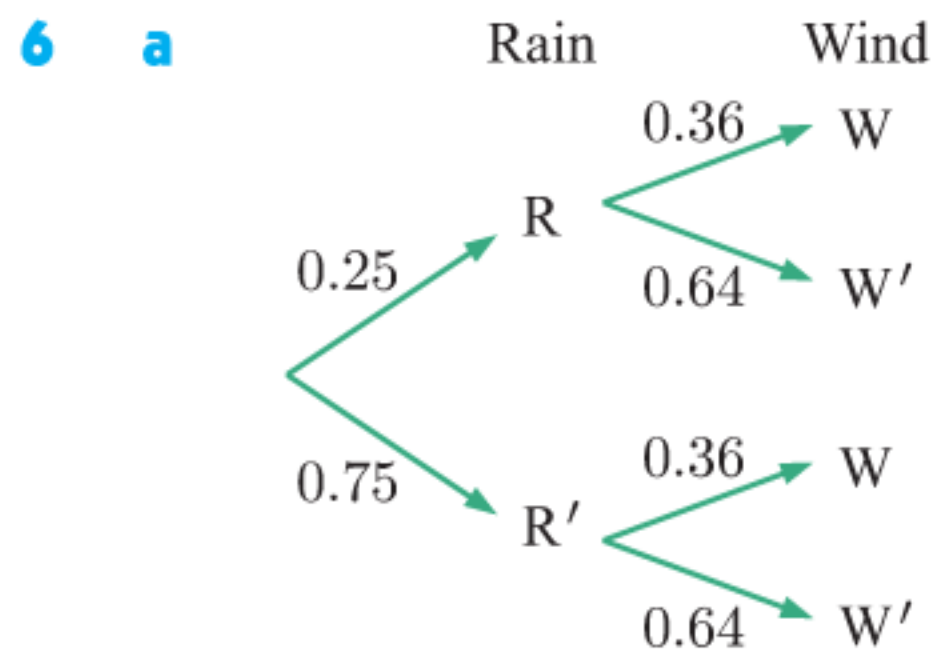
- mean
- standard deviation.

Bill (€ b)	Number of families
$140 \leq b < 160$	27
$160 \leq b < 180$	32
$180 \leq b < 200$	48
$200 \leq b < 220$	25
$220 \leq b < 240$	37
$240 \leq b < 260$	21
$260 \leq b < 280$	18
$280 \leq b < 300$	7

14 Friends Kevin and Felicity each selected a sample of 20 crossword puzzles. The times they took, in minutes, to complete each puzzle were:

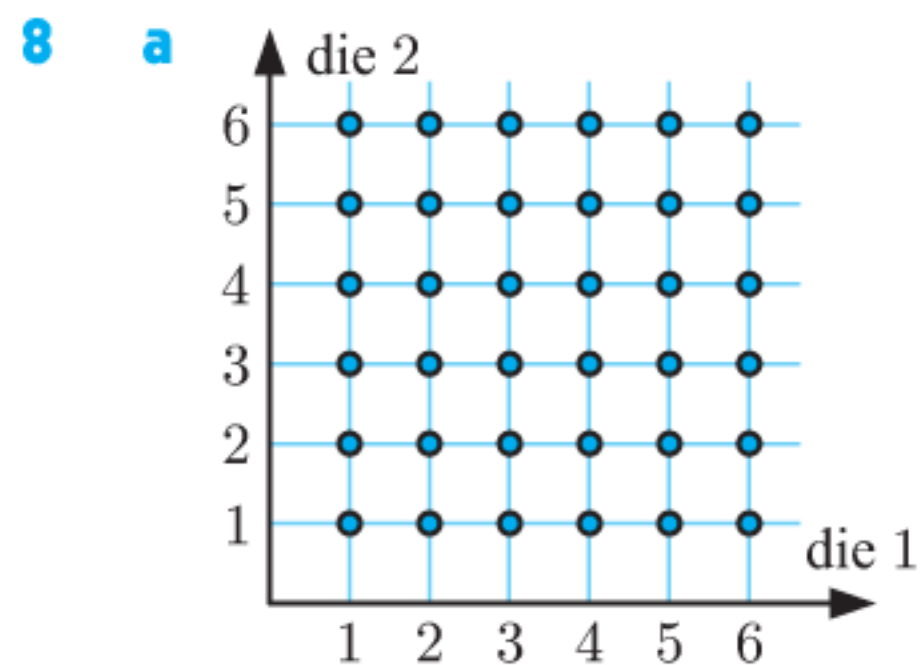
Kevin					Felicity				
37	53	47	33	39	33	36	41	26	52
49	37	48	32	36	38	49	57	39	44
39	42	34	29	52	48	25	34	27	53
48	33	56	39	41	38	34	35	50	31

- Find the mean of each data set.
 - Find the standard deviation for each person.
 - Who generally solves crossword puzzles faster?
 - Who is more consistent in their time taken to solve the puzzles?
- 15** A data set has $s^2 = 4.1$ and $\sigma^2 = 3.69$. How many data values does the data set have?

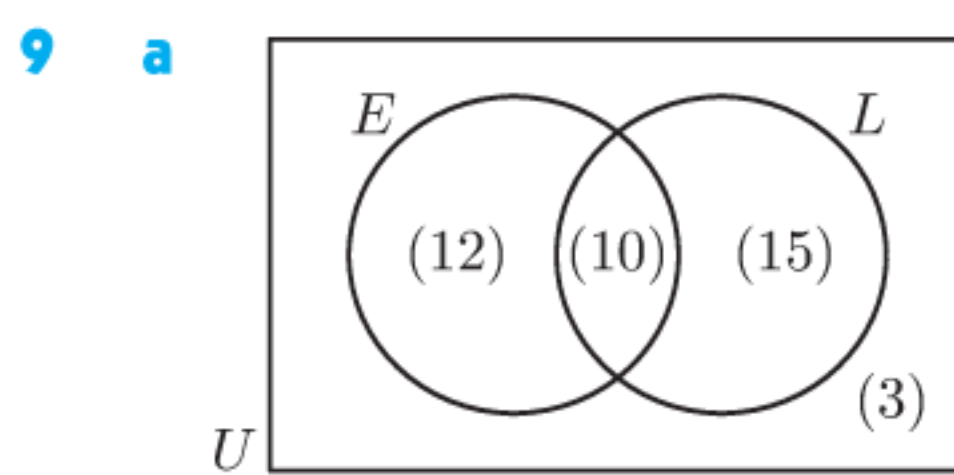


- b** i 0.09
 ii 0.52
c It is assumed that the events are independent.

- 7 a** 0 **b** 0.45 **c** 0.8



- b** i $\frac{2}{9}$
 ii $\frac{5}{12}$



- b** i $\frac{1}{4}$
 ii $\frac{37}{40}$
 iii $\frac{2}{5}$

- 10** 4350 seeds **11 a** $\frac{25}{144}$ **b** $\frac{25}{72}$ **c** $\frac{7}{16}$ **d** $\frac{4}{9}$

- 12** ≈ 0.127

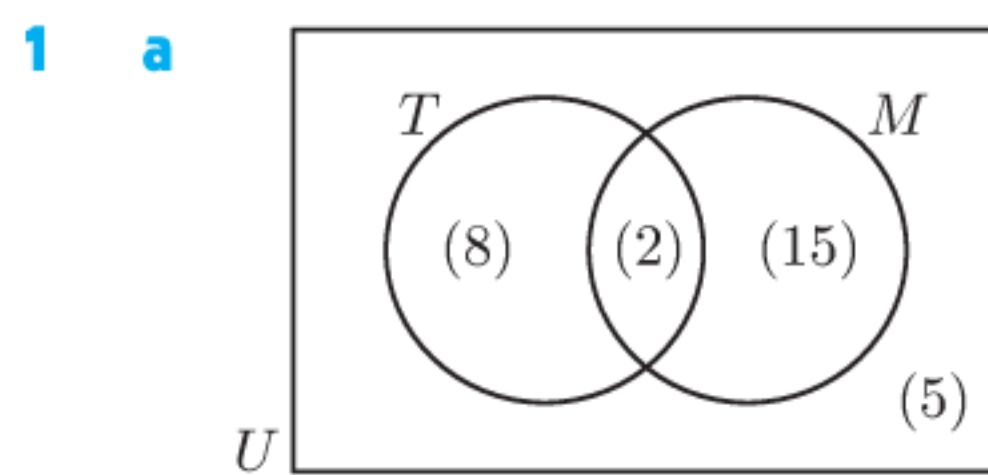
13 a

	Female	Male	Total
Smoker	20	40	60
Non-smoker	70	70	140
Total	90	110	200

- b** i $\frac{7}{20}$
 ii $\frac{1}{2}$
c i ≈ 0.121
 ii ≈ 0.422

- 14** $\frac{69}{95}$ **15 a** $\frac{1}{5}$ **b** $P(B | A) \neq P(B)$ **c** $\frac{2}{3}$ **16** $\frac{5}{324}$

REVIEW SET 11B



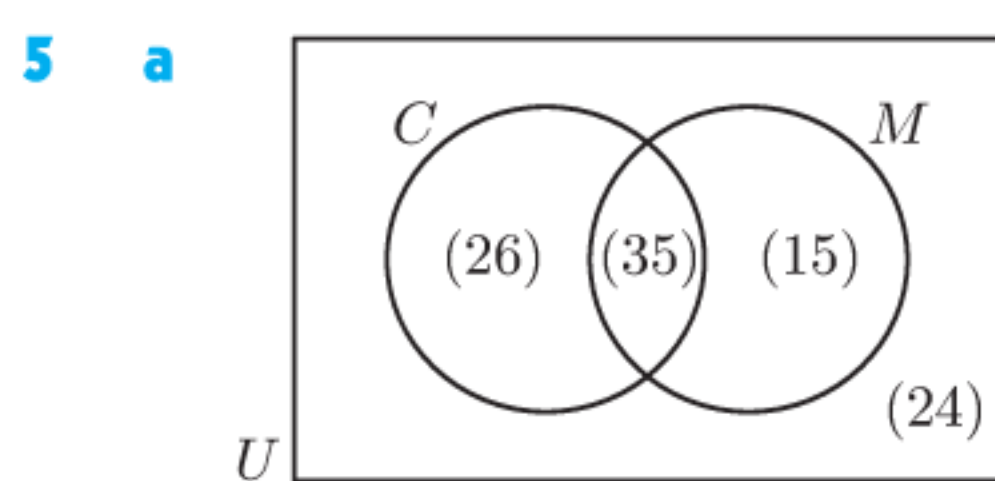
- b** i $\frac{1}{15}$
 ii $\frac{2}{17}$

- 2** 0.9975

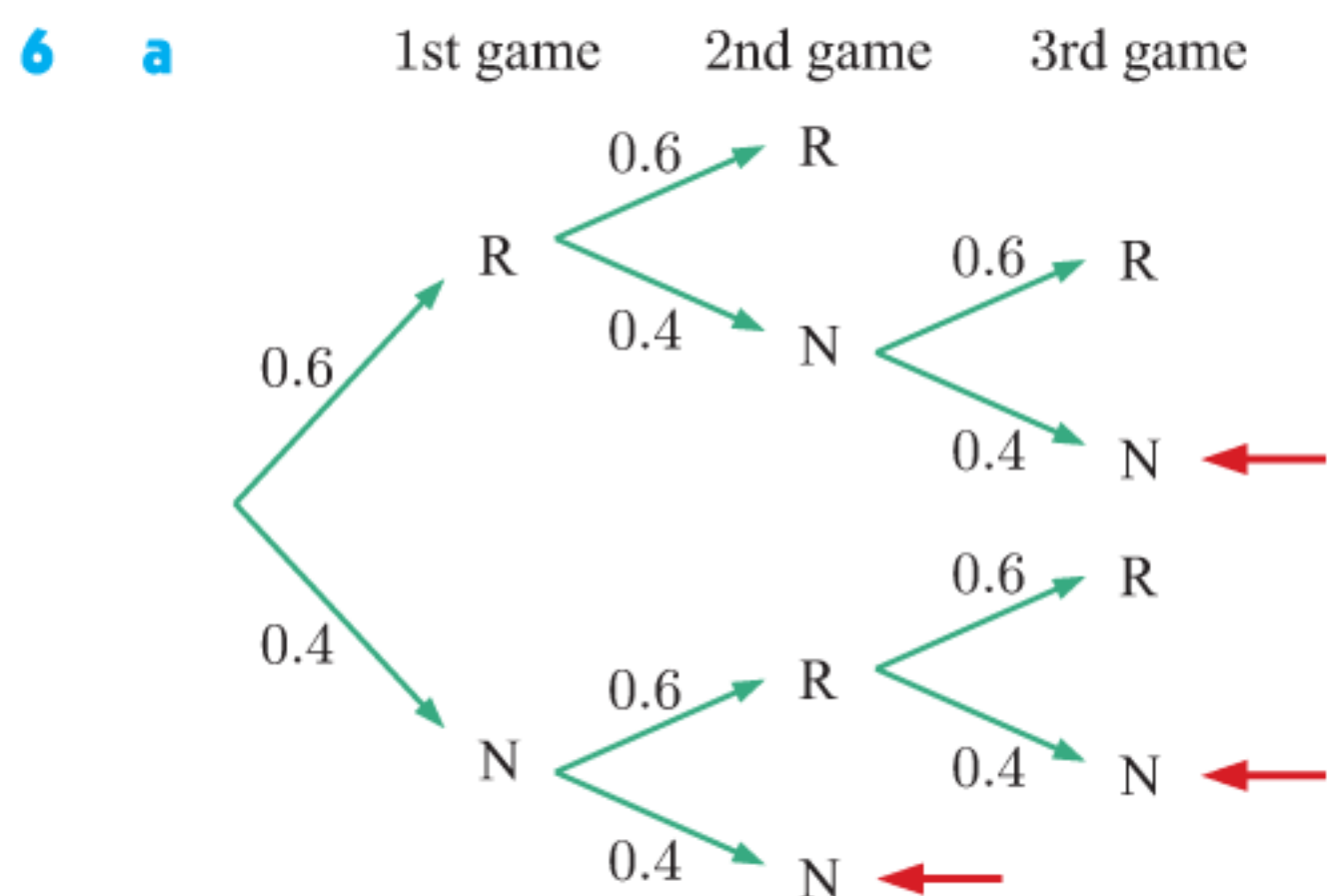
- 3 a** $P(A \cap B) = 0.28$ which is not equal to 0.
 $\therefore A$ and B are not mutually exclusive.

- b** $P(A \cup B) = 0.82$

- 4** $\frac{5}{9}$



- b** $\frac{35}{61}$



- b** $P(N \text{ wins}) = 0.352$

- 7 a** 0.93 **b** 0.8 **c** 0.2 **d** 0.65

- 8 a** $\frac{4}{500} \times \frac{3}{499} \times \frac{2}{498} \approx 0.000\,000\,193$

- b** $1 - \frac{496}{500} \times \frac{495}{499} \times \frac{494}{498} \approx 0.0239$

- 9 a** 0.2588 **b** ≈ 0.703

- 10 a** $P(B) = \frac{1}{3}$ **b** i $\frac{16}{21}$ ii $\frac{13}{21}$

- 11** $\frac{4}{9}$ **12 a** ≈ 0.660 **b** $\frac{100}{7} \approx 14.3$ or 14 pieces

- 13 a** $\frac{31}{70}$ **b** $\frac{21}{31}$ **14** $\frac{1}{2}$

- 15 a** i 100 balloons ii 33 balloons

- b** i $\frac{19}{25}$ ii $\frac{37}{50}$

- c** i $\frac{17}{66}$ ii $\frac{2701}{4950}$ iii $\frac{25}{66}$ iv $\frac{29}{66}$

- d** i $\frac{1}{980}$ ii $\frac{17}{308}$

- 16 b** ≈ 0.988 **c** i ≈ 0.547 ii ≈ 0.266 **d** females

e A 20 year old is expected to live much longer than 30 more years, so it is unlikely the insurance company will have to pay out the policy. A 50 year old however is expected to live for only another 26.45 years (males) or 31.59 years (females), so the insurance company may have to pay out the policy.

g For "third world" countries with poverty, lack of sanitation, and so on, the tables would show a significantly lower life expectancy.

EXERCISE 12A

- This sample is too small to draw reliable conclusions from.
- The sample size is very small and may not be representative of the whole population.
 - The sample was taken in a Toronto shopping mall. People living outside of the city are probably not represented.
- a** The sample is likely to under-represent full-time weekday working voters.

b The members of the golf club may not be representative of the whole electorate.

c Only people who catch the train in the morning such as full-time workers or students will be sampled.

d The voters in the street may not be representative of those in the whole electorate.
- a** The sample size is too small.

b With only 10 sheep being weighed, any errors in the measuring of weights will have more impact on the results.
- a** The whole population is being considered, not just a sample. There will be no sampling error as this is a census.

b measurement error
- a** Many of the workers may not return or even complete the survey.

b There may be more responses to the survey as many workers would feel that it is easier to complete a survey online rather than on paper and mailing it back. Responses would also be received more quickly however some workers may not have internet access and will therefore be unable to complete the survey.
- a** Yes; members with strong negative opinions regarding the management structure of the organisation are more likely to respond.

b No; the feedback from the survey is still valid. Although it might be biased, the feedback might bring certain issues to attention.

EXERCISE 12B

- 1 Note:** Sample answers only - many answers are possible.

- a** 12, 6, 23, 10, 21, 25

- b** 11, 2, 10, 17, 24, 14, 25, 1, 21, 7
c 14, 24, 44, 34, 27, 1 **d** 166, 156, 129, 200, 452
- 2 a** 17, 67, 117, 167, 217 **b** 1600 blocks of chocolate
- 3 a** Select 5 random numbers between 1 and 365 inclusive. For example, 65, 276, 203, 165, and 20 represent 6th March, 3rd October, 22nd July, 14th June, and 20th January.
b Select a random number between 1 and 52 inclusive. Take the week starting on the Monday that lies in that week.
c Select a random number between 1 and 12 inclusive.
d Select 3 random numbers between 1 and 12 inclusive.
e Select a random number between 1 and 10 inclusive for the starting month.
f Select 4 random numbers between 1 and 52 inclusive. Choose the Wednesday that lies in that week.
- 4 a** convenience sampling
b The people arriving first will spend more time at the show, and so are more likely to spend more than €20. Also, the sample size is relatively small.
c For example, systematic sample of every 10th person through the gate.
- 5 a** systematic sampling **b** 14 days
c Only visitors who use the library on Mondays will be counted. Mondays may not be representative of all of the days.
- 6 a** 160 members
b 20 tennis members, 15 lawn bowls members, 5 croquet members
- 7** 1 departmental manager, 3 supervisors, 9 senior sales staff, 13 junior sales staff, 4 shelf packers
- 8 a** It is easier for Mona to survey her own home room class, so this is a convenience sample.
b Mona's sample will not be representative of all of the classes in the school. Mona's survey may be influenced by her friends in her class.
c For example, a stratified sample of students from every class.
- 9 a** Not all students selected for the sample will be comfortable discussing the topic.
b quota sample
- 10 a** All students in Years 11 and 12 were asked, not just a sample.
b 0.48
c **i** Sample too small to be representative.
ii Sample too small to be representative.
iii Valid but unnecessarily large sample size.
iv Useful and valid technique.
v Useful and valid technique.
vi Useful and valid technique.
d v is simple random sampling, while **iii** and **iv** are systematic sampling, and **vi** is stratified or quota sampling.

EXERCISE 12C

Note: Sample answers only - many answers are possible.

