Chapter 11 — Sampling, Data Presentation and Interpretation

1. Populations and Sampling

For any statistical investigation, there will be a group of something (e.g. people, items, animals, etc.) that you want to find out about. This whole group is called the population. Populations are finite if the members are countable (in practice, not just in theory), and infinite if they are uncountable.

To collect information about your population, you can carry out a survey. This means questioning the people or examining the items.

Censuses and Sampling
When you collect information from every member of a population, it's called a census — it's a survey of the whole population. It's easier to carry out if the population is fairly small and easily accessible.

**Advantages**
- It's an accurate representation of the population because every member has been surveyed — it's unbiased.

**Disadvantages**
- For large populations, it takes a lot of time and effort (and possibly money) to carry out.
- It can be difficult to make sure all members are surveyed. If some are missed, the survey may be biased.
- If the tested items are used up or damaged in some way by doing a census, a census is impractical.

If doing a census is impossible or impractical, you can find out about a population by questioning or examining a selected group of the people or items — called a sample.

**Advantages**
- Sample surveys are quicker and cheaper than a census, and it's easier to get hold of all the required information.
- It's the only option when surveyed items are used up or damaged.

**Disadvantages**
- Each possible sample will give different results, so you could just happen to select one which doesn't accurately reflect the population.
- Samples can easily be affected by sampling bias.

It's important that the sample is as similar to the population as possible so that it is representative. Otherwise, it may be biased, and conclusions about the population based on your sample may not be correct.

To avoid sampling bias:
- Select from the correct population and make sure no member of the population is excluded.
- Select your sample at random — if members are linked in some way, it can cause bias.
- Make sure all your sample members respond.

Exercise 1.1

Q1 For each population described say whether it is finite or infinite.
   a) The grains of sand on a beach.
   b) The 2016 Olympic gold medalists.
   c) The stars in the Milky Way galaxy.
   d) The cells in a human body.
   e) The members of the Ulverston Musical Appreciation Society.
   f) The jalapeno chilli plants on sale at Church Lane Garden Centre.

Q2 Members of a local book club have to be consulted about the next book they'll read.
   a) What is the population? b) Explain whether a census or a sample survey should be used.

Q3 For each situation below, explain whether a census or a sample survey should be done:
   a) Marcel is in charge of a packaging department of 8 people.
   b) A toy manufacturer produces batches of 500 toys. As part of a safety check, they want to test the toys to work out the strength needed to pull them apart.
   c) Tara has a biased dice. She wants to find the proportion of dice rolls that will result in a 'three'.

Q4 Pooja is doing a survey on whether people buy ethically-sourced products. She asked her mother to hand out questionnaires to 20 of her friends. Why might this sample be biased?

Sampling Methods

**Simple random sampling**
Simple random sampling is where every person or item in the population has an equal chance of being in the sample, and each selection is independent of the others.

To choose a simple random sample:
- Give a number to each population member, from a full list of the population.
- Generate a list of random numbers and match them to the numbered members to select your sample.

**Advantage**
Every member of the population has an equal chance of being selected, so it's completely unbiased.

**Disadvantage**
It can be inconvenient if the population is spread over a large area — it might be difficult to track down the selected members (e.g. in a nationwide sample).

**Example**
A zoo has 80 cotton-top tamarins. Describe how the random-number table given could be used to select a sample of five of them, for a study on tail length.

1. Draw up a list of the 80 cotton-top tamarins, and give each tamarin a unique 2-digit number between 01 and 80.
2. Use the random-number table to choose five numbers. Five random numbers can be chosen by reading off every two digits that give a unique number between 01 and 80.
   The first five numbers are: 69 (too big), 30, 39, 99 (too big), 18, 40, 03
3. Select the cotton-top tamarins with the numbers 30, 39, 18, 40 and 03.
Systematic sampling

Systematic sampling selects every \( n \)th member from the population you're investigating.

**Advantages**
- It can be used for quality control on a production line — a machine can be set up to sample every \( n \)th item.
- It should give an unbiased sample.

**Disadvantage**
- If the interval coincides with a pattern in the population, the sample could be biased.

**Example**

50,000 fans attended a football match. Describe how systematic sampling could be used to select a sample of 100 people.

1. Give each fan a 5-digit number between 00001 and 50000.
2. 50000 ÷ 100 = 500, so select every 500th fan.
3. Use a calculator to randomly generate a starting point between 1 and 500.
4. Keep adding 500 to the starting point to find the rest of the sample.

E.g. If 239 is randomly generated, then select the fans numbered:
00239, 06739, 01239, ..., 48739, 49239, 49739

Exercise 1.2

Q1 The animals in a zoo are given a unique 3-digit ID number between 001 and 500. Describe how you could use a random number generator to choose a simple random sample of 20 of the zoo's animals.

Q2 A teacher wants to interview a sample of 10 students from his school of 700 students.
   a) Explain why this is not a simple random sample.
   b) The teacher has a list of every student in the school, sorted by age.
   c) Give an advantage of the sampling method described in part b) over his original sampling method.

Q3 All dogs which are admitted to the Greyway Animal Sanctuary are microchipped with a unique identification number. 108 of the dogs at the sanctuary were admitted between 2015 and 2016. A sample of 12 of these dogs is selected for long-term monitoring.
   a) What is the population?
   b) Explain how to select this sample using (i) simple random sampling (ii) systematic sampling.

Q4 A high-street store is investigating the number of customers they get each day over the course of one year. They record how many people make a purchase from their shop per day, on each of 50 days during the year. They use systematic sampling to choose which 50 days to sample. Explain why this sampling method might lead to bias in their data.

Stratified sampling

If a population is divided into categories (e.g. age or gender), you can use a stratified sample — this uses the same proportion of each category in the sample as there is in the population.

**To choose a stratified sample:**
- Divide the population into categories.
- Calculate the number needed for each category in the sample using the formula:
  \[
  \frac{\text{size of category in pop.}}{\text{size of category in sample}} = \frac{\text{size of pop.}}{\text{total sample size}}
  \]
- Randomly select the sample for each category.

**Advantages**
- If the categories are disjoint (i.e. there is no overlap, e.g. age groups), this should give a representative sample.
- It's useful when results may vary depending on categories.

**Disadvantage**
- The extra detail needed can make it expensive.

**Example**

A teacher takes a sample of 20 pupils from her school, stratified by year group. The table below shows the number of pupils in each year group. Calculate how many pupils from each year group should be in her sample.

<table>
<thead>
<tr>
<th>Year</th>
<th>Group</th>
<th>No. of pupils</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Group</td>
<td>120</td>
</tr>
<tr>
<td>8</td>
<td>Group</td>
<td>80</td>
</tr>
<tr>
<td>9</td>
<td>Group</td>
<td>95</td>
</tr>
<tr>
<td>10</td>
<td>Group</td>
<td>63</td>
</tr>
<tr>
<td>11</td>
<td>Group</td>
<td>42</td>
</tr>
</tbody>
</table>

Quota sampling

Quota sampling is often used in market research. The interviewer will be given a quota of people in each category to interview (e.g. 20 men and 20 women).

They then choose people to interview until the quotas are fulfilled.

**To choose a quota sample:**
- Divide the population into categories.
- Give each category a quota (number of members to sample).
- Collect data until the quotas are met in all categories (without using random sampling).

**Advantages**
- It is easy for the interviewer as they don't need access to the whole population, or a list of every member.
- The interviewer continues to sample until all the quotas are met, so non-response is less of a problem.

**Disadvantage**
- It can be biased by the interviewer — selection isn't random, so they might exclude some of the population.

**Example**

A video-game company wants to gather opinions on a new game. The interviewer is asked to interview 75 people aged under 30, and 25 people aged 30+. Give one advantage and one disadvantage of this quota sample.

**Advantages**
- The company doesn't have a full list of everyone who has played the game, so random sampling isn't possible.

**Disadvantage**
- People with strong views on the game are more likely to respond to the interviewer, which may cause sampling bias.
Exercise 1.3

Q1 A sports centre selects a sample of 10 members, stratified by age. The table shows the total number of members in each age group.

<table>
<thead>
<tr>
<th>Age (a)</th>
<th>Under 20</th>
<th>20 to 40</th>
<th>41 to 60</th>
<th>Over 60</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of members</td>
<td>45</td>
<td>33</td>
<td>15</td>
<td>57</td>
</tr>
</tbody>
</table>

Calculate how many people from each age group should be sampled.

Q2 Martin is investigating the heights of athletes at a competition. He wants to measure 10 males and 10 females taking part in each of the high jump, the long jump and the pole vault. Give one advantage and one disadvantage of using quota sampling in this situation.

Q3 Numbria council wants to collect residents' opinions on public transport, using a stratified sample of 200 residents. The table shows the number of residents, to the nearest 100, in each area.

<table>
<thead>
<tr>
<th>Area</th>
<th>Addipati</th>
<th>Loosemoor</th>
<th>Coniston</th>
<th>Logby Bridge</th>
<th>Anglesea</th>
<th>Gradebeck</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of residents</td>
<td>7000</td>
<td>8600</td>
<td>4200</td>
<td>6000</td>
<td>5400</td>
<td>8000</td>
</tr>
</tbody>
</table>

a) Explain why stratified sampling might be better than simple random sampling in this case.
b) Calculate how many residents from each area should be in the sample.
c) They use a simple random sample to select the required number of residents in each area, and send them a survey through the post, asking them to fill in and send back. Suggest why this could lead to bias in the sample.

Opportunity sampling

Opportunity (convenience) sampling is where the sample is chosen from a section of the population that is most convenient for the sampler.

To choose an opportunity sample:
Choose members of the population that are the easiest to sample — e.g. ask the first people you meet or sample whatever products you can find.

Advantage: Data can be gathered very quickly and easily.
Disadvantage: It isn't random and can be very biased — there's no attempt to make the sample representative.

Examples:
Mel thinks that most people watch her favourite television programme. She asks 20 friends whether they watch the television programme.

a) Name the sampling method Mel used.
Mel asked her friends because they're easily available to sample, so:
Opportunity (convenience) sampling

b) Give two reasons why Mel's sample may be biased.
1. Mel's friends could be of a similar age or the same gender, which is not representative of the whole population.
2. Because this is Mel's favourite television programme, she might have encouraged her friends to watch it too.

Exercise 1.4

Q1 James wants to find out what colour of car is most common. He stands by a busy road and records the colour of the first 100 cars that pass by. Give an advantage and a disadvantage of this sampling method.

Q2 A biologist is collecting data on the butterflies in a forest. She needs to capture 30 butterflies and measure their wingspan. Suggest why opportunity sampling might be appropriate in this case.

Q3 Hattie has a biased coin. To investigate the probability of getting heads, she takes the results of the first 50 times she flips the coin. Is this data likely to be biased? Explain your answer.

Q4 For a school project, Neville is investigating the types of music people in the UK like to listen to. He collects data by asking friends from his year group. Is this sample likely to be representative of the population? Give one way in which the sample could be improved.

Cluster sampling

Cluster sampling is another method that's useful when the population can be divided into distinct groups. The clusters should be groups that you expect to give similar results to each other.

To choose a cluster sample:
- Divide the population into clusters covering the whole population, where no member of the population belongs to multiple clusters.
- Randomly select clusters to use in the sample, based on the required sample size.
- Either use all the members of the selected clusters (a one-stage cluster sample), or randomly sample within each cluster to form the sample (a two-stage cluster sample).

Advantages
- It can be more practical (e.g. quicker or cheaper) in certain situations.
- You can incorporate other sampling methods, making it quite adaptable.

Disadvantages
- Because you only sample certain clusters, the results could be less representative.
- It's not always possible to separate a population into clusters in a natural way.

You can use any of the other sampling techniques for either stage of the sample (selecting the clusters or choosing which members of the cluster to use in the sample), depending on the circumstances.