- 1 a** It does not allow for colours which are different from those given.
b What colour is your shirt?
c For example, one person might interpret a colour as blue whereas another person may interpret it as purple. A shirt may also be more than one colour which could lead to difficulties in interpreting the colour.
- 2 a** The question could be interpreted as:
 - “Do you have any medically diagnosed allergies?”
 - “Do you have any life threatening allergies?”
 - “Do you have any food allergies?”

- “Do you think you have any allergies?”

The question also does not specify if it is a structured yes/no type of question or if the respondent should list specific allergies.

- b** “Do you have any food or other type of allergies (medically diagnosed or otherwise), and if so, what are they?”
- 3 a** The question could be interpreted as:
 - “Do you have any animals in your household?”
 - “Do you have any animals in your care at home or elsewhere?”

The question does not specify if it is a structured yes/no type of question or if it includes livestock or only domestic animals.

- b** “Do you have any domesticated animals in your household (not including livestock), and if so, what are they?”
- 4 a** The journalist's question is misleading as it only mentions the proposed cuts to education, not the proposal to move those funds to health. This may produce a measurement error as the respondents are unlikely to give their views about the whole proposal.
b For example, “What are your views about the Government's proposal to move funding from education to health?”
- 5 a** Many respondents would not be comfortable giving their address to someone they do not know.
b The question could be more specific, asking for the general area or suburb only. For example:
 - “Which suburb do you live in?”
 - “Which state do you live in?”

Giving a reason why this information is needed will also improve the response rate.

- 6 a i** The question contains a double negative which could confuse respondents. The word “infectious” suggests that *not* immunising is undesirable behaviour, thus the question is biased.
ii “Have you been immunised against meningococcal disease?”
- b i** The question asks for two things:
 - whether climate change is a major issue
 - the respondent's political opinion on climate change.
It is not clear whether the respondent's response will reflect their general or political opinion on the issue. The phrase “thrown around by politicians” is also rather emotive, thus the question is biased.
ii “Do you believe that climate change is an important issue?”
- c i** The question uses a positive fact about fair trade cocoa to try to persuade the respondent into answering “yes”. So the question is biased. It is also very long and takes a long time to get to the point.
ii “Do you believe that fair trade certified chocolate should be more expensive than uncertified chocolate?”

EXERCISE 12D

- 1 a** discrete; 0, 1, 2, 3, ...
b categorical; red, yellow, orange, green
c continuous; 0 - 15 minutes
d continuous; 0 - 25 m
e categorical; Ford, BMW, Renault **f** discrete; 1, 2, 3, ...
g categorical; Australia, Hawaii, Dubai
h discrete; 0.0 - 10.0 **i** continuous; 0 - 4 L
j continuous; 0 - 80 hours **k** continuous; -20°C - 35°C

l categorical; cereal, toast, fruit, rice, eggs

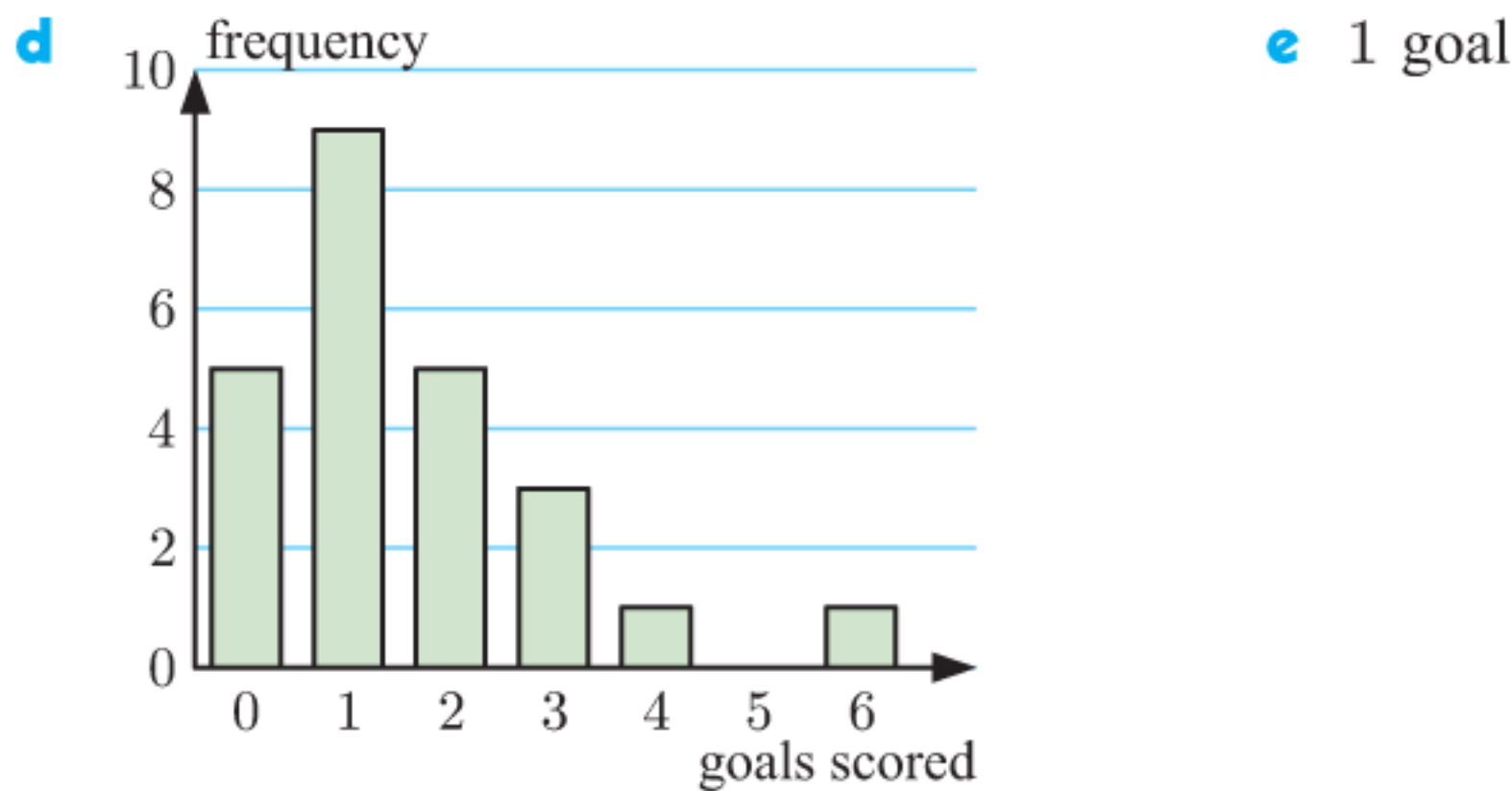
m discrete; 0, 1, 2, ...

2 *Name*: categorical, *Age*: continuous, *Height*: continuous, *Country*: categorical, *Wins*: discrete, *Speed*: continuous, *Ranking*: discrete, *Prize money*: discrete

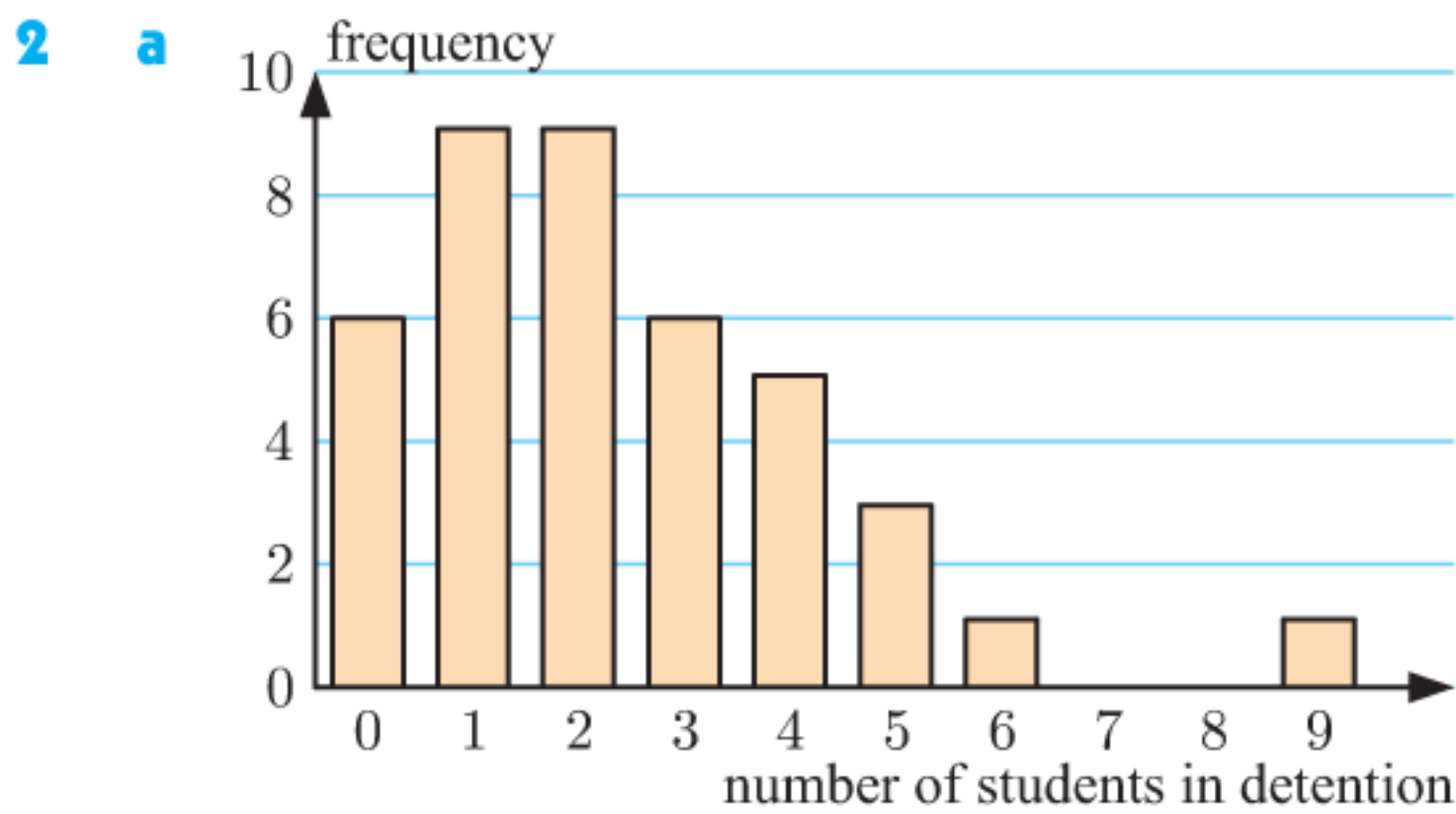
EXERCISE 12E

- 1** **a** the number of goals scored in a game
b variable is counted, not measured

Goals scored	Tally	Frequency	Rel. Frequency
0		5	≈ 0.208
1		9	0.375
2		5	≈ 0.208
3		3	0.125
4		1	≈ 0.042
5		0	0
6		1	≈ 0.042
<i>Total</i>		24	



f positively skewed, one outlier (6 goals) **g** ≈ 20.8%

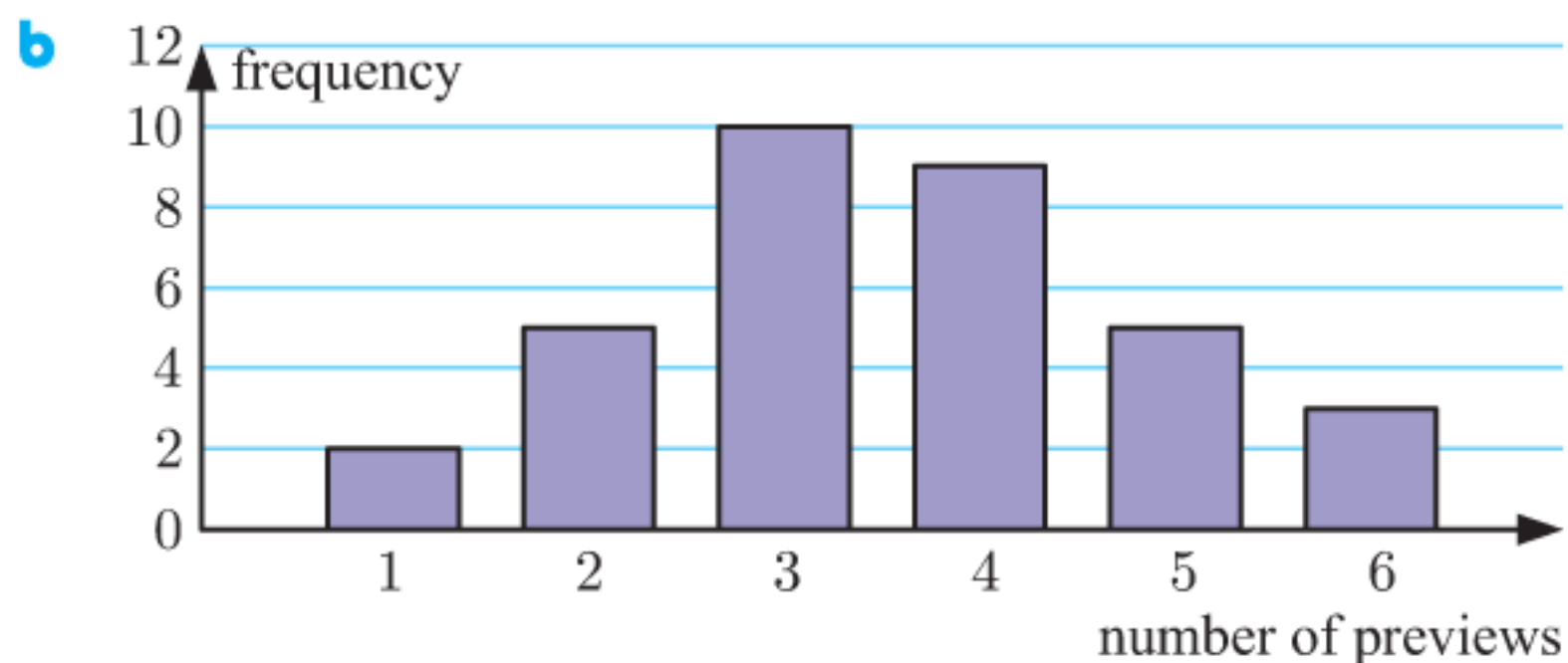


b 1 and 2 **c** positively skewed, one outlier (9 students)

d 12½%

3 a

Number of previews	Tally	Frequency
1		2
2		5
3		10
4		9
5		5
6		3
<i>Total</i>		34



c 3 previews **d** symmetrical, no outliers **e** ≈ 79.4%

- 4 a** 45 people **b** 1 time **c** 8 people **d** 20%
e positively skewed, no outliers

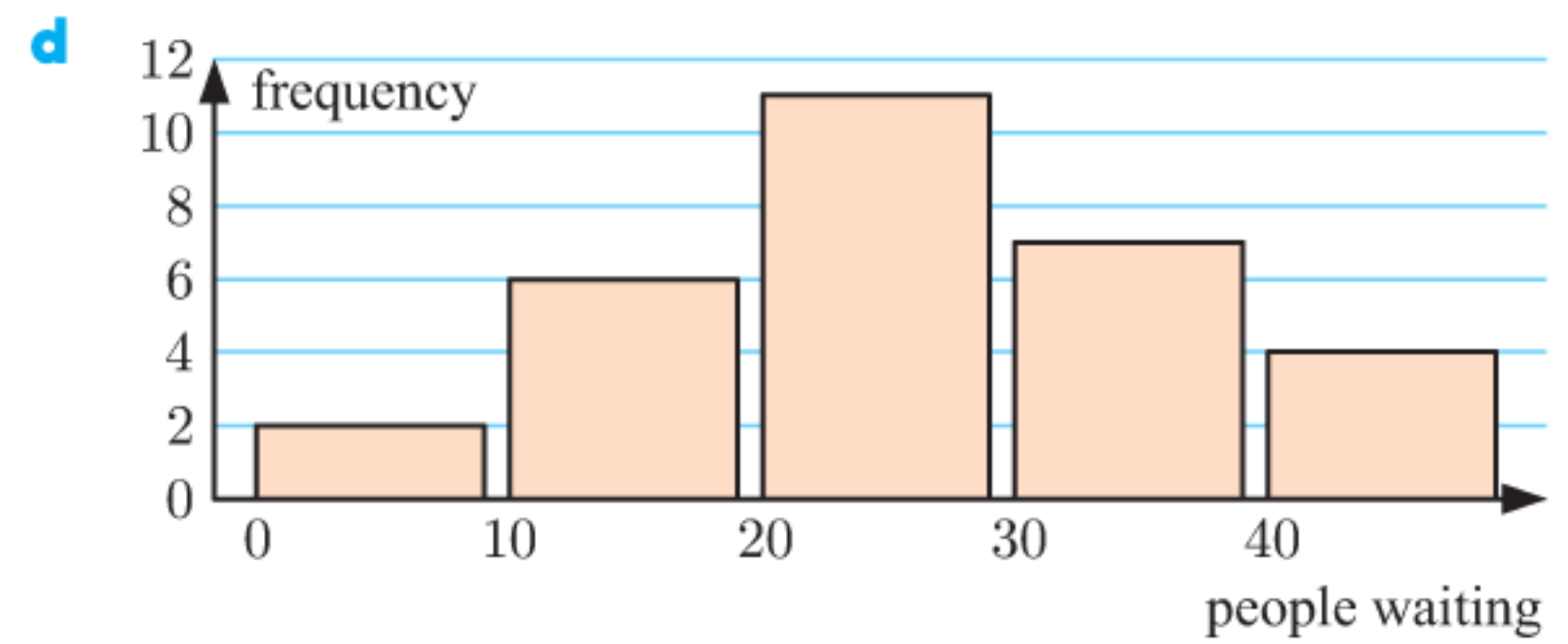
EXERCISE 12F

- 1 a** 37 businesses **b** 40 - 49 employees
c negatively skewed **d** ≈ 37.8%
e No, only that it was in the interval 50 - 59 employees.