Examples:
A researcher wants to conduct a taste test of a new energy drink on university students in the UK.

a) Explain why a cluster sample might be suitable in this situation.
Taking universities in the UK as the clusters, all university students are included, and you can assume that no student belongs to multiple universities. You would also expect different universities to give similar results.

b) Explain whether a one-stage or a two-stage cluster sample would be more appropriate.
Conducting a one stage test on every student at each university would be expensive and difficult to organise. It would be better to select a sample of the students and conduct the tests on them, so a two-stage cluster sample would be more appropriate.
Exercise 1.5

Q1 Will wants to know about the number of people that cycle to work across his company. He suggests using a one-stage cluster sample, with the different departments as the clusters.
   a) Give a possible reason why Will’s suggested method may not be suitable in this situation.
   b) Suggest a different sampling method Will could use and give one advantage of this method.

Q2 A manufacturing company makes machines that produce widgets. They have a list of all of the machines that they have made, and the addresses of the companies that own them. They want to collect data on the proportion of misshapen widgets that their machines produce, minimizing the amount of time spent travelling from site to site while still getting a representative sample.
   Describe how a cluster sample could be used to collect this data, stating whether a one-stage or a two-stage cluster sample would be more appropriate.

Self-selection sampling
Self-selection (or volunteer) sampling is where people choose to be part of the sample — e.g. they choose to complete a questionnaire or volunteer to take part in a study.

Advantages
- It requires little time or effort in finding sample members, as they contact you.
- People who have volunteered are less likely to not respond.
- It could be the only way to get people to take part in a study, or to find members of a population.

Disadvantage
There can easily be trends within the respondents, such as people having strong opinions, which would lead to bias.

Example
A website puts an advert on their homepage, asking visitors to complete a short survey about the site for a payment of £20.

Give three ways that this sampling method could cause bias in their results.

1. People who are willing to complete the survey might have stronger opinions than other visitors, which would introduce bias into the results.
2. Certain people might not see the advert if the website does not display properly for them, or if they block adverts through their browser, which would exclude parts of the population from the sample, making it less representative.
3. Since the website is offering money for responses, people might be more likely to try to fill out the survey multiple times or not take the questions seriously, which would make the responses less representative of the views of all of their visitors.

Exercise 1.6

Q1 The manager of a park wants to collect opinions from visitors about a potential building project. Describe how she could use self-selection sampling to do this and give an advantage of this method.

Q2 David gives out a questionnaire about reading habits to the students in his year group at school. He gets 70 responses from boys and 40 from girls. As his sample, he decides to use all of the responses from the girls, and 40 randomly selected responses from the boys, to match the proportion of boys and girls in his year group. Give an advantage and a disadvantage of this method.

Chapter 11 Sampling, Data Presentation and Interpretation

Chapter 11 Sampling, Data Presentation and Interpretation
You can represent data using different diagrams, depending on whether the data is discrete or continuous. **Frequency polygons** are one type that are used for continuous data.

**Example**
The table shows the maximum daily temperature (°C) in a town over 100 days.

<table>
<thead>
<tr>
<th>Maximum daily temperature (°C)</th>
<th>10 &lt; t ≤ 15</th>
<th>15 &lt; t ≤ 20</th>
<th>20 &lt; t ≤ 25</th>
<th>25 &lt; t ≤ 30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>21</td>
<td>98</td>
<td>56</td>
<td>5</td>
</tr>
</tbody>
</table>

**Exercise 2.1**

Q1  A mechanic collects the following information about cars he services:

- Make, Mileage, Colour, Number of doors, Cost of service

Write down all the variables from this list that are:

a) qualitative  
b) quantitative

Q2  Amy is an athletics coach. She records the following information about each of the athletes she trains.

- Number of medals won last season, Height, Mass, Shoe size

Write down all the variables from this list that are examples of:

a) discrete quantitative variables  
b) continuous quantitative variables

Q3  A group of 50 people were given 30 seconds to see how many words they could make out of a set of 10 letters. The bar chart shows the results.

a) Is the data discrete or continuous?

b) How many people found at least 11 words?

Q4  The heights of the members of a history society are shown in the table.

- Explain why ‘height’ is a continuous variable.
- For each class, find:
  - (i) the lower class boundary
  - (ii) the upper class boundary
  - (iii) the class width
- Show the information in the table in a frequency polygon.

<table>
<thead>
<tr>
<th>Height, H (cm)</th>
<th>Number of members</th>
</tr>
</thead>
<tbody>
<tr>
<td>140 ≤ H &lt; 150</td>
<td>3</td>
</tr>
<tr>
<td>150 ≤ H &lt; 160</td>
<td>9</td>
</tr>
<tr>
<td>160 ≤ H &lt; 170</td>
<td>12</td>
</tr>
<tr>
<td>170 ≤ H &lt; 180</td>
<td>5</td>
</tr>
<tr>
<td>180 ≤ H &lt; 190</td>
<td>4</td>
</tr>
<tr>
<td>190 ≤ H &lt; 200</td>
<td>1</td>
</tr>
</tbody>
</table>

Histograms

Histograms look like bar charts. However, as they're used for continuous variables, there are **no gaps** between the bars, and the heights of the bars show **frequency density**, rather than frequency, where:

$$\text{Frequency density} = \frac{\text{frequency}}{\text{class width}}$$

**Example**
The masses (to the nearest 10 g) of 1000 parcels are given in the table below. Draw a histogram to show this data.