2 a

People waiting	Tally	Frequency	Rel. Freq.
0 - 9		2	≈ 0.067
10 - 19		6	0.200
20 - 29		11	≈ 0.367
30 - 39		7	≈ 0.233
40 - 49		4	≈ 0.133
<i>Total</i>		30	

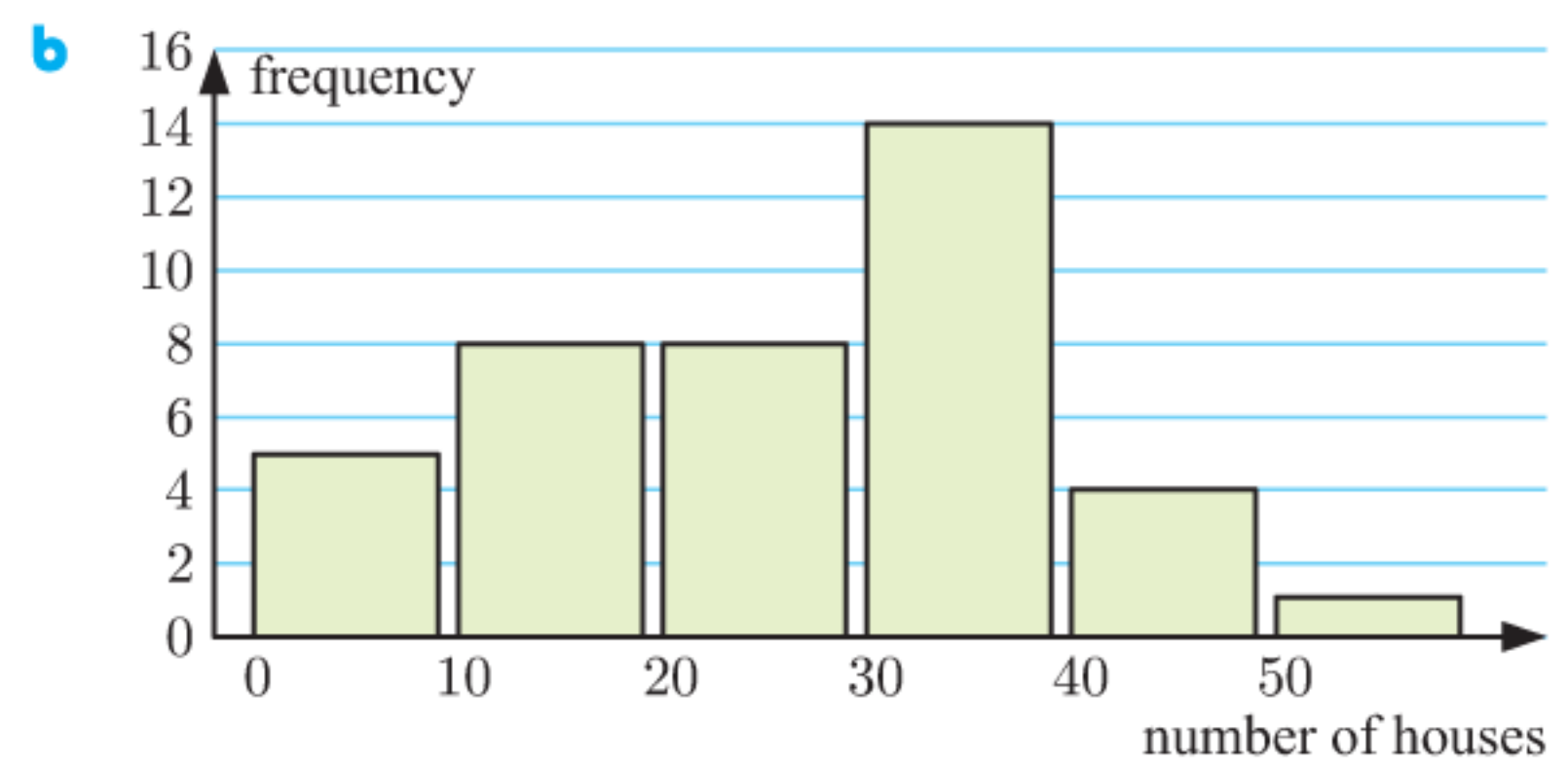
b 2 days **c** ≈ 36.7%



e 20 - 29 people

3 a

Number of houses	Tally	Frequency
0 - 9		5
10 - 19		8
20 - 29		8
30 - 39		14
40 - 49		4
50 - 59		1
<i>Total</i>		40

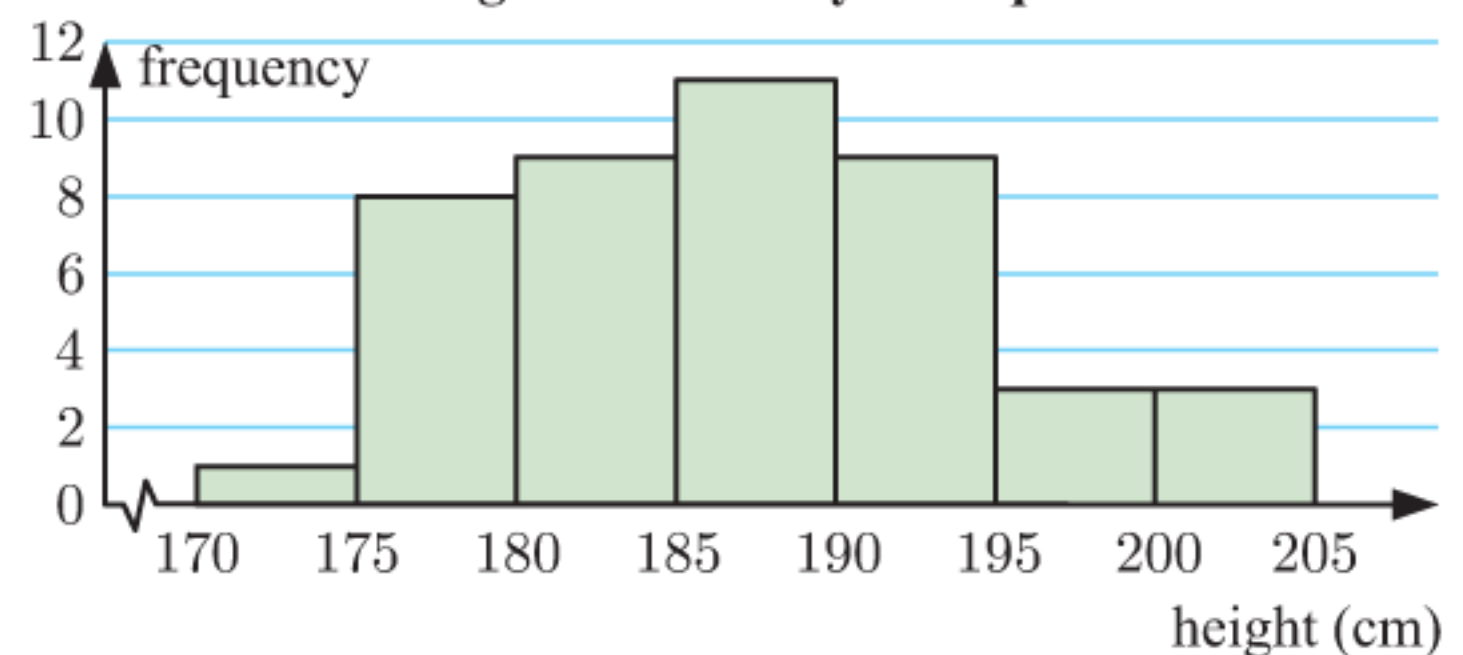


c 30 - 39 houses **d** 67.5%

EXERCISE 12G

- 1 a** Height is measured on a continuous scale.

b **Heights of a volleyball squad**

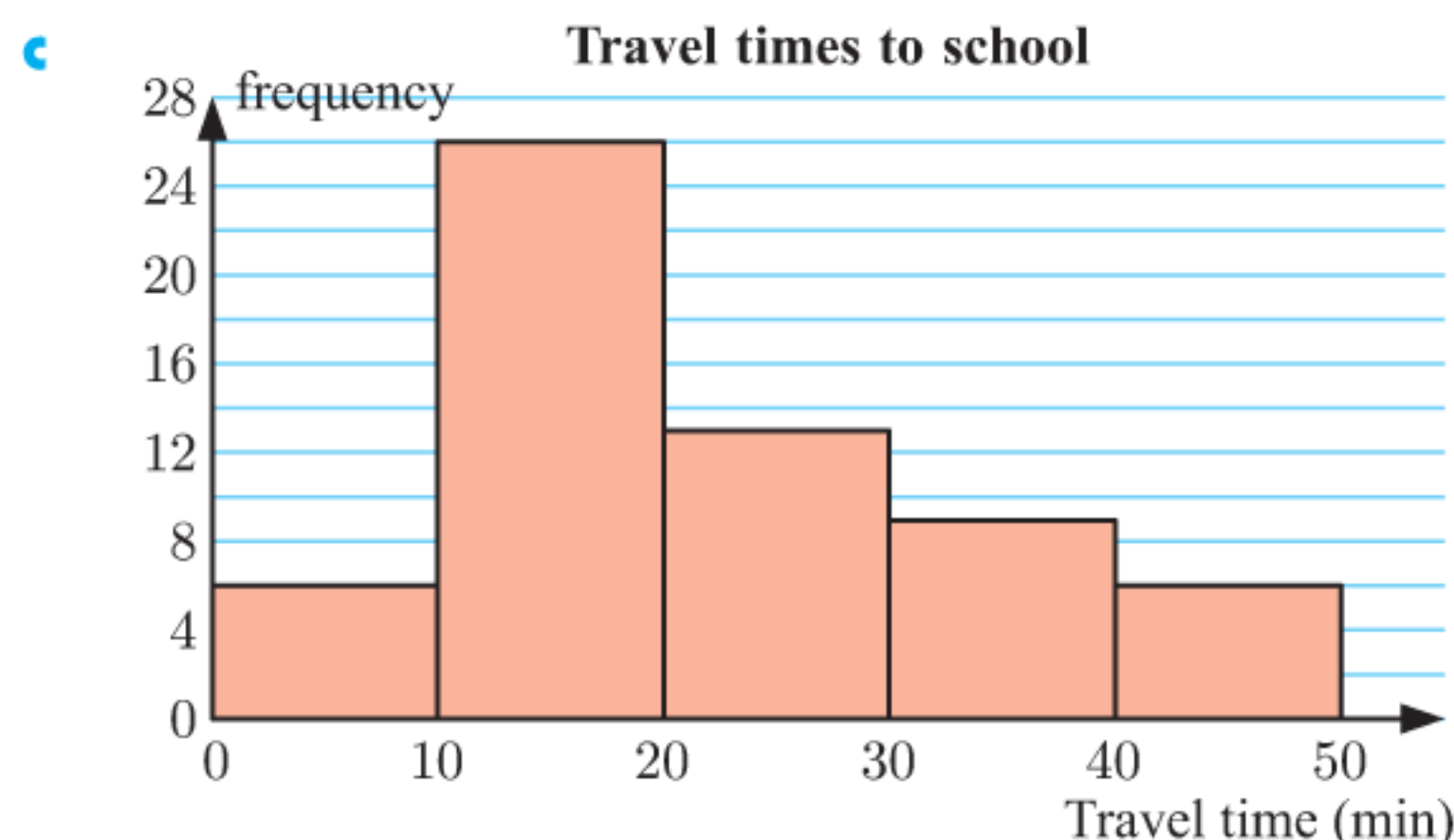


c 185 ≤ H < 190 cm. This is the class of values that appears most often.

d slightly positively skewed

2 a continuous

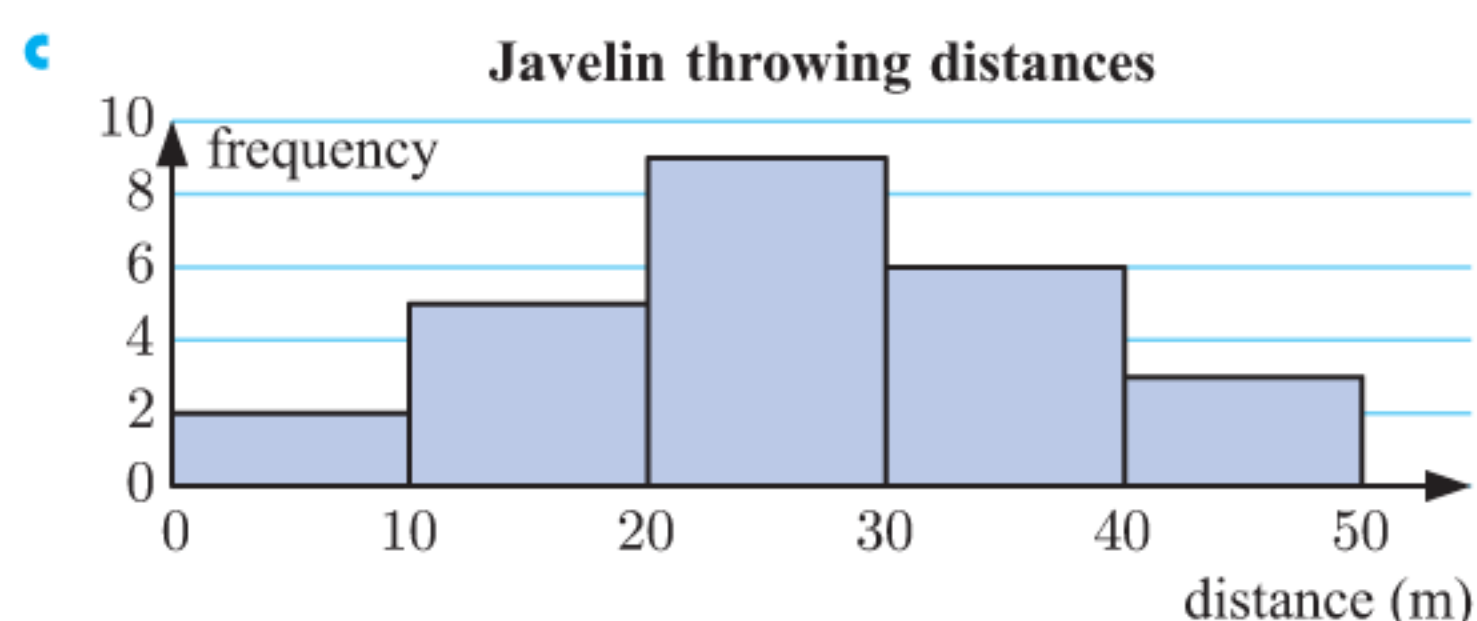
Travel time (min)	Tally	Frequency
$0 \leq t < 10$		6
$10 \leq t < 20$		26
$20 \leq t < 30$		13
$30 \leq t < 40$		9
$40 \leq t < 50$		6
<i>Total</i>		60



d positively skewed e $10 \leq t < 20$ minutes

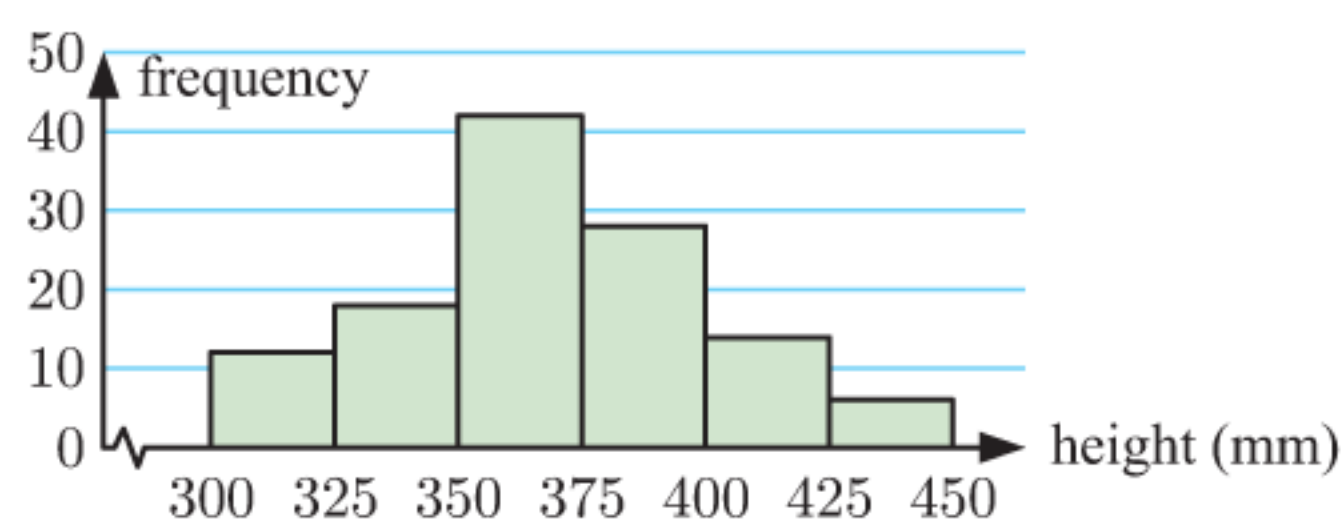
3 a, b

Distance (m)	Tally	Frequency
$0 \leq d < 10$		2
$10 \leq d < 20$		5
$20 \leq d < 30$		9
$30 \leq d < 40$		6
$40 \leq d < 50$		3
<i>Total</i>		25



d $20 \leq d < 30$ m e 36%

4 a Heights of 6-month old seedlings at a nursery

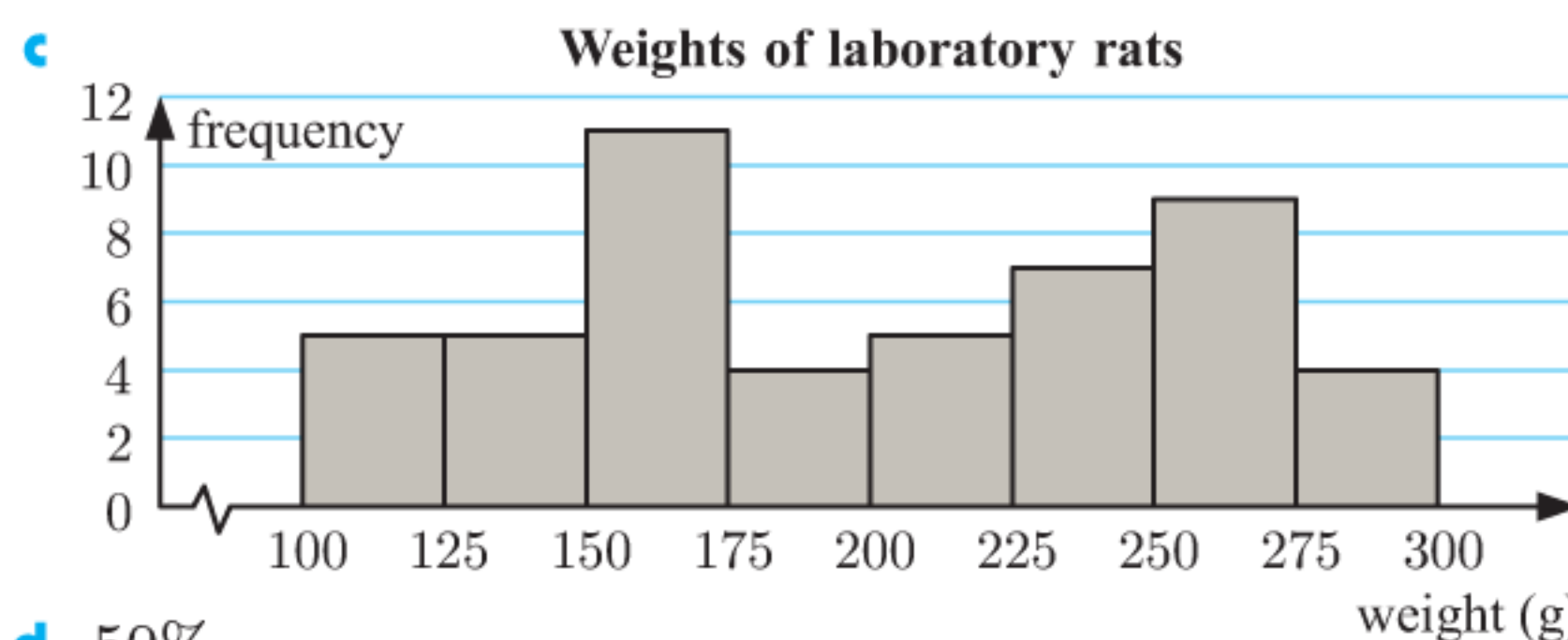


b 20 seedlings c $\approx 58.3\%$

d i ≈ 1218 seedlings ii ≈ 512 seedlings

5 a, b

Weight (g)	Tally	Frequency
$100 \leq w < 125$		5
$125 \leq w < 150$		5
$150 \leq w < 175$		11
$175 \leq w < 200$		4
$200 \leq w < 225$		5
$225 \leq w < 250$		7
$250 \leq w < 275$		9
$275 \leq w < 300$		4
<i>Total</i>		50



d 50%

REVIEW SET 12A

1 a Students studying Italian may have an Italian background so surveying these students may produce a biased result.

b For example, Andrew could survey a randomly selected group of students as they entered the school grounds one morning.