<table>
<thead>
<tr>
<th>Mass of parcel (to nearest 10 g)</th>
<th>100-200</th>
<th>200-300</th>
<th>300-400</th>
<th>400-500</th>
<th>500-600</th>
<th>600-700</th>
<th>700-800</th>
<th>800-900</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of parcels</td>
<td>100</td>
<td>250</td>
<td>300</td>
<td>400</td>
<td>500</td>
<td>600</td>
<td>700</td>
<td>800</td>
</tr>
</tbody>
</table>

1. Draw a table showing the class boundaries, class width and frequency density.

<table>
<thead>
<tr>
<th>Mass of parcel (g)</th>
<th>Lower class boundary (g)</th>
<th>Upper class boundary (g)</th>
<th>Class width</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>100-200</td>
<td>100</td>
<td>200</td>
<td>100</td>
<td>0.5</td>
</tr>
<tr>
<td>200-300</td>
<td>200</td>
<td>300</td>
<td>100</td>
<td>1.25</td>
</tr>
<tr>
<td>300-400</td>
<td>300</td>
<td>400</td>
<td>100</td>
<td>2.0</td>
</tr>
<tr>
<td>400-500</td>
<td>400</td>
<td>500</td>
<td>100</td>
<td>3.0</td>
</tr>
<tr>
<td>500-600</td>
<td>500</td>
<td>600</td>
<td>100</td>
<td>4.0</td>
</tr>
<tr>
<td>600-700</td>
<td>600</td>
<td>700</td>
<td>100</td>
<td>5.0</td>
</tr>
<tr>
<td>700-800</td>
<td>700</td>
<td>800</td>
<td>100</td>
<td>6.0</td>
</tr>
<tr>
<td>800-900</td>
<td>800</td>
<td>900</td>
<td>100</td>
<td>7.0</td>
</tr>
</tbody>
</table>

2. Draw the histogram — plot frequency density on the vertical axis and mass along the horizontal axis.

On a histogram, the frequency in a class is proportional to the area of its bar. In other words, frequency = k x area of bar (where k is a number).

**Example**
The histogram shows the heights of a group of people. There were 12 people between 155 cm and 160 cm tall.

a) How many people in the group are between 150 cm and 155 cm tall?

1. Use the bar for 155-160 cm to find the frequency.

Area of 155-160 cm bar = 1 x 6 = 6 squares

Frequency = k x area ⇒ k = 12 ÷ 6 = 2

2. Now find the area of the 130-155 cm bar and multiply by k to get the frequency.

Area of 130-155 cm bar = 5 x 2 = 10 ⇒ frequency = 2 x 10 = 20 people

b) Estimate the number of people in the group who are over 175 cm tall.

Assuming that the heights are evenly spread throughout the 170-190 cm class, one-fourth of the people in the class would be over 175 cm tall.

Area of 170-190 cm bar = 4 x 4 = 16 ⇒ frequency = k x area = 2 x 16 = 32 people

The estimated number over 175 cm = 32 x 3 = 96 people.
Exercise 2.2

Q1 The table shows the percentage of water remaining in a cyclist's bottle after their first lap, over 20 races. Draw a histogram to show the data.

Q2 The histogram shows the audition times (in seconds) for contestants applying for a place on a television talent show. The auditions for 54 contestants lasted between 30 and 45 seconds.
   a) Work out the number of contestants whose auditions lasted less than 30 seconds.
   b) How many contestants auditioned altogether?
   c) Estimate the number of contestants whose auditions lasted less than 40 seconds.
   d) A contestant is chosen at random. Find the probability that their audition lasted at least one minute.
   e) Estimate the probability that a randomly chosen contestant's audition lasted less than 25 seconds.

Q3 A butterfly enthusiast measures the wingspan (w, in mm), to the nearest millimetre, of a sample of tortoiseshell butterflies. She groups her measurements and displays the data in a histogram.
   The group containing butterflies with 44.5 ≤ w < 47.5 has a frequency of 12. This group is represented on the histogram by a bar of width 1.5 cm and height 9 cm.
   a) Show that each butterfly is represented on the histogram by an area of 1.125 cm².
   b) The bar representing the butterflies with 51.5 ≤ w ≤ 53.5 has an area of 22.5 cm².
   Work out the frequency for this group.
   c) The frequency for butterflies with 53.5 ≤ w ≤ 58.5 is 14.
   Work out the width and the height of the bar used to represent this group.

Chapter 11 Sampling, Data Presentation and Interpretation

Stem and leaf diagrams

Stem and leaf diagrams are another way to represent data. Each value is split into a 'stem' and a 'leaf', like this:

A stem and leaf diagram always needs a key to tell you what the stems and leaves represent. In this example, the stems represent the 'tens', so the first row shows the values 11, 13 and 17.

The leaves in each row should be sorted from lowest to highest.

Example

The lengths, in m, of cars in a car park were measured to the nearest 10 cm.

The results were: 2.9, 3.5, 4.0, 2.8, 4.1, 2.7, 3.1, 3.6, 3.8 and 3.7.

Draw a stem and leaf diagram to show the data.

1. Use the numbers before the decimal point as the stems, and the numbers after the decimal point as the leaves.

2. Put the leaves in order and add a key.

A back-to-back stem and leaf diagram is two normal stem and leaf diagrams drawn either side of the same stems. The left-hand diagram is read 'backwards' because the stems are on the opposite side.

Example

The heights of boys and girls in a Year 11 class are given to the nearest cm in the back-to-back stem and leaf diagram on the right.

Write out the data in full.

1. The boys’ heights are read backwards, so the first value, 15[1], is a height of 165 cm.

2. The first value for the girls, [15][1], is a height of 159 cm.

The girls’ heights in cm are: 159, 161, 165, 169, 170, 172, 173, 175, 180

Exercise 2.3

Q1 The stem and leaf diagram shows the number of members in each of the chains taking part in a choir competition.
   a) How many choirs were competing in this competition altogether?
   b) How many choirs had 45 members?
   c) How many choirs had more than 52 members?
   d) What was the size of the largest choir?

Q2 Construct a stem and leaf diagram for the following data:


Chapter 11 Sampling, Data Presentation and Interpretation
Q3 The stem and leaf diagram shows the minimum recorded temperature (to the nearest 0.1 °C) in 24 towns on a June night.
   a) How many towns had a minimum temperature below 10 °C?
   b) Which temperature was recorded most frequently?
   c) What was the difference between the highest and lowest temperatures recorded in these towns?

Q4 Sixteen children belong to an orchestra. The distances (in kilometres) that the children live from the practice hall are shown below.

   2.4 3.5 4.3 1.5 1.0 8.7 5.8 5.4 1.6 1.2
   3.2 1.6 4.3 2.5 3.9 2.1 4.0 2.5 6.4

   a) Draw a stem and leaf diagram to show this information.
   b) (i) Find the furthest distance that a child lives from the practice hall.
   (ii) How many children live less than 2 km from the practice hall?
   c) How many of his dramas last more than 90 minutes?
   d) How many of his films last between 70 and 80 minutes?

Q5 Freddie has a collection of films on DVD.
   The back-to-back stem and leaf diagram shows the running times (in the nearest minute) of these films.
   a) What is the running time of the shortest comedy film?
   b) How many of his dramas last more than 90 minutes?
   c) How many of his films last between 70 and 80 minutes?

Describing distributions
There are a number of terms you can use to describe the distribution of a quantitative data set:
- **Symmetrical**: the data is symmetrical about the mean and median, which are roughly equal.
- **Positively Skewed**: the data is concentrated in the lower part of the range (to the left).
- **Negatively Skewed**: the data is concentrated in the upper part of the range (to the right).
- **Unimodal**: the data has one peak where the distribution 'peaks'.
- **Bimodal**: the data has two peaks where the distribution 'peaks'.

Exercise 2.4

Q1 The diagram to the right shows the number of pets in each house in two different neighbourhoods, A and B. For each neighbourhood, is the data positively skewed, negatively skewed, or symmetrical?

Q2 A company collects data on the heights of 1000 40-year-olds (500 men and 500 women). They plot a histogram to show the data, which appears to be bimodal. Explain why this might be the case.

3. Location: Mean, Median and Mode

**Measures of location** (or of central tendency) summarise where the 'centre' of the data lies. The most common measure of location is the mean — the formula for the mean (\(\bar{x}\); read 'x-bar') is:

\[
\text{Mean} = \bar{x} = \frac{\sum x}{n} \quad \text{or} \quad \bar{x} = \frac{\sum fx}{f}
\]

where each \(x\) is a data value, \(f\) is the frequency of each \(x\) (the number of times it occurs), and \(n\) is the total number of data values.

The \(\sum\) (sigma) means "sum" (see p.293) — so \(\sum x\) means add up all the values of \(x\), and \(\sum f = n\).

**Example** A scientist counts the number of eggs in some honey bee nests.
His data is shown in this table. Calculate the mean number of eggs in these nests.

<table>
<thead>
<tr>
<th>Number of eggs, (x)</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of nests, (f)</td>
<td>4</td>
<td>9</td>
<td>16</td>
<td>8</td>
<td>3</td>
</tr>
</tbody>
</table>

1. Add a row showing the values of \(fx\) and a column showing the totals \(\sum f\) and \(\sum fx\) to the table.

<table>
<thead>
<tr>
<th>(x)</th>
<th>(f)</th>
<th>(fx)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>27</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>64</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>40</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>18</td>
</tr>
</tbody>
</table>

2. Now use the formula for the mean.

\[
\bar{x} = \frac{\sum fx}{\sum f} = \frac{157}{40} = 3.925 \text{ eggs}
\]

If you know a data set of size \(n_1\) has mean \(\bar{x}_1\) and another data set of size \(n_2\) has mean \(\bar{x}_2\), then the combined mean is:

\[
\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}
\]

**Example** A scientist is looking at the amount of rainfall over a week. The mean of the first 5 days is \(\bar{x} = 4.1\) mm and the mean of the next 2 days is \(\bar{x} = 19.9\) mm. Find the combined mean (\(\bar{x}\)) of the rainfall over the week.

\[
\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{(5 \times 4.1) + (2 \times 19.9)}{5 + 2} = \frac{60.3}{7} = 8.6142... = 8.61 \text{ mm (3 s.f.)}
\]

Exercise 3.1

Katia visits 12 shops and records the price of a loaf of bread. Her results are given below:

- £1.01, £1.19, £1.25, £1.19, £1.26, £1.24, £1.15, £1.09, £1.16, £1.20, £1.05, £1.10

Work out the mean price of a loaf of bread in these shops.
Q2: In a competition, the 7 members of team A scored a mean of 35 points, and the 6 members of team B scored 252 points altogether. Find the combined mean score.

Q3: The total number of goals scored in 20 football matches are shown in the table. Calculate the mean number of goals scored in these matches.

<table>
<thead>
<tr>
<th>Number of goals</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>

Q4: A drama group has 15 members. The mean age of the members is 47.4 years.
   a) A 17-year-old leaves the drama group. Find the new mean age.
   b) A new member then joins, making the mean age 47.6 years. How old is the new member?

There are two other important measures of location: the median and the mode.

- **Median**: The value in the middle of the data set when all the data values are placed in order of size.

To find the median, put the data values in order, then:
- If \( \frac{n}{2} \) is not a whole number, round it up to find the position of the median.
- If \( \frac{n}{2} \) is a whole number, the median is halfway between the value in this position and the next value.

**Mode**: Most frequently occurring data value.

A data set can have multiple modes, if it has only one, it’s called unimodal, and if it has two or more, it’s bimodal. If each data value occurs only once (common when the data is continuous), then the data set has no mode.

**Example:** Find the median and mode of the following data set:

4, 3, 11, 4, 8, 9, 3, 6, 7, 8

1. Put the data in order and use the rule for \( \frac{n}{2} \):
   - It’s 5.5, so the median is halfway between the 5\( \text{th} \) and 6\( \text{th} \) values.
   - Mode = 7

Finding the mode and median for data in a frequency table is easy, as long as the data isn’t grouped.

**Example:** The table shows the number of letters received one day in a sample of houses.

<table>
<thead>
<tr>
<th>No. of Letters</th>
<th>No. of Houses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>27</td>
</tr>
<tr>
<td>5</td>
<td>21</td>
</tr>
</tbody>
</table>

a) Find the modal number of letters. The modal number of letters is the one received by the most houses. **Mode = 2 letters**

b) Find the median number of letters.
   1. Add a column for the cumulative frequency — see p.154 — a running total of the frequency.
   2. Find \( \frac{n}{2} \):
      - \( n = 84 \), so \( \frac{n}{2} = 42 \)
      - The median is halfway between the 42\( \text{nd} \) and 43\( \text{rd} \) data values.
   3. Using the cumulative frequency, you can see that the data values in positions 37 to 63 are all 2, so: **Median = 2 letters**

**Exercise 3.2**

Q1: Seventeen friends took part in a charity fun run. Below is the amount, in £, that each friend raised.