2 a It would be too time consuming and expensive.

b

Age range	< 18	18 - 39	40 - 54	55 - 70	> 70
Sample size	50	82	123	69	26

3 a discrete b continuous c categorical

d categorical e categorical f continuous

g continuous h discrete i discrete

4 a convenience sampling

b Yes, the sample will be biased as people are more likely to be drinking on a Saturday night. It is sensible for this sample to be biased since drink-driving is illegal.

5 a The question could be interpreted as:

- "Do you consider yourself to be healthy?"
- "Are you not currently suffering from any health conditions?"
- "Do you eat a balanced diet and exercise regularly?"
- "Do you take any medication for any health conditions?"

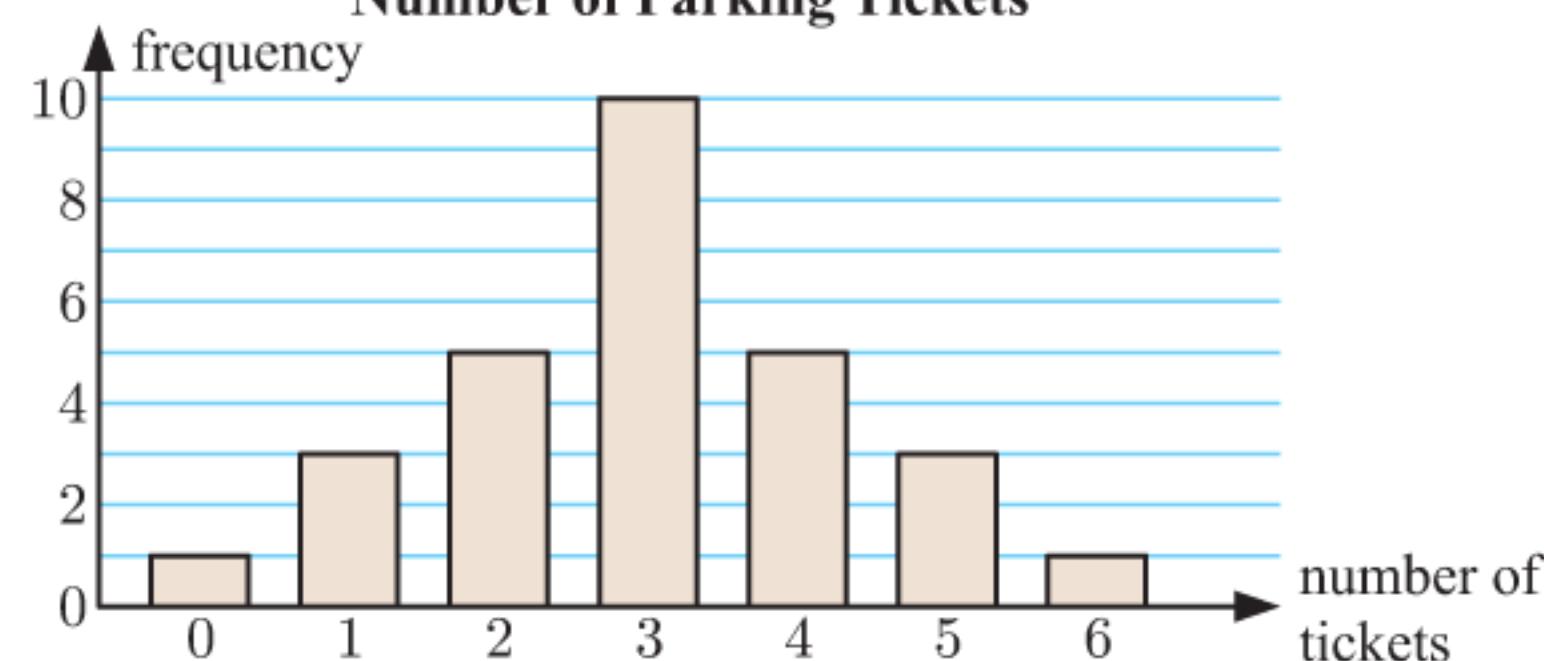
b "Do you eat a balanced diet and exercise regularly?"

6 a discrete b 1 round c positively skewed

7 a

Number of tickets	Tally	Frequency
0		1
1		3
2		5
3		10
4		5
5		3
6		1

b Number of Parking Tickets

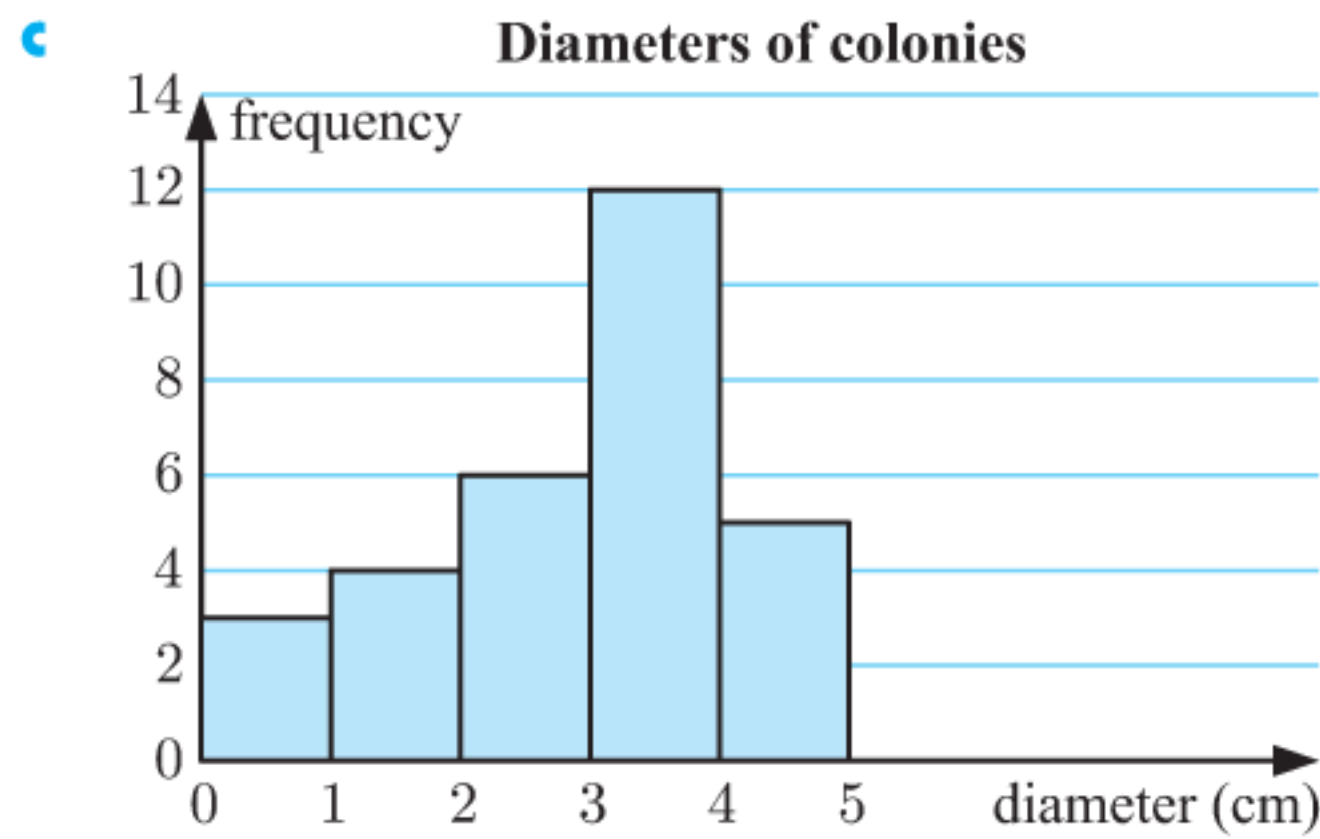


c The data is symmetric with no outliers.

8 a continuous

b

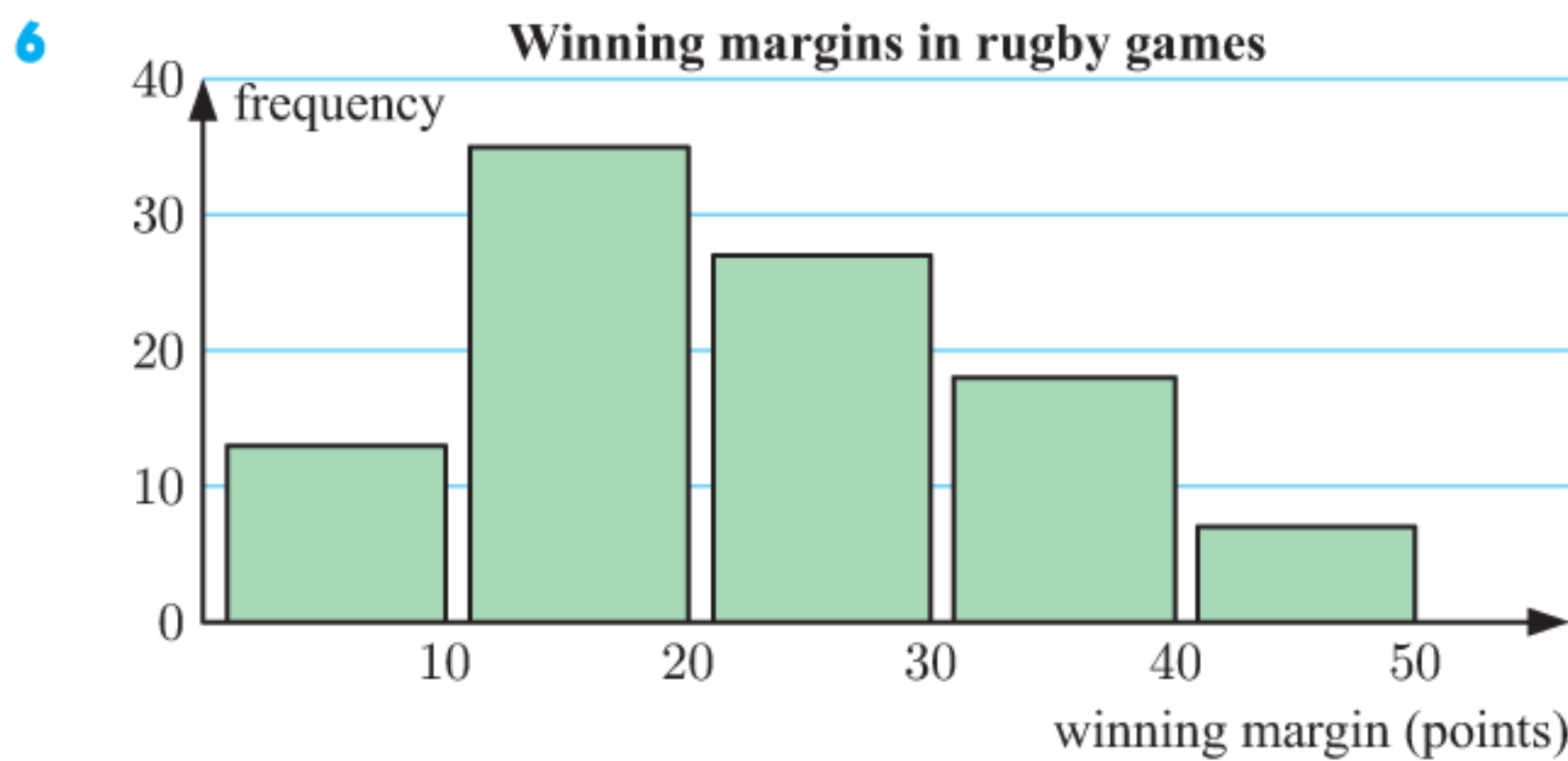
Diameter (d cm)	Tally	Frequency
$0 \leq d < 1$		3
$1 \leq d < 2$		4
$2 \leq d < 3$		6
$3 \leq d < 4$		12
$4 \leq d < 5$		5
<i>Total</i>		30



- d $3 \leq d < 4$ cm e slightly negatively skewed

REVIEW SET 12B

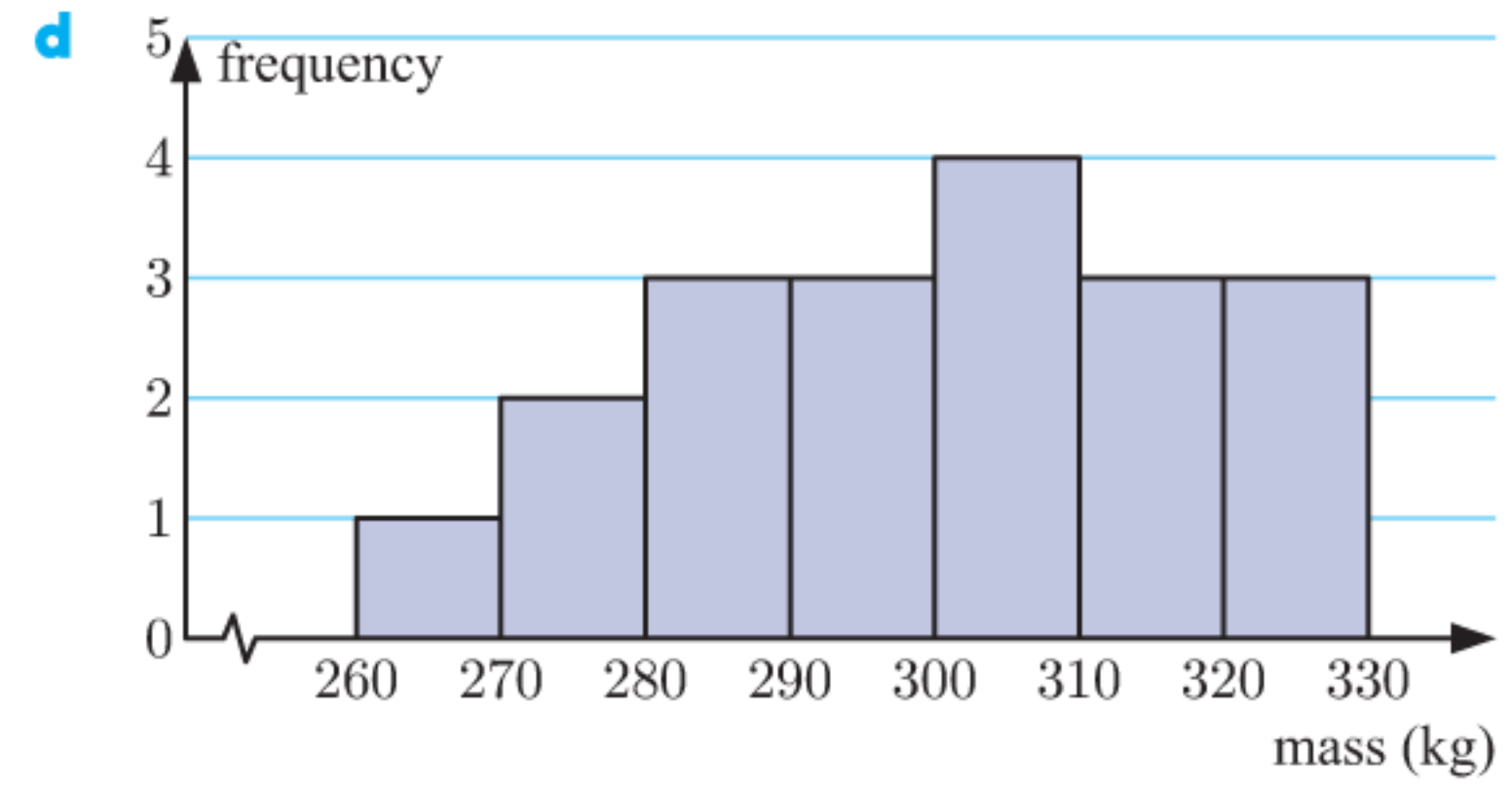
- 1 a discrete b continuous c discrete
- 2 a systematic sampling
- b A house will be visited if the last digit in its number is equal to the random number chosen by the promoter, with the random number 10 corresponding to the digit 0. Each house therefore has a 1 in 10 chance of being visited.
- c Once the first house number has been chosen, the remaining houses chosen must all have the same second digit in their house number, that is, they are not randomly chosen. For example, it is impossible for two consecutively numbered houses to be selected for the sample.
- 3 a Petra's teacher colleagues are quite likely to ignore the emailed questionnaire as emails are easy to ignore.
- b It is likely that the teachers who have responded will have strong opinions either for or against the general student behaviour. These responses may therefore not be representative of all teachers' views.
- 4 Did you learn about our services via:
- friends/family • the internet • newspaper
 - television • elsewhere?
- 5 a The tone is not neutral and it is a structured question. The only responses possible are yes or no.
- b How would you describe your general behaviour when you were a child?



- 7 a Mass is measured on a continuous scale.

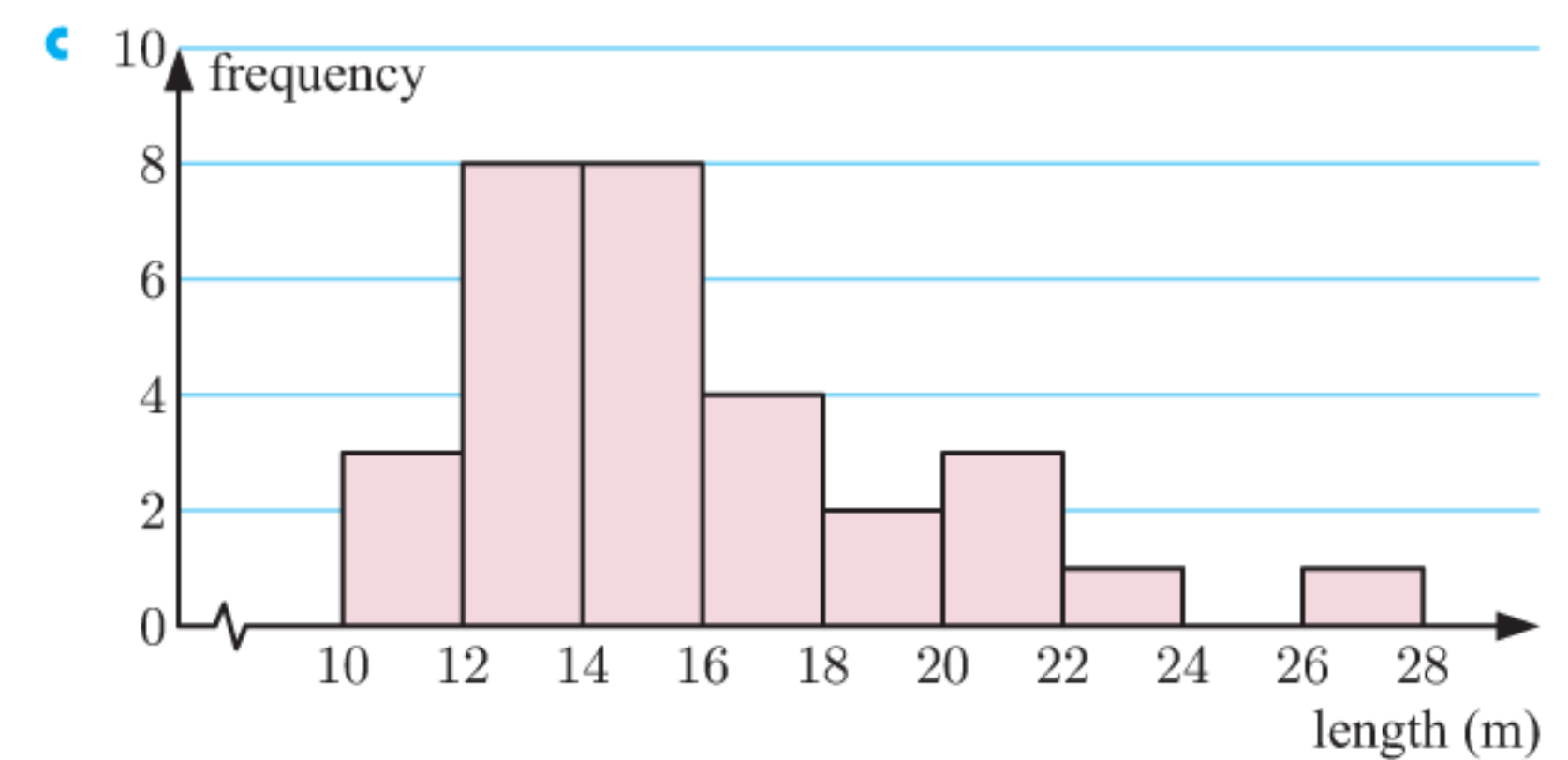
Mass (m kg)	Frequency
$260 \leq m < 270$	1
$270 \leq m < 280$	2
$280 \leq m < 290$	3
$290 \leq m < 300$	3
$300 \leq m < 310$	4
$310 \leq m < 320$	3
$320 \leq m < 330$	3