<table>
<thead>
<tr>
<th>Amount (£)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>

a) Find the median amount of money raised by these friends.

b) Explain why it is not possible to find the mode for this data.

Q2: A financial adviser records the interest rates charged by 12 different banks to customers taking out a loan. His findings are below.

<table>
<thead>
<tr>
<th>Bank</th>
<th>Interest Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
</tr>
</tbody>
</table>

Find the modal interest rate and the median interest rate charged by these banks.

Q3: An online seller has received the ratings shown in the table. Find the median and the median customer rating.

<table>
<thead>
<tr>
<th>Rating</th>
<th>No. of customers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

Q4: A theatre stages 35 performances of its pantomime one year. The stem and leaf diagram shows the number of empty seats for each performance.

<table>
<thead>
<tr>
<th>Number of empty seats</th>
</tr>
</thead>
<tbody>
<tr>
<td>00-09</td>
</tr>
<tr>
<td>10-19</td>
</tr>
<tr>
<td>20-29</td>
</tr>
<tr>
<td>30-39</td>
</tr>
<tr>
<td>40-49</td>
</tr>
<tr>
<td>50-59</td>
</tr>
<tr>
<td>60-69</td>
</tr>
<tr>
<td>70-79</td>
</tr>
<tr>
<td>80-89</td>
</tr>
</tbody>
</table>

a) Write down the value of the modal number of empty seats.

b) Work out the median number of empty seats.

Key: 3 | 1 = 31 empty seats

Q5: Kwasi and Ben each check the download speeds on their computers on a number of different occasions. Their results are shown in the back-to-back stem and leaf diagram. Find the modal and the median download speeds for:

a) Kwasi

b) Ben

<table>
<thead>
<tr>
<th>Kwasi</th>
<th>Ben</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>

Key: 3 | 5 | 9 means 5.3 Mbit/s for Kwasi and 5.9 Mbit/s for Ben.

**Averages of grouped data**

You can only estimate the mean and median for grouped data, since you no longer have the exact data values. Instead of the mean, you can only find the modal class.

- **Modal class**: The class with the highest frequency density (see p.143). If all the classes are the same width, then this will just be the class with the highest frequency.
- To estimate the mean, assume that every reading takes the value of its class midpoint (see p.141). Then you can use the formula \( \bar{x} = \frac{\sum fx}{\sum f} \), where \( x \) is now the midpoint of each class.
- To estimate the median, find which class the median is in, using the value of \( \frac{n}{2} \), then use linear interpolation — see the next example for the method.
Examples

The table shows the heights of a number of trees in a park.

<table>
<thead>
<tr>
<th>Height of tree to nearest m</th>
<th>0-5</th>
<th>6-10</th>
<th>11-15</th>
<th>16-20</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of trees, ( f )</td>
<td>26</td>
<td>17</td>
<td>11</td>
<td>6</td>
<td>60</td>
</tr>
<tr>
<td>Total, ( \sum f )</td>
<td>75</td>
<td>86</td>
<td>88</td>
<td>60</td>
<td>259</td>
</tr>
</tbody>
</table>

a) Estimate the mean height of the trees.

1. Add the class midpoints (\( x \)) and the values of \( f \) to the table.

\[
\text{Class midpoint, } x = \frac{0.5 + 6 + 11.5 + 16}{4} = 7.125 \text{ m}
\]

2. Use the formula: \( \text{Mean} = \frac{\sum f x}{\sum f} \)

\[
\text{Mean} = \frac{45.75}{60} = 0.7625 \text{ m (3 s.f.)}
\]

b) Estimate the median height of the trees.

1. Draw a table showing the cumulative frequency for each class.

2. \( \frac{N}{2} = 30 \), so there are 30 values that are less than or equal to the median.

3. Use what you know to draw a diagram.

Class boundaries: lower = 5.5, upper = 10.5
Cumulative frequency at 6 = 26, at cb = 43
Cumulative frequency at median = 30

4. Assumming the values are evenly spread, \( \frac{30 - 26}{43 - 26} \approx \frac{m - 5.5}{10.5 - 5.5} \)

\[
m = 5.5 + \frac{43 - 26}{43 - 26} \times (10.5 - 5.5) = 5.5 + 4 = 9.5 \text{ m (3 s.f.)}
\]

Exercise 3.3

Q1 A postman records the number of letters delivered to each of 50 houses one day. The results are shown in the table.

<table>
<thead>
<tr>
<th>No. of letters</th>
<th>0-2</th>
<th>3-5</th>
<th>6-8</th>
<th>9-11</th>
<th>12-14</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of houses</td>
<td>20</td>
<td>16</td>
<td>17</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

a) State the modal class.
b) Estimate the mean number of letters delivered to these houses.

Q2 The time that 60 students took to change after PE is shown below.

<table>
<thead>
<tr>
<th>Time to change, ( t ) (s)</th>
<th>3 s or less</th>
<th>4 s or less</th>
<th>5 s or less</th>
<th>6 s or less</th>
<th>7 s or less</th>
<th>( f )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency, ( f )</td>
<td>7</td>
<td>14</td>
<td>24</td>
<td>10</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

a) Copy and complete the table.
b) Work out estimates of the mean and the median time it took these students to change.

Comparing measures of location

The mean, the median and the mode are all useful for different kinds of data:

Mean
- The mean is a good average as it uses all of the data.
- It can be heavily affected by extreme values / outliers (see p.155) and by skew (see p.146).
- It can only be used with quantitative data (i.e. numbers).

Median
- The median is not affected by extreme values or by data that is skewed.
- Estimating the median for grouped data requires a lot of work.