- c $300 \leq m < 310$ kg



- e slightly negatively skewed
- 8 a continuous

Length (l m)	Frequency
$10 \leq l < 12$	3
$12 \leq l < 14$	8
$14 \leq l < 16$	8
$16 \leq l < 18$	4
$18 \leq l < 20$	2
$20 \leq l < 22$	3
$22 \leq l < 24$	1
$24 \leq l < 26$	0
$26 \leq l < 28$	1



- d positively skewed, one outlier (27.4 m)

EXERCISE 13A

- 1 a 1 cup b 2 cups c 1.8 cups
- 2 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
- 3 9 4 Ruth
- 5 a data set A: ≈ 6.46 , data set B: ≈ 6.85
- b data set A: 7, data set B: 7
- c Data sets A and B differ only by their last value. This affects the mean, but not the median.
- 6 a i motichoor ladoo: ≈ 67.1 , malai jamun: ≈ 53.6
- ii motichoor ladoo: 69, malai jamun: 52
- b The mean and median were much higher for the motichoor ladoo, so the motichoor ladoo were more popular.
- 7 a Bus: mean = 39.7, median = 40.5
Tram: mean ≈ 49.1 , median = 49
- b The tram data has a higher mean and median, but since there are more bus trips per day and more people travel by bus in total, the bus is more popular.
- 8 a 44 points b 44 points
- c i Decrease, since 25 is lower than the mean of 44 for the first four matches.
- ii 40.2 points
- 9 €185 604 10 116 11 17.25 goals per game
- 12 $x = 15$ 13 $a = 5$ 14 37 marks
- 15 ≈ 14.8 16 6 and 12

EXERCISE 13B

- a** mean = \$363 770, median = \$347 200
The mean has been affected by the extreme values (the two values greater than \$400 000).

b **i** the mean **ii** the median
- a** mode = \$33 000, mean = \$39 300, median = \$33 500

b The mode is the lowest value in the data set.

c No, it is too close to the lower end of the distribution.
- a** mean \approx 3.19 mm, median = 0 mm, mode = 0 mm

b The median is not the most suitable measure of centre as the data is positively skewed.

c The mode is the lowest value.

d 42 mm and 21 mm **e** no
- a** mean \approx 2.03, median = 2, mode = 1 and 2

b Yes, as Esmé can then offer a “family package” to match the most common number of children per family.

c 2 children, since this is one of the modes; it is also the median, and very close to the mean.

EXERCISE 13C

- a** 1 person **b** 2 people **c** \approx 2.03 people
- a** **i** 2.96 phone calls **ii** 2 phone calls
iii 2 phone calls

b

Phone calls in a day

frequency

number of phone calls

mode, median (2) mean (2.96)

c positively skewed

d The mean takes into account the larger numbers of phone calls.

e the mean
- a** **i** \approx 2.61 children **ii** 2 children **iii** 2 children

b This school has more children per family than the average British family.

c positively skewed

d The values at the higher end increase the mean more than the median and the mode.

Pocket money (€)	Frequency
1	4
2	9
3	2
4	6
5	8

b 29 children

c **i** \approx €3.17
ii €3
iii €2

d the mode

- 10.1 cm
- a** **i** \$63 000 **ii** \$56 000 **iii** \$66 600 **b** the mean
- a** $x = 5$ **b** 75%

EXERCISE 13D

- a** 40 phone calls **b** \approx 15 minutes **c** \approx 31.7
- a** 70 service stations **b** \approx 411 000 litres (\approx 411 kL)

c \approx 5870 L

d $6000 < P \leq 7000$ L. This is the most frequently occurring amount of petrol sales at a service station in one day.

- | Runs scored | Tally | Frequency |
|-------------|-------|-----------|
| 0 - 9 | | 11 |
| 10 - 19 | | 8 |
| 20 - 29 | | 8 |
| 30 - 39 | | 2 |
| Total | | 29 |

b \approx 14.8 runs

c \approx 14.9 runs; the estimate in **b** was very accurate.
- a** $p = 24$ **b** \approx 3.37 minutes **c** \approx 15.3%
- a** 125 people **b** \approx 119 marks **c** $\frac{3}{25}$ **d** 28%

EXERCISE 13E

- a** **i** 13 **ii** $Q_1 = 9, Q_3 = 18$ **iii** 16 **iv** 9

b **i** 18.5 **ii** $Q_1 = 13, Q_3 = 23$ **iii** 19 **iv** 10

c **i** 26.5 **ii** $Q_1 = 20, Q_3 = 35$ **iii** 28 **iv** 15

d **i** 37 **ii** $Q_1 = 28, Q_3 = 52$ **iii** 49 **iv** 24
- a** Jane: mean = \$35.50, median = \$35.50
Ashley: mean = \$30.75, median = \$26.00

b Jane: range = \$18, IQR = \$9
Ashley: range = \$40, IQR = \$14

c Jane **d** Ashley
- a** range = 60, IQR = 8.5 **b** ‘67’ is an outlier.

c range = 18, IQR = 8 **d** the range
- a** Derrick: range = 240 minutes, IQR = 30 minutes
Gareth: range = 170 minutes, IQR = 120 minutes

b **i** Gareth’s **ii** Derrick’s

c The IQR is most appropriate as it is less affected by outliers.
- a** g **b** **i** $m - a$ **ii** $\left(\frac{j+k}{2}\right) - \left(\frac{c+d}{2}\right)$

Measure	median	mode	range	interquartile range
a	11	9	13	6
b	18	14	26	12

EXERCISE 13F

- a** 35 points **b** 78 points **c** 13 points **d** 53 points

e 26 points **f** 65 points **g** 27 points
- a** **i** 98, 25 marks **ii** 70 marks **iii** 85 marks
iv 55, 85 marks

b range = 73, IQR = 30
- a** **i** min = 3, $Q_1 = 5$, med = 6, $Q_3 = 8$, max = 10

ii

iii range = 7, IQR = 3

b **i** min = 0, $Q_1 = 4$, med = 7, $Q_3 = 8$, max = 9

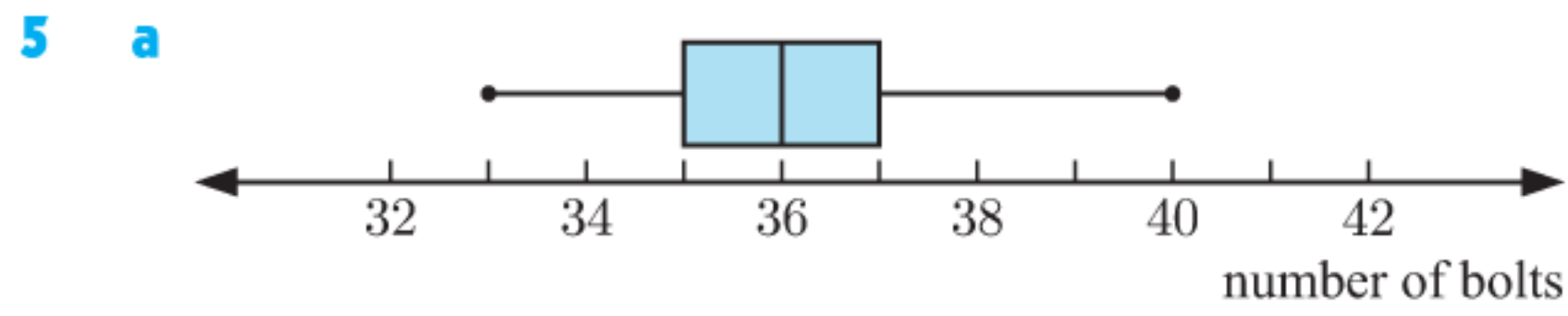
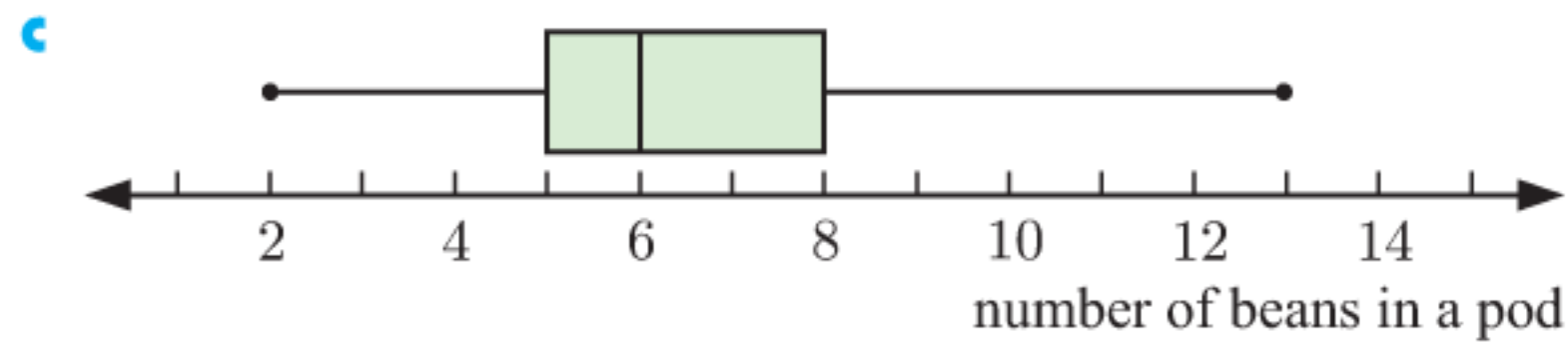
ii

iii range = 9, IQR = 4

c **i** min = 17, $Q_1 = 26$, med = 31, $Q_3 = 47$, max = 51

ii

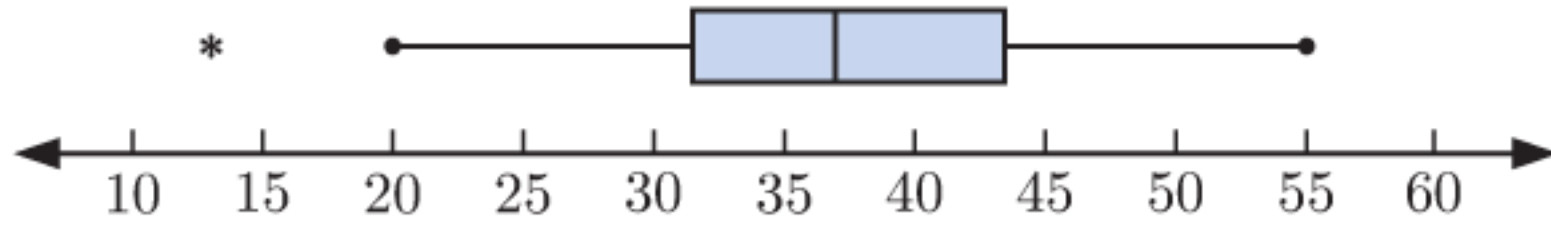
iii range = 34, IQR = 21
- a** median = 6, $Q_1 = 5, Q_3 = 8$ **b** IQR = 3



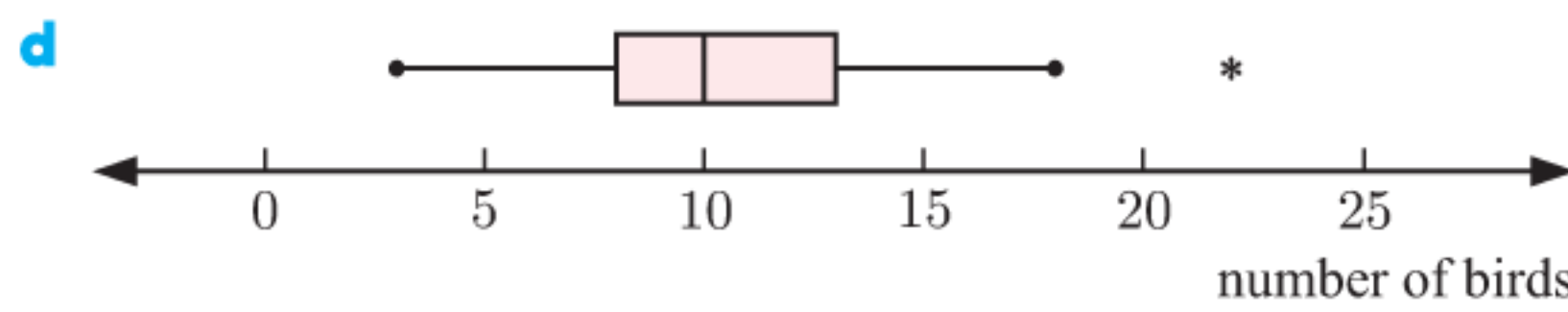
b range = 7, IQR = 2

EXERCISE 13G

1 a 12 **b** lower = 13.5, upper = 61.5 **c** 13
d *

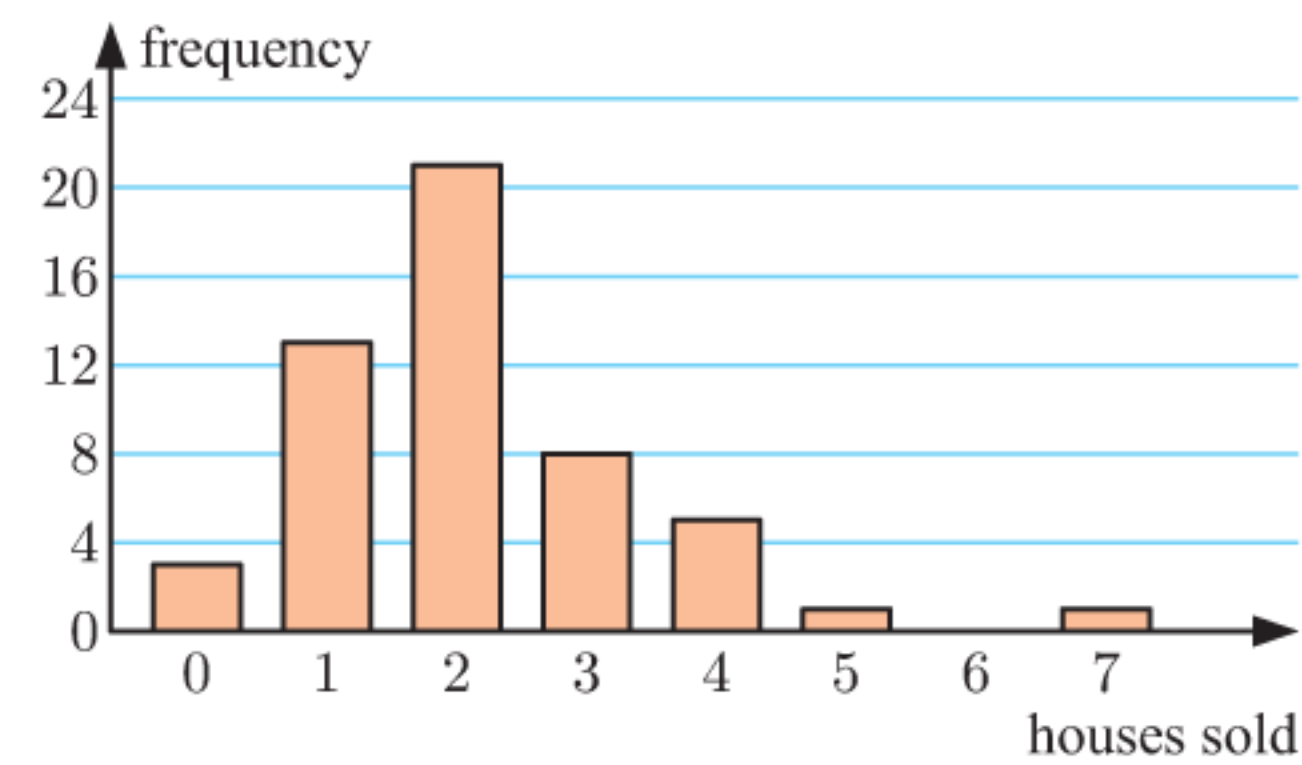


2 a median = 10, $Q_1 = 8$, $Q_3 = 13$ **b** IQR = 5
c lower = 0.5, upper = 20.5, 22 is an outlier

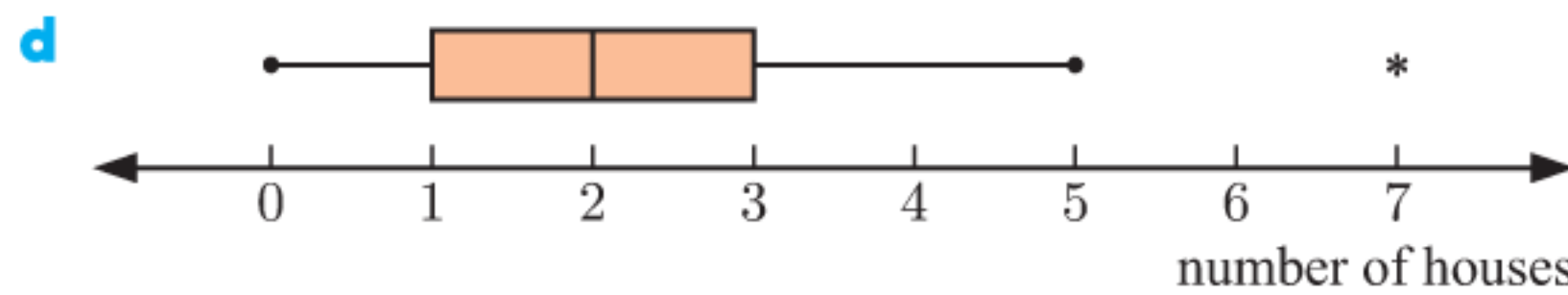


3 a **A** **b** **D** **c** **C** **d** **B**

4 a Houses sold by a real estate agent



b 7 houses appears to be an outlier.
c lower boundary = -2, upper boundary = 6
 7 houses is an outlier



EXERCISE 13H

1 a

Statistic	Year 9	Year 12
minimum	6	36
Q_1	30	60
median	45	84
Q_3	60	96
maximum	72	105

b i Year 9: 66 min
 Year 12: 69 min
ii Year 9: 30 min
 Year 12: 36 min

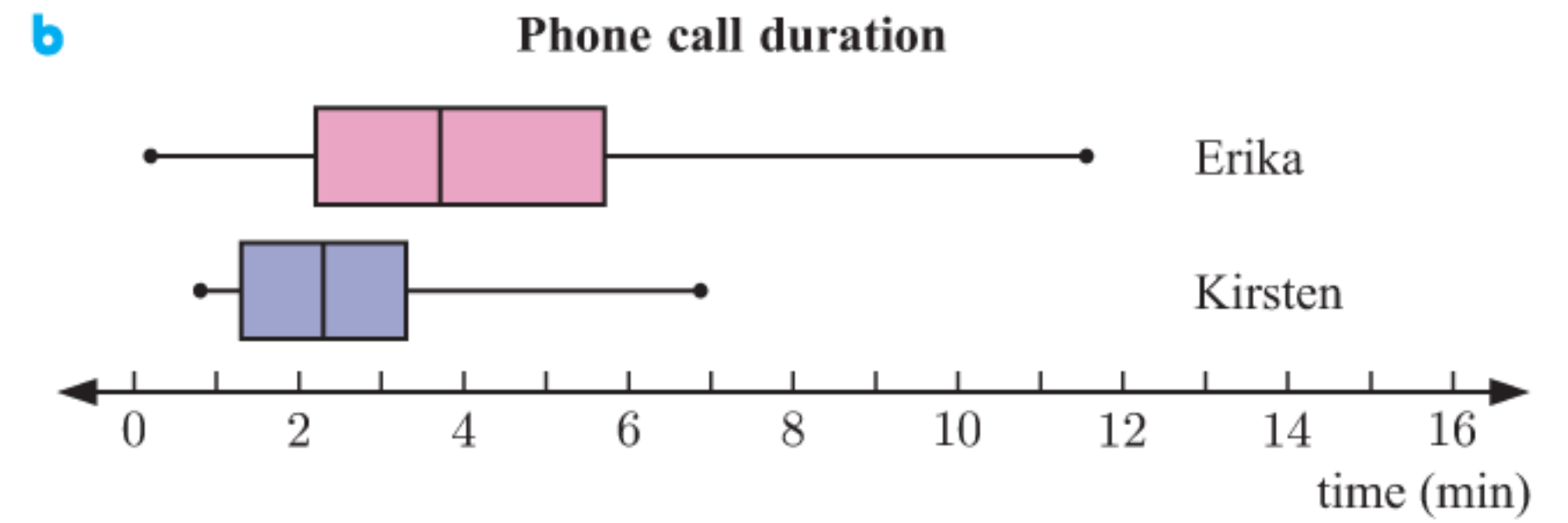
c i cannot tell **ii** true, since Year 9 $Q_1 <$ Year 12 min

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

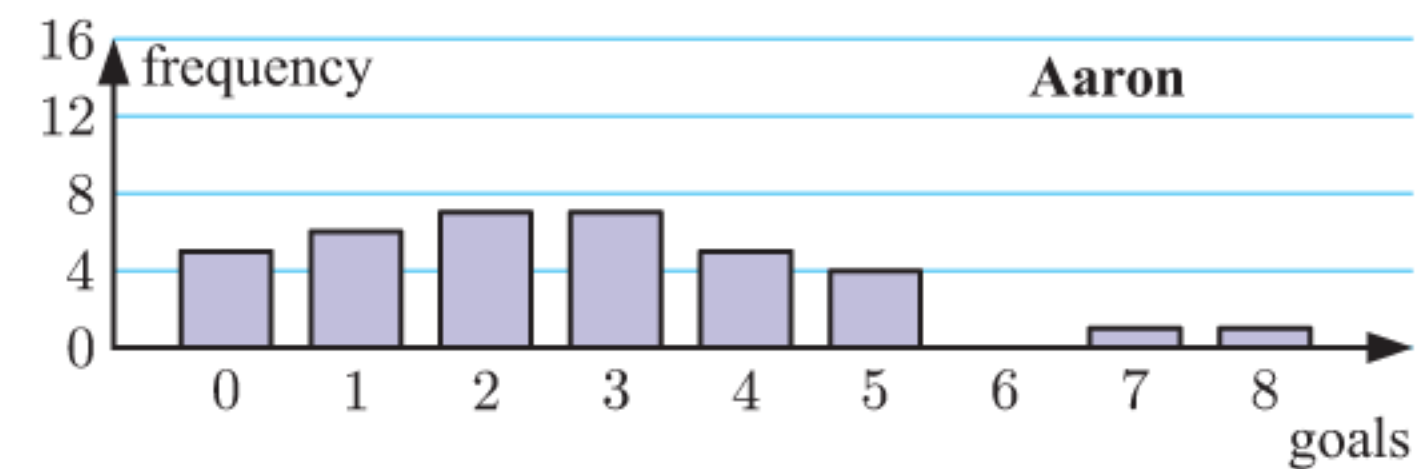
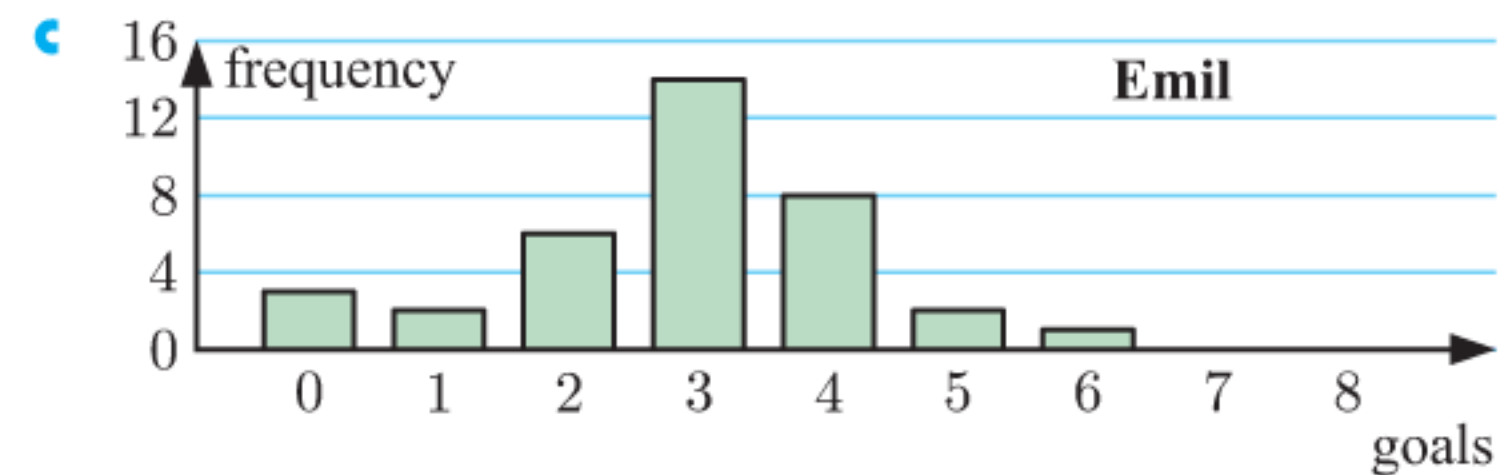
3 a i class 1 (96%) **ii** class 1 (37%) **iii** class 1
b 18% **c** 55% **d i** 25% **ii** 50%
e i slightly positively skewed **ii** negatively skewed
f class 2, class 1

4 a Kirsten: min = 0.8 min, $Q_1 = 1.3$ min, med = 2.3 min, $Q_3 = 3.3$ min, max = 6.9 min
 Erika: min = 0.2 min, $Q_1 = 2.2$ min, med = 3.7 min, $Q_3 = 5.7$ min, max = 11.5 min



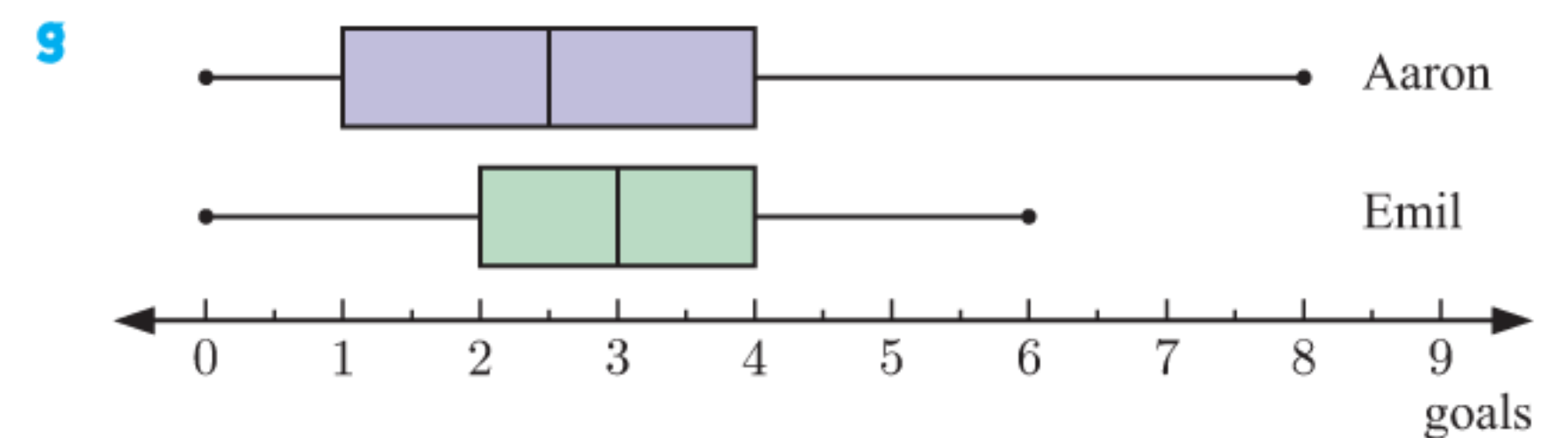
c Both are positively skewed (Erika's more so than Kirsten's). Erika's phone calls were more varied in duration.

5 a discrete



d Emil: approximately symmetrical
 Aaron: positively skewed
e Emil: mean ≈ 2.89 , median = 3, mode = 3
 Aaron: mean ≈ 2.67 , median = 2.5, mode = 2, 3
 Emil's mean and median are slightly higher than Aaron's. Emil has a clear mode of 3, whereas Aaron has two modes (2 and 3).

f Emil: range = 6, IQR = 2
 Aaron: range = 8, IQR = 3
 Emil's data set demonstrates less variability than Aaron's.



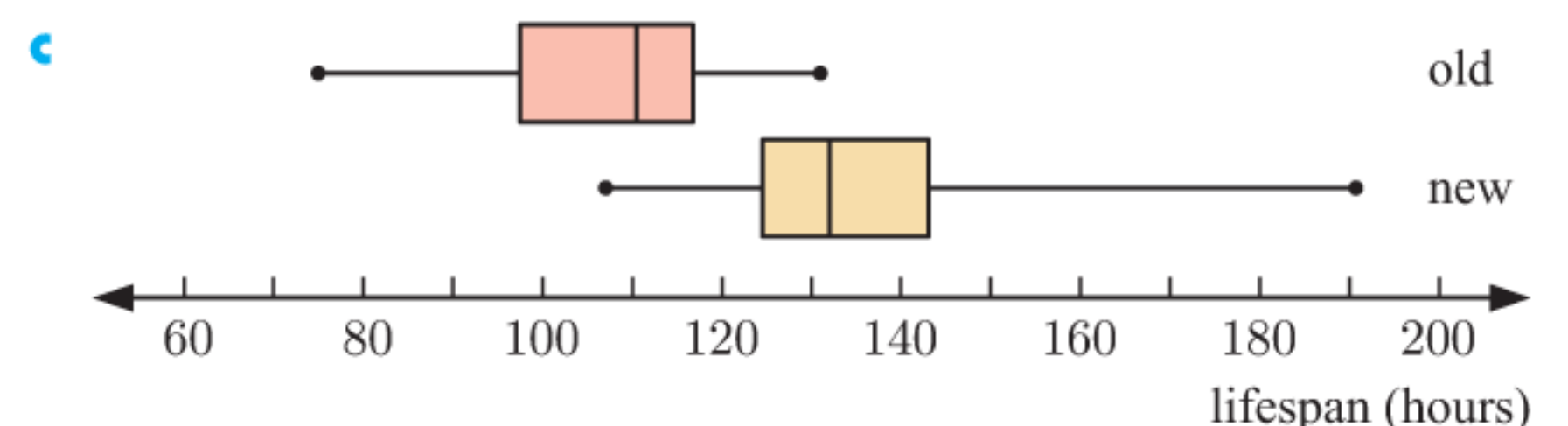
h Emil is more consistent with his scoring (in terms of goals) than Aaron.

6 a continuous (the data is measured)

b Old type: mean = 107 hours, median = 110.5 hours, range = 56 hours, IQR = 19 hours
 New type: mean = 134 hours, median = 132 hours, range = 84 hours, IQR = 18.5 hours

The "new" type of light globe has a higher mean and median than the "old" type.

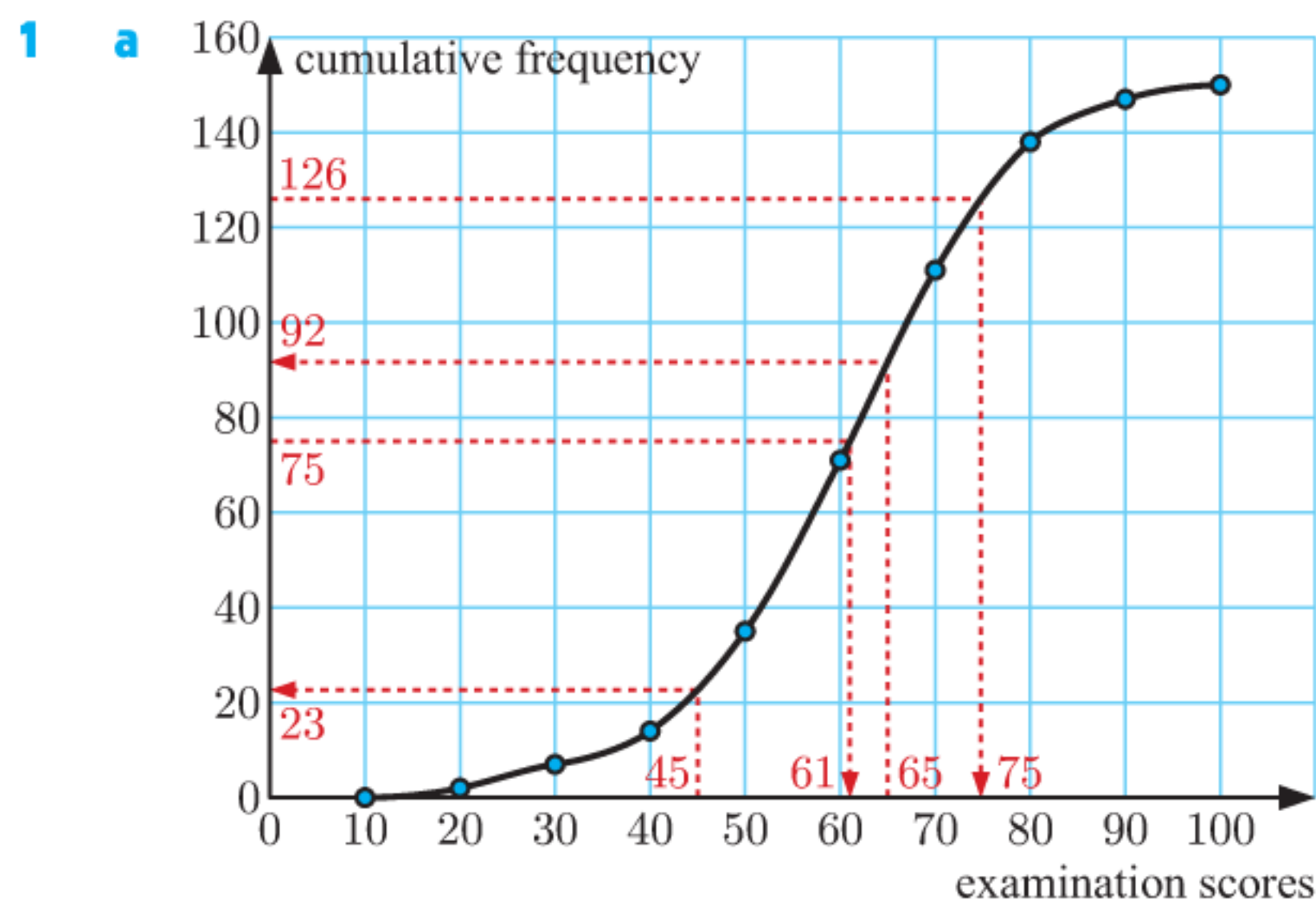
The IQR is relatively unchanged going from "old" to "new", however, the range of the "new" type is greater, suggesting greater variability.



d Old type: negatively skewed
 New type: positively skewed

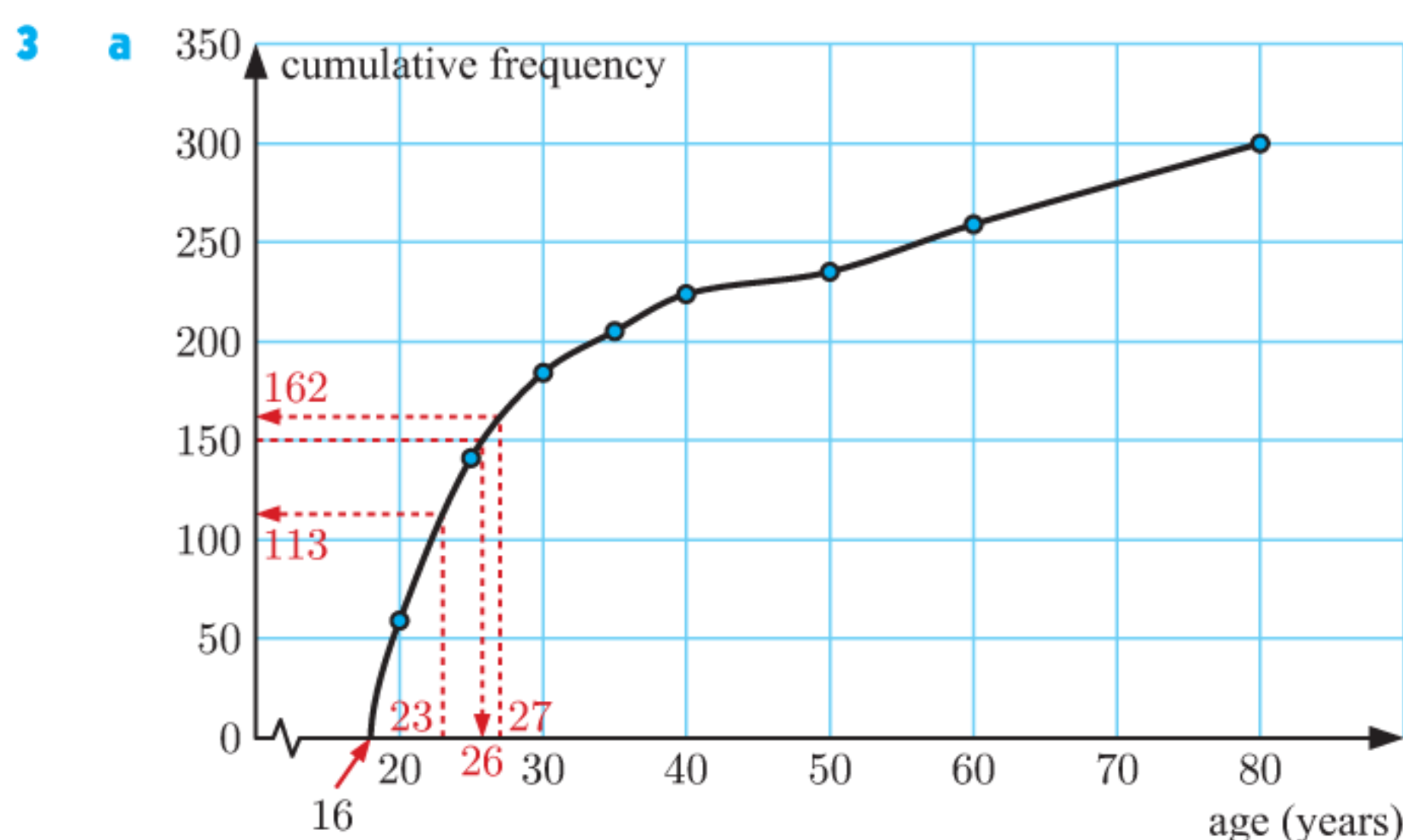
e The “new” type of light globes do last longer than the “old” type. From c, both the mean and median for the “new” type are close to 20% greater than that of the “old” type. The manufacturer’s claim appears to be valid.