Mode
- The mode can be used with qualitative (non-numerical) data.
- Data can have more than one mode, and if every value occurs only once then there is no mode.

Exerciso 3.4

Q1 Explain whether the mean, median or mode would be most suitable as a summary of each of the following data sets.
a) Salaries of each employee at a company.
b) Length of adult female adder snakes.
c) Make of cars parked in a car park.
d) Weight of all newborn full-term babies born one year at a hospital.
e) Distance a firm's employees travel to work each morning.

Q2 Host records the number of bedrooms in the houses lived in by a sample of 10 adults. His results are shown in the table.

<table>
<thead>
<tr>
<th>Number of bedrooms</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Explain why the mean may not be the most suitable measure of location for the data.

4. Dispersion

Range, interquartile range and percentile range

Dispersion measures how spread out the data values are. The simplest measure of dispersion is range:

\[
\text{Range} = \text{highest value} - \text{lowest value}
\]

Tip: The midrange (halfway between the highest and lowest values) is another measure of location.

Chapter 11 Sampling, Data Presentation and Interpretation
Another measure of dispersion is the interquartile range. Quartiles divide the data into four quarters. 25% of the data is less than or equal to the lower quartile ($Q_1$), 50% of the data is less than or equal to the median ($Q_2$), and 75% of the data is less than or equal to the upper quartile ($Q_3$).

To find $Q_1$ and $Q_3$, find their position in the ordered list of the data:
- For $Q_1$, work out $\frac{3}{4}$ of and for $Q_3$, work out $\frac{3}{4}$.
- If this value is not a whole number, round it up to find the position of the quartile.
- If it is a whole number, the quartile is halfway between the value in this position and the next value.

You can then find the interquartile range using the formula:

\[
\text{Interquartile range (IQR)} = \text{upper quartile (Q}_3\text{)} - \text{lower quartile (Q}_1\text{)}
\]

**Example**

Find the median and interquartile range of the following data set:
2, 5, 4, 11, 6, 8, 3, 8, 1, 6, 2, 23, 9, 11, 18, 19, 22, 7

1. Put the list in order. 1, 2, 2, 3, 3, 5, 6, 6, 7, 8, 8, 9, 11, 11, 18, 19, 22, 23
2. Find the position of the median. $\frac{7 + 8}{2} = 7.5$, so the median is halfway between the 7th and 8th values. Median = 7
3. Find the positions of $Q_1$ and $Q_3$.
   - $\frac{3}{4} \times 7 = 5.25$, so $Q_1$ is the 5th value = 3
   - $\frac{3}{4} \times 11 = 8.25$, so $Q_3$ is the 12th value = 13
4. Subtract $Q_1$ from $Q_3$.
   - IQR = $Q_3 - Q_1 = 13 - 3 = 10$

To estimate quartiles for grouped data, use linear interpolation.

**Example**

Estimate the interquartile range for the tree heights in this table.

<table>
<thead>
<tr>
<th>Height (m) to nearest m</th>
<th>0-5</th>
<th>6-10</th>
<th>11-15</th>
<th>16-20</th>
<th>No. of trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height (m)</td>
<td>0.5</td>
<td>6-10</td>
<td>11-15</td>
<td>16-20</td>
<td></td>
</tr>
<tr>
<td>Cumulative frequency</td>
<td>26</td>
<td>17</td>
<td>11</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

1. Add the cumulative frequencies to the table.
2. Draw a diagram showing the position of $Q_1$.
   - Cumulative frequency at $Q_1 = \frac{3}{4} = 15$, so $Q_1$ is in the 0.5 class.
   - Class boundaries: lower = 0, upper = 5.5
   - Cumulative frequency at $Q_1 = 0$, at $Q_1 = 0$
3. Solve $\frac{a}{b} = \frac{0}{6}$ to find $Q_1$.
   - $Q_1 = 0.5 \times 0 + 10 = 0.5 \times 11 + 6.5 = 3.1730...$

4. Draw another diagram showing the position of $Q_3$.
   - Cumulative frequency at $Q_3 = \frac{3}{4} = 15$, so $Q_3$ is in the 11.5 class.
   - Class boundaries: lower = 10.5, upper = 15.5
   - Cumulative frequency at $Q_3 = 0$, at $Q_3 = 0$
5. Solve $\frac{a}{b} = \frac{0}{6}$ to find $Q_3$.
   - $Q_3 = 0.5 \times 0 + 15 = 0.5 \times 11.5 + 8.5 = 11.409...$
6. Find the interquartile range: $\text{IQR} = Q_3 - Q_1 = 11.409... - 3.1730... = 8.2364$ (3 s.f.)

**Percentiles** are similar to quartiles, but they divide the data into 100 parts.

The position of the $p$th percentile $(P_p)$ is $\frac{p}{100} \times \text{total frequency (n)}$.

Tip: The median is the 50th percentile, $Q_1$ is the 25th percentile, etc.

Use linear interpolation to estimate percentiles for grouped data, just like for quartiles.

You can find interpercentile ranges by subtracting two percentiles.

The $a%$ to $b%$ interpercentile range is $P_a - P_b$.

**Example**

A reptile specialist records the mass (in kilograms) of 150 tortoises. Her results are shown in the table.

Estimate the 10th percentile for this data.

1. Draw a diagram showing the position of $P_{10}$.
   - Cumulative frequency at $P_{10} = \frac{10}{150} = 15$, so $P_{10}$ is in the 0.2 ≤ $m$ < 0.6 class.
   - Class boundaries: lower = 0.2, upper = 0.6
   - Cumulative frequency at $P_{10} = 0$, at $P_{10} = 0$
2. Solve $\frac{a}{b} = \frac{0}{6}$ to find $P_{10}$.
   - $P_{10} = 0.5 \times 0 + 0.2 = 0.1$, at $P_{10} = 0.1$

**Exercise 4.1**

Find the range and interquartile range of the following data set:
3, 41, 49, 26, 20, 31, 9, 32, 39, 4, 21, 9, 12, 48, 23, 26, 10

The diameters (in miles) of the eight planets in the Solar System are given below:
3032, 7251, 7926, 4222, 88846, 74898, 31763, 30778

For this data set, calculate: a) the range b) the upper and lower quartiles c) the interquartile range.
Q3 The data sets below show the speeds (in mph) of 18 different cars observed at the given time and place.