EXERCISE 13I



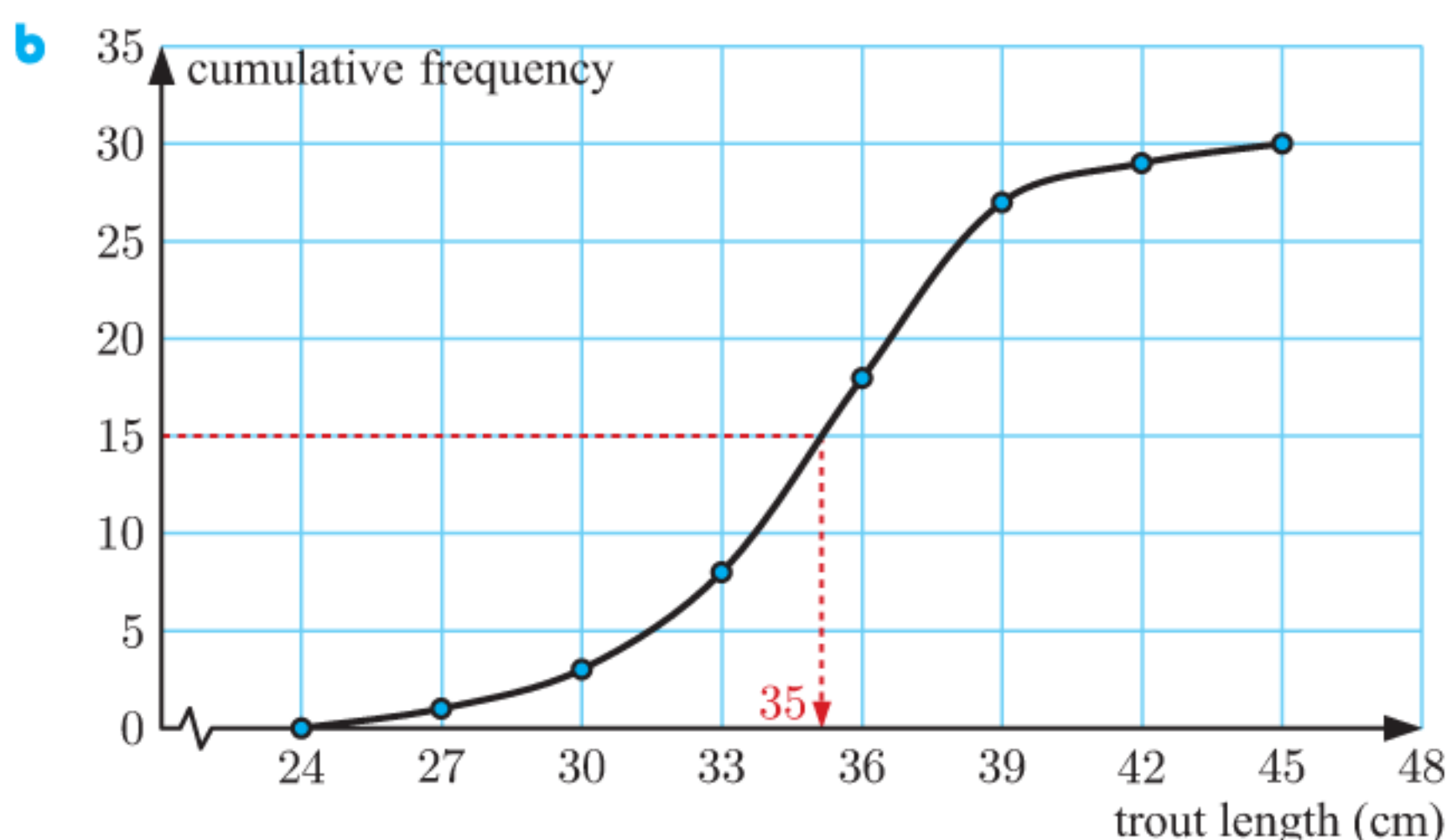
b ≈ 61 marks c ≈ 92 students d 76 students
 e ≈ 23 students f ≈ 75 marks

2 a ≈ 9 seedlings b $\approx 28.3\%$ c ≈ 7.1 cm
 d ≈ 2.4 cm
 e 10 cm, which means that 90% of the seedlings are shorter than 10 cm.



b ≈ 26 years c $\approx 37.7\%$ d i ≈ 0.54 ii ≈ 0.04

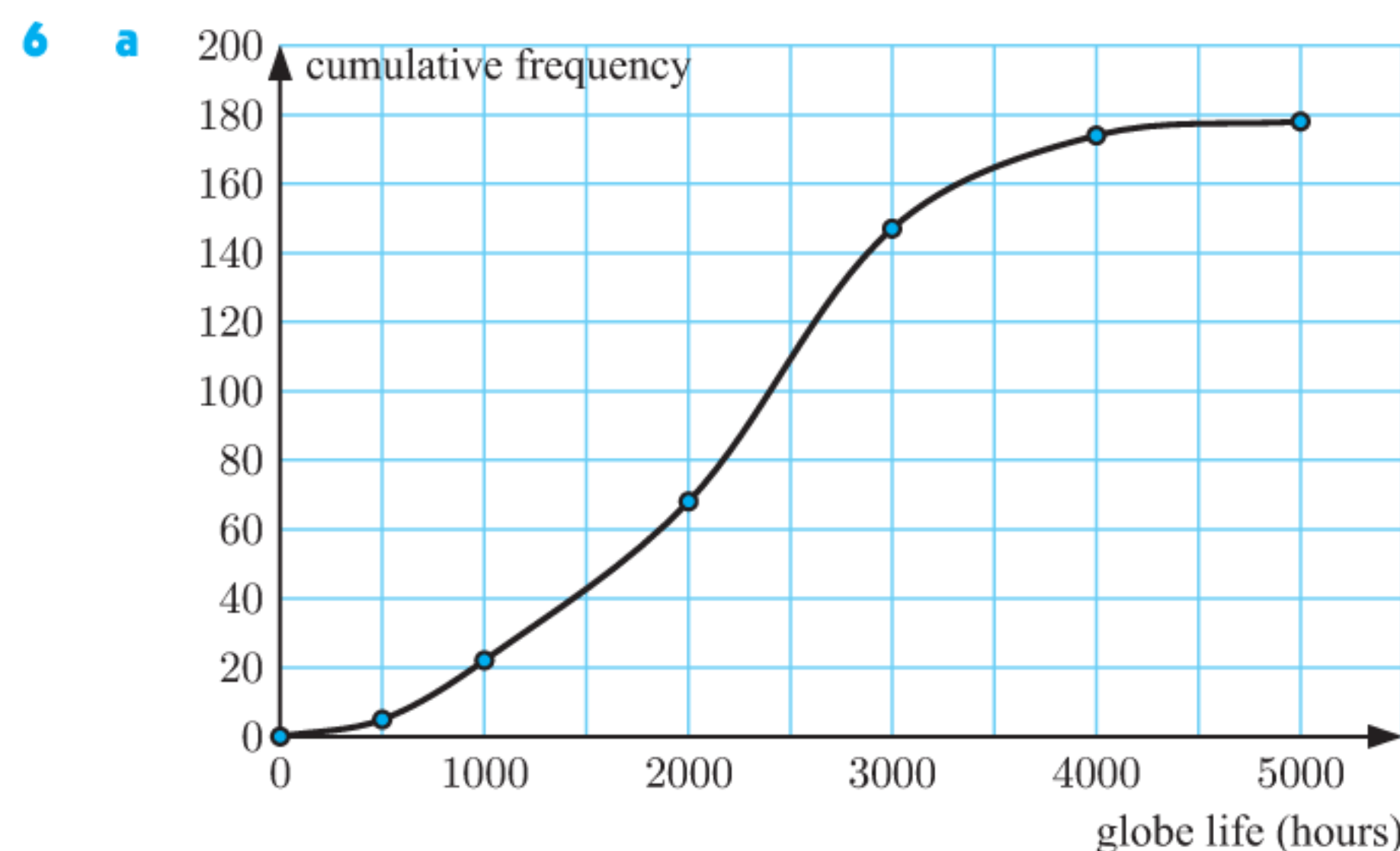
Length (cm)	Frequency	Cumulative frequency
$24 \leq x < 27$	1	1
$27 \leq x < 30$	2	3
$30 \leq x < 33$	5	8
$33 \leq x < 36$	10	18
$36 \leq x < 39$	9	27
$39 \leq x < 42$	2	29
$42 \leq x < 45$	1	30



c median ≈ 35 cm
 d median = 34.5 cm; the median found from the graph is a good approximation.
 5 a ≈ 27 min b ≈ 29 min c ≈ 31.3 min
 d ≈ 4.3 min e ≈ 28.2 min

Time (t min)	$21 \leq t < 24$	$24 \leq t < 27$	$27 \leq t < 30$
Number of competitors	5	15	30

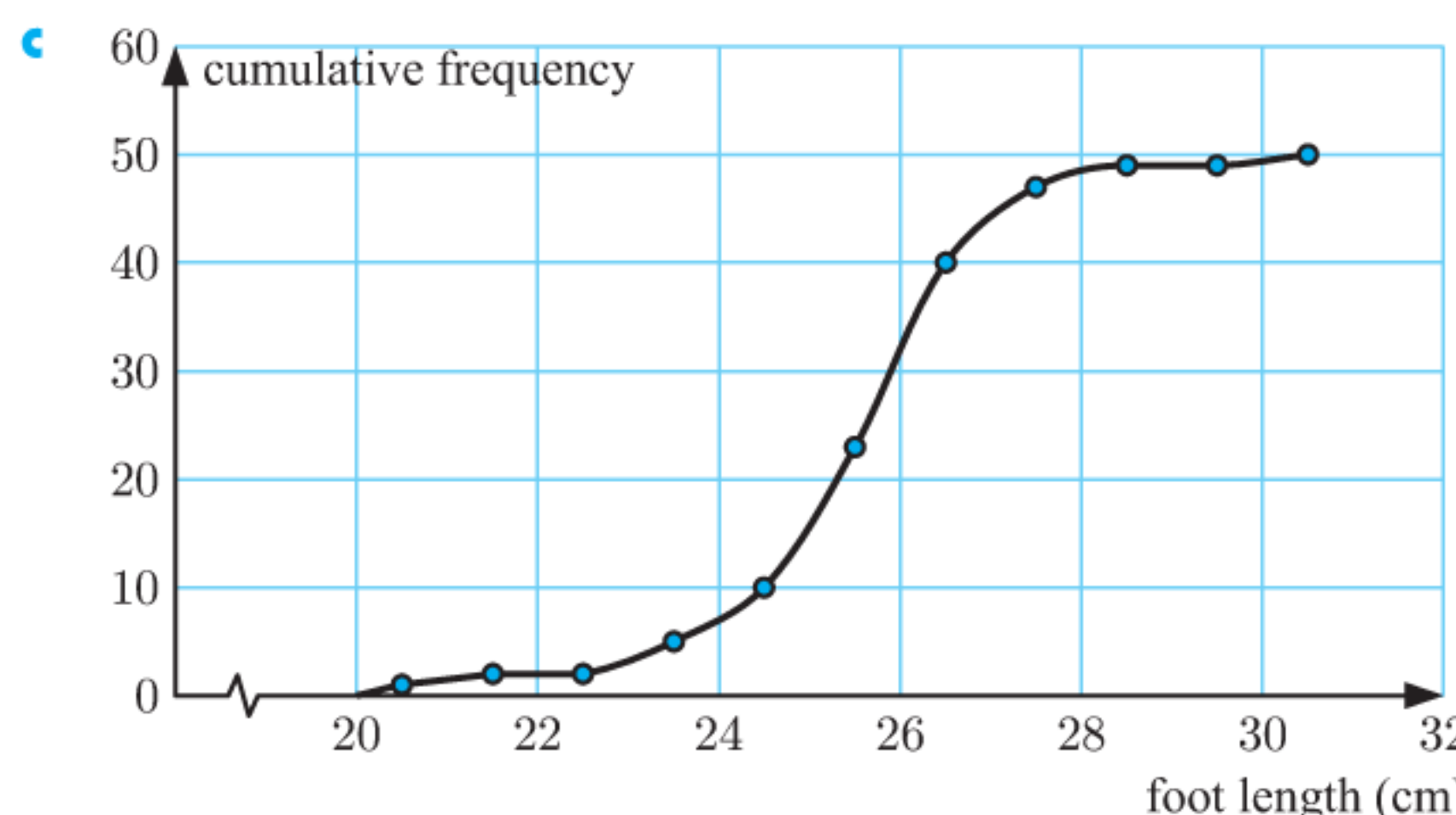
Time (t min)	$30 \leq t < 33$	$33 \leq t < 36$
Number of competitors	20	10



b ≈ 2280 hours c $\approx 71\%$ d ≈ 67

7 a $19.5 \leq l < 20.5$ cm

Foot length (cm)	Frequency	Cumulative frequency
$19.5 \leq l < 20.5$	1	1
$20.5 \leq l < 21.5$	1	2
$21.5 \leq l < 22.5$	0	2
$22.5 \leq l < 23.5$	3	5
$23.5 \leq l < 24.5$	5	10
$24.5 \leq l < 25.5$	13	23
$25.5 \leq l < 26.5$	17	40
$26.5 \leq l < 27.5$	7	47
$27.5 \leq l < 28.5$	2	49
$28.5 \leq l < 29.5$	0	49
$29.5 \leq l < 30.5$	1	50



d i ≈ 25.2 cm ii ≈ 18 people

EXERCISE 13J

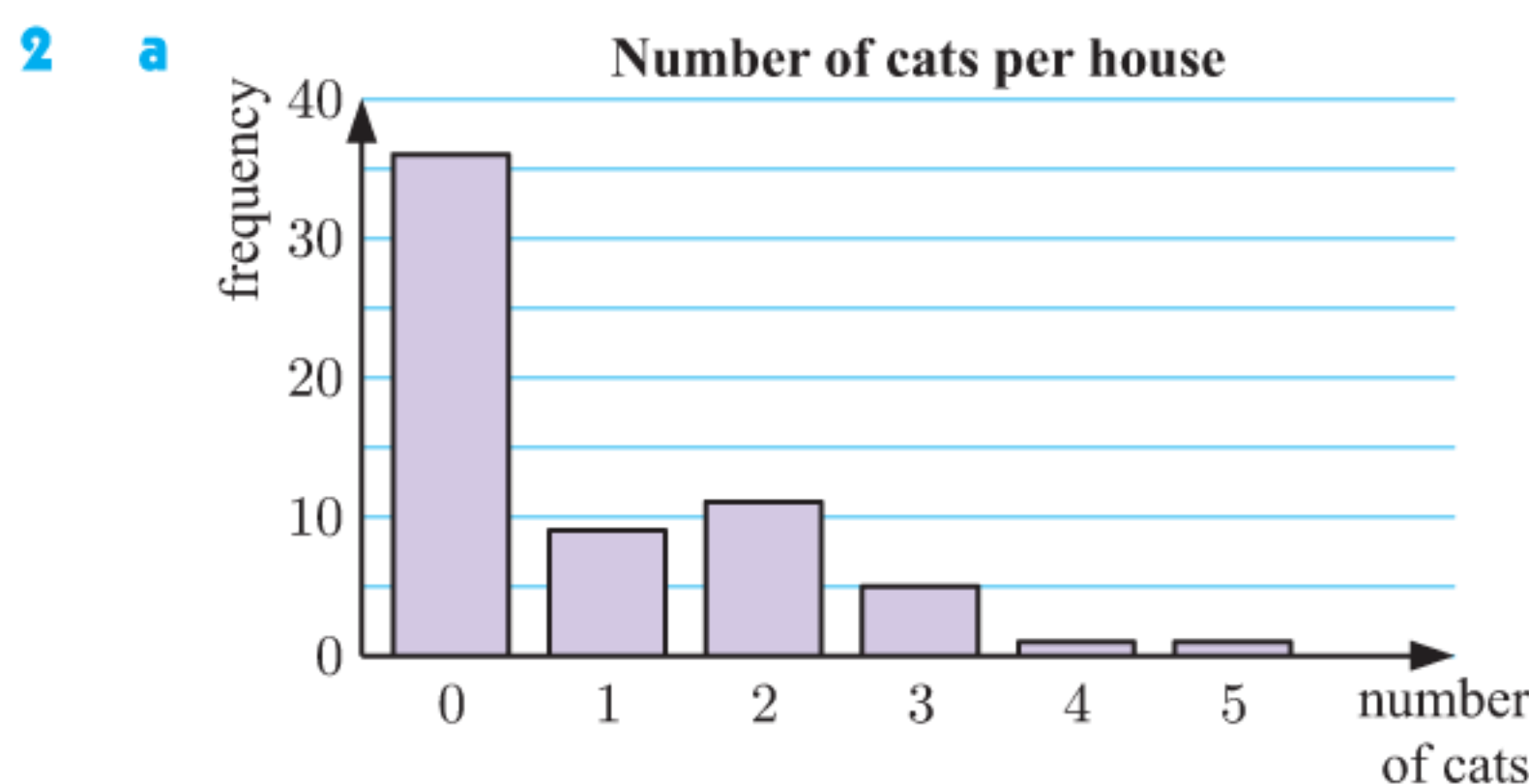
1 a Data set A: mean = $\frac{10 + 7 + 5 + 8 + 10}{5} = 8$
 Data set B: mean = $\frac{4 + 12 + 11 + 14 + 1 + 6}{6} = 8$

- b** Data set B appears to have a greater spread than data set A, as data set B has more values that are a long way from the mean, such as 1 and 14.
- c** Data set A: $\sigma^2 = 3.6$, $\sigma \approx 1.90$
Data set B: $\sigma^2 \approx 21.7$, $\sigma \approx 4.65$
- 2 a** The data is positively skewed. **b** $\sigma \approx 1.59$
c $\sigma^2 \approx 2.54$
- 3 a** $\mu = 24.25$, $\sigma \approx 3.07$ **b** $\mu = 28.25$, $\sigma \approx 3.07$
c If each data value is increased or decreased by the same amount, then the mean will also be increased or decreased by that amount, however the population standard deviation will be unchanged.
- 4** $\sigma \approx 2.64$, $s \approx 2.71$
- 5 a** Danny: ≈ 3.21 hours; Jennifer: 2 hours
b Danny
c Danny: $\sigma \approx 0.700$ hours, $s \approx 0.726$ hours;
Jennifer: $\sigma \approx 0.423$ hours, $s \approx 0.439$ hours
d Jennifer
- 6 a**
- | | Mean \bar{x} | Median | Standard deviation | | Range |
|-------|----------------|--------|--------------------|----------------|-------|
| | | | σ | s | |
| Boys | 32.02 | 31.05 | ≈ 4.52 | ≈ 4.77 | 13.8 |
| Girls | 34.77 | 35.85 | ≈ 3.76 | ≈ 3.96 | 11.7 |
- b i** boys **ii** boys
c Tyson could increase his sample size.
- 7 a** Rockets: mean = 5.7, range = 11
Bullets: mean = 5.7, range = 11
b We suspect the Rockets, since they twice scored zero runs.
Rockets: $\sigma = 3.9$, $s \approx 4.11$ ← greater variability
Bullets: $\sigma \approx 3.29$, $s \approx 3.47$
c standard deviation
- 8 a i** Museum: ≈ 934 visitors; Art gallery: ≈ 1230 visitors
ii Museum: $\sigma \approx 208$ visitors, $s \approx 211$ visitors;
Art gallery: $\sigma \approx 84.6$ visitors, $s \approx 86.0$ visitors
b the museum
c i '0' is an outlier.
ii This outlier corresponded to Christmas Day, so the museum was probably closed which meant there were no visitors on that day.
iii Yes, although the outlier is not an error, it is not a true reflection of a visitor count for a particular day.
iv Museum: mean ≈ 965 visitors, $\sigma \approx 121$ visitors,
 $s \approx 123$ visitors
v The outlier had greatly increased the population standard deviation.
- 9** $s_A > s_B$ does not imply that $\sigma_A > \sigma_B$.
Hint: Find a counter example.
- 10** $p = 6$, $q = 9$ **11** $a = 8$, $b = 6$ **12 b** $\mu = \pm 8.7$
- 13** $\sigma \approx 0.775$ **14** $\mu = 14.48$ years, $\sigma \approx 1.75$ years
- 15 a** Data set A **b** Data set A: 8, Data set B: 8
c Data set A: 2, Data set B: ≈ 1.06
Data set A does have a wider spread.
d The standard deviation takes all of the data values into account, not just two.
- 16 a** The female students' marks are in the range 16 to 20 whereas the male students' marks are in the range 12 to 19.
i the females **ii** the males
b Females: $\mu \approx 17.5$, $\sigma \approx 1.02$
Males: $\mu \approx 15.5$, $\sigma \approx 1.65$

- 17** The results for the mean will differ by 1, but the results for the standard deviation will be the same. Jess' question is worded so that the respondent will not include themselves.
- 18 a** $\bar{x} \approx 48.3$ cm **b** $\sigma \approx 2.66$ cm, $s \approx 2.70$ cm
- 19 a** $\bar{x} \approx 17.45$ **b** $\sigma \approx 7.87$, $s \approx 7.91$
- 20 a** $\bar{x} \approx \$780.60$ **b** $\sigma \approx \$31.74$, $s \approx \$31.82$
- 21 a** $\bar{x} = 40.35$ hours, $\sigma \approx 4.23$ hours, $s \approx 4.28$ hours
b $\bar{x} = 40.6$ hours, $\sigma \approx 4.10$ hours, $s \approx 4.15$ hours
The mean increases slightly; the standard deviation decreases slightly. These are good approximations.

REVIEW SET 13A

- 1 a i** ≈ 4.67 **ii** 5 **b i** 3.99 **ii** 3.9



- b** positively skewed
c i 0 cats **ii** ≈ 0.873 cats **iii** 0 cats
d The mean, as it suggests that some people have cats. (The mode and median are both 0.)
- 3 a**
- | Distribution | Girls | Boys |
|--------------|---------------|---------------|
| median | 36 s | 34.5 s |
| mean | 36 s | 34.45 s |
| modal class | 34.5 - 35.5 s | 34.5 - 35.5 s |
- b** The girls' distribution is positively skewed and the boys' distribution is approximately symmetrical. The median and mean swim times for boys are both about 1.5 seconds lower than for girls. Despite this, the distributions have the same modal class because of the skewness in the girls' distribution. The analysis supports the conjecture that boys generally swim faster than girls with less spread of times.
- 4** $a = 8$, $b = 6$
- 5 b** $k + 3$
- 6 a** We do not know each individual data value, only the intervals they fall in, so we cannot calculate the mean winning margin exactly.
b ≈ 22.6 points
- 7 a** min = 3, $Q_1 = 12$, med = 15, $Q_3 = 19$, max = 31
b range = 28, IQR = 7
c
-
- 8 a** 101.5 **b** 7.5 **c** 100.2 **d** ≈ 7.59
- 9 a** A: min = 11 s, $Q_1 = 11.6$ s, med = 12 s, $Q_3 = 12.6$ s, max = 13 s
B: min = 11.2 s, $Q_1 = 12$ s, med = 12.6 s, $Q_3 = 13.2$ s, max = 13.8 s
b A: range = 2.0 s, IQR = 1.0 s
B: range = 2.6 s, IQR = 1.2 s
c i A, the median time is lower.
ii B, the range and IQR are higher.
- 10 a** ≈ 58.5 s **b** ≈ 6 s **c** ≈ 53 s

11 a ≈ 88 students

b $m \approx 24$

Time (t min)	Frequency
$5 \leq t < 10$	20
$10 \leq t < 15$	40
$15 \leq t < 20$	48
$20 \leq t < 25$	42
$25 \leq t < 30$	28
$30 \leq t < 35$	17
$35 \leq t < 40$	5

12 a $\sigma^2 \approx 63.0, \sigma \approx 7.94$ b $\sigma^2 \approx 0.969, \sigma \approx 0.984$

13 a $\bar{x} \approx 49.6$ matches, $\sigma \approx 1.60$ matches, $s \approx 1.60$ matches

b The claim is not justified, but a larger sample is needed.

14 a $\bar{x} \approx 33.6$ L b $\sigma \approx 7.63$ L, $s \approx 7.66$ L

15 a No, extreme values have less effect on the standard deviation of a larger population.

b i mean ii standard deviation

c A low standard deviation means that the weight of biscuits in each packet is, on average, close to 250 g.

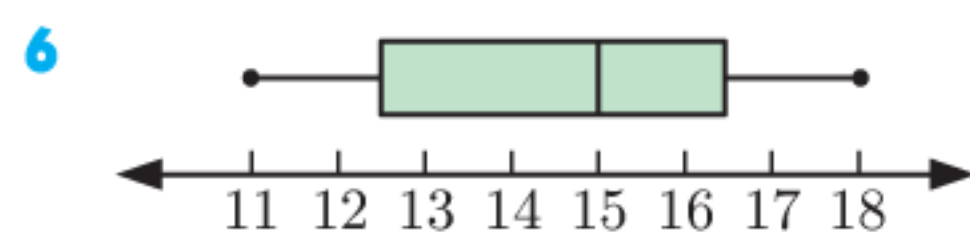
REVIEW SET 13B

	mean (seconds)	median (seconds)
Week 1	≈ 16.0	16.3
Week 2	≈ 15.1	15.1
Week 3	≈ 14.4	14.3
Week 4	14.0	14.0

b Yes, Heike's mean and median times have gradually decreased each week which indicates that her speed has improved over the 4 week period.

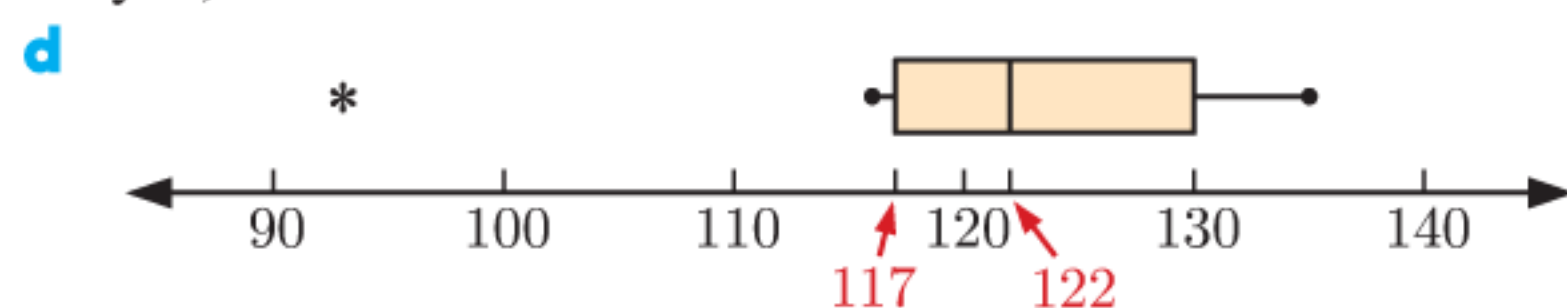
2 a 5 b 3.52 c 3.5 3 a $x = 7$ b 6

4 $p = 7, q = 9$ (or $p = 9, q = 7$) 5 ≈ 414 patrons

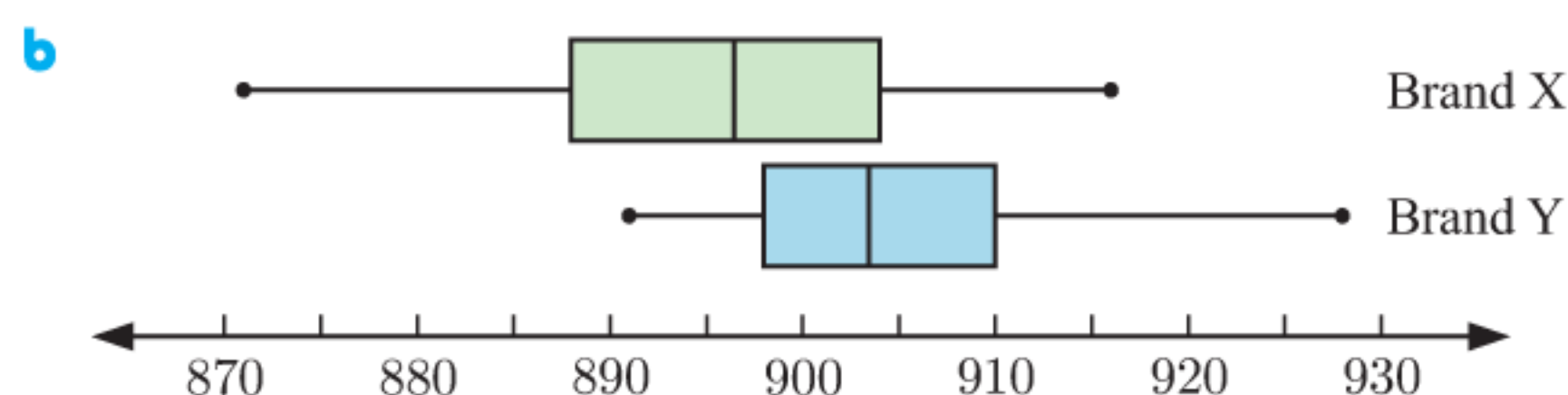


7 a $\sigma \approx 11.7, s \approx 12.4$ b $Q_1 = 117, Q_3 = 130$

c yes, 93



	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



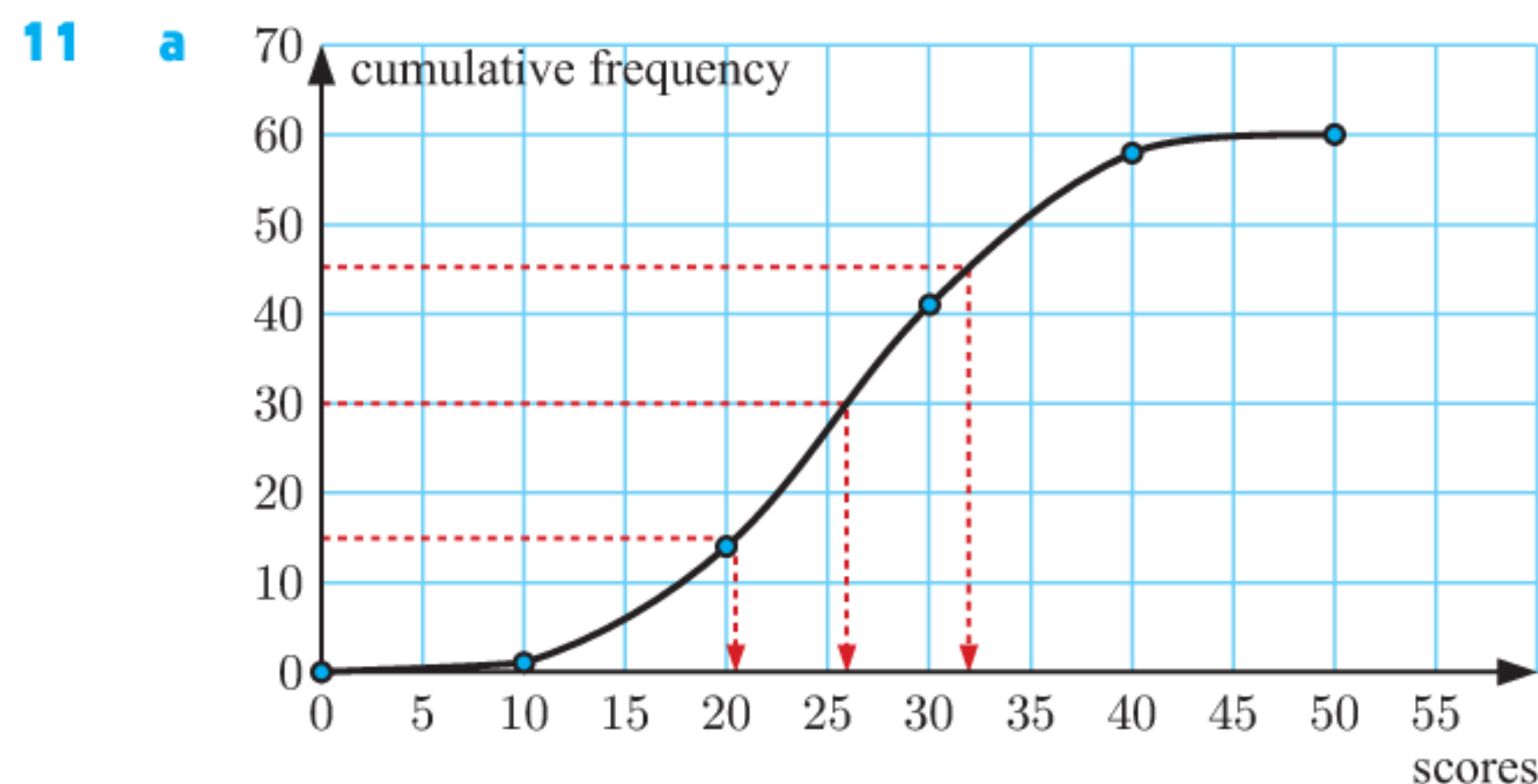
c i Brand Y, as the median is higher.
ii Brand Y, as the IQR is lower, so less variation.

9 a $p = 12, m = 6$

c $\bar{x} = \frac{254}{30} = \frac{127}{15}$

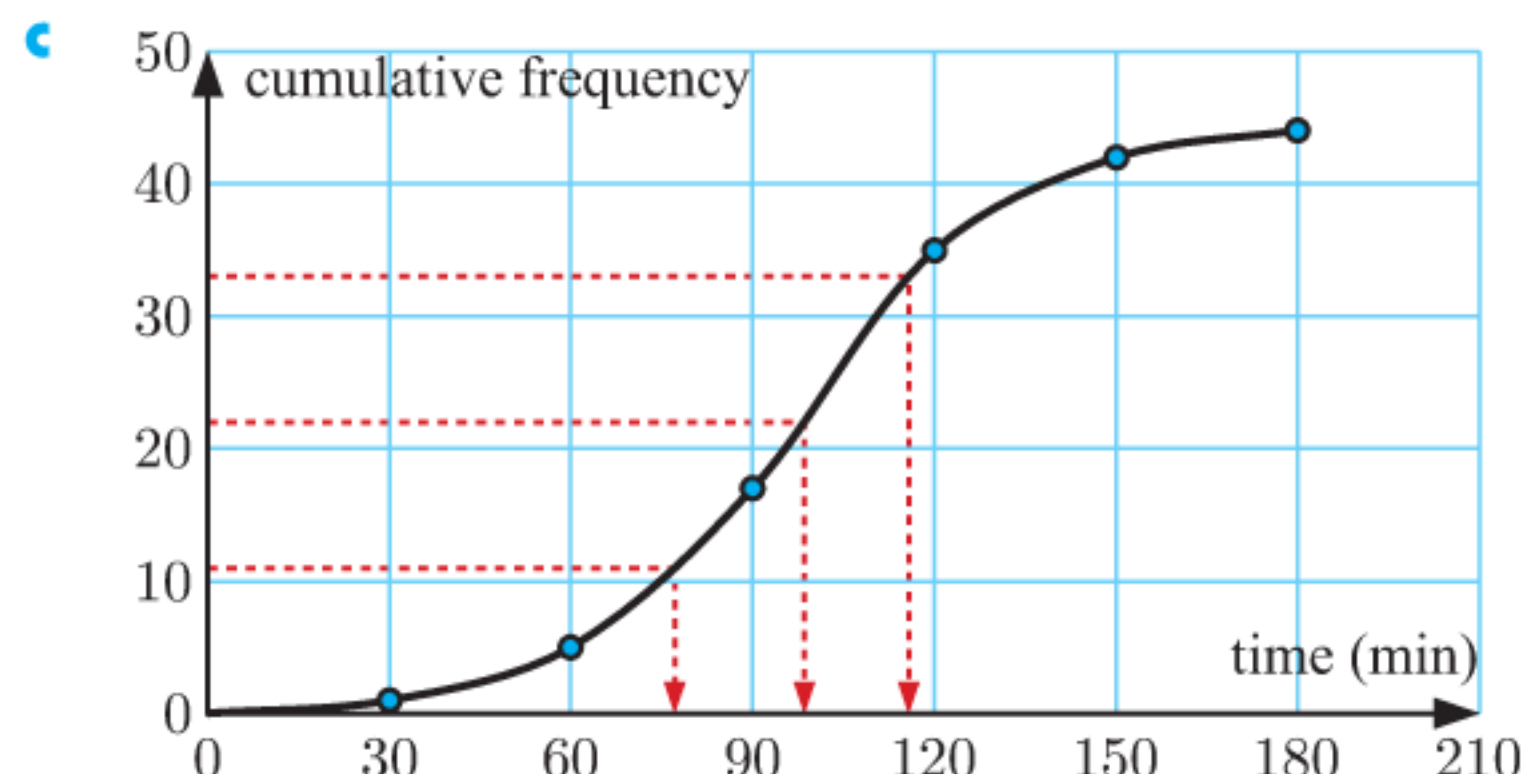
Measure	Value
mode	9
median	9
range	4

10 a ≈ 77 days b ≈ 12 days



b i median ≈ 26 ii IQR ≈ 11.5
iii $\bar{x} \approx 26.0$ iv $\sigma \approx 8.31$

12 a 44 players b $90 \leq t < 120$ min



d i ≈ 98.6 min ii ≈ 96.8 min iii no
e "... between 77.2 and 115.7 minutes."

13 a $\bar{x} \approx \text{€}207.02$ b $\sigma = \text{€}38.80, s \approx \text{€}38.89$

14 a Kevin: $\bar{x} = 41.2$ min; Felicity: $\bar{x} = 39.5$ min

b Kevin: $\sigma \approx 7.61$ min, $s \approx 7.81$ min;
Felicity: $\sigma \approx 9.22$ min, $s \approx 9.46$ min

c Felicity d Kevin

15 10 data values

EXERCISE 14A

1 a $y = x^2 - 3x + 1$

x	-2	-1	0	1	2
y	11	5	1	-1	-1

b $y = x^2 + 2x - 5$

x	-2	-1	0	1	2
y	-5	-6	-5	-2	3

c $y = 2x^2 - x + 3$

x	-4	-2	0	2	4
y	39	13	3	9	31

d $y = -3x^2 + 2x + 4$

x	-4	-2	0	2	4
y	-52	-12	4	-4	-36

2 a no b yes c yes d yes e no f yes

3 a $x = -1$ or -2 b $x = 2$ c $x = 1$ or 5

d $x = -3$ or $\frac{1}{2}$ e $x = -6$ or 1 f no real solutions

EXERCISE 14B.1