<table>
<thead>
<tr>
<th>Time</th>
<th>Speeds (mph)</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:45 am</td>
<td>20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38</td>
</tr>
<tr>
<td>10:45 am</td>
<td>19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37</td>
</tr>
</tbody>
</table>

For each set of data, calculate:

a) the range and the interquartile range
b) the 30% to 70% interpercentile range
c) the 10% to 90% interpercentile range
d) the 45% to 55% interpercentile range

Q4 The lengths (l) of a zoot's beetles measured to the nearest mm are shown in the table below. Estimate:

<table>
<thead>
<tr>
<th>Length (l)</th>
<th>Number of beetles</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-5</td>
<td>38</td>
</tr>
<tr>
<td>6-10</td>
<td>28</td>
</tr>
<tr>
<td>11-15</td>
<td>44</td>
</tr>
<tr>
<td>16-20</td>
<td>30</td>
</tr>
<tr>
<td>21-25</td>
<td>16</td>
</tr>
</tbody>
</table>

Cumulative frequency diagrams

A cumulative frequency diagram plots the running total of the frequencies at each upper class boundary. It makes it easy to read off the median and the quartiles.

**Examples**

a) Draw a cumulative frequency diagram for the data below.

<table>
<thead>
<tr>
<th>Age in years</th>
<th>Number of students</th>
</tr>
</thead>
<tbody>
<tr>
<td>11-12</td>
<td>50</td>
</tr>
<tr>
<td>13-14</td>
<td>65</td>
</tr>
<tr>
<td>15-16</td>
<td>56</td>
</tr>
<tr>
<td>17-18</td>
<td>27</td>
</tr>
<tr>
<td>Cumulative frequency</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>115</td>
</tr>
<tr>
<td></td>
<td>173</td>
</tr>
<tr>
<td></td>
<td>200</td>
</tr>
</tbody>
</table>

1. There are 0 students under the age of 11, so the first point is (11, 0). Plot the point (upper class boundary, cumulative frequency) for each class: (11, 0), (13, 50), (15, 115), (17, 173) and (19, 200).
2. Join up the points with straight lines.

b) Estimate the median and interquartile range from the graph.

Exercise 4.2

Q1 The weights of a group of boys were measured and summarised in the cumulative frequency diagram. Estimate how many of the boys weigh:

a) Less than 55 kg
b) More than 73 kg
c) Explain why your answers are estimates.

Q2 Draw a cumulative frequency diagram for the data given below.

<table>
<thead>
<tr>
<th>Distance walked (d)</th>
<th>Number of walkers</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 &lt; d ≤ 2</td>
<td>1</td>
</tr>
<tr>
<td>2 &lt; d ≤ 4</td>
<td>10</td>
</tr>
<tr>
<td>4 &lt; d ≤ 6</td>
<td>7</td>
</tr>
<tr>
<td>6 &lt; d ≤ 8</td>
<td>2</td>
</tr>
</tbody>
</table>

Use your diagram to estimate the median and interquartile range.

Q3 The cumulative frequency diagram shows the monthly earnings of some sixteen-year-olds.

a) How many people were sampled?
b) Estimate the median earnings.
c) Estimate the interquartile range.
d) Estimate how many earned:
   (i) less than £64
   (ii) at least £94
   (iii) between £46 and £84

Outliers and box plots

An outlier is a piece of data that lies a long way from the majority of the readings in a data set. To decide whether a reading is an outlier, you have to test whether it falls outside certain limits, called fences. It's common to use the fences $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$.

A box plot is a 'visual summary' of a data set — it shows the median, quartiles and outliers.

**Example**

The stem and leaf diagram shows the IQs of Year 11 students at a high school.

Draw a box plot to represent this data, using the fences $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$ to identify outliers.

1. Find the quartiles:
   $n = 21$  \[ \frac{21}{4} = 5.25 \]
   $Q_1 = 92$, $Q_2 = 105$, $Q_3 = 118$, $Q_4 = 131$, $Q_5 = 144$

2. Find the fences: $IQR = 124 - 112 = 12$
   Lower fence $= 112 - (1.5 \times 12) = 94$, upper fence $= 124 + (1.5 \times 12) = 146$

3. Check for outliers: 93 < 94 so 93 is an outlier, 144 > 142 so 144 is an outlier.

4. Draw the box plot:
Exercise 4.3

In this exercise use the fences $Q_1 - (1.5 \times IQR)$ and $Q_3 + (1.5 \times IQR)$ to test for outliers.

Q1. The lower and upper quartiles of a data set are 19 and 31. Are the data values 4 and 52 outliers?

Q2. A set of data was analysed and the following values were found.

- minimum value = 4
- maximum value = 49
- $Q_1 = 16$
- median = 24
- $Q_3 = 37$

a) Find the interquartile range.
b) Are there any outliers in this data set?
c) Draw a box plot to illustrate the data set.

Q3. The numbers of items of junk mail received in six months by people living in the towns of Goose & Pigham are shown.

- Goose: 0, 2, 4, 13, 15, 17, 19, 24, 27, 28, 29, 31, 32, 32, 33, 35, 39, 41, 42, 46, 48, 52, 54, 55

Are any of the data values from Pigham outliers?

- a) Draw a box plot to illustrate the data from Goose.
- b) Draw a box plot to illustrate the data from Pigham.
- c) Draw a box plot to illustrate the data from Goose & Pigham.

Exercise 4.4

The attendance figures (a) and the number of TV viewers (b) for a football club's first six matches of the season were:

- attendance figures: 756, 755, 764, 776, 754, 759
- TV viewers: 1, 2, 3, 4

Q1. Find the mean (x̄) of these attendance figures.

Q2. Calculate the sum of the squares of the attendance figures, $\sum x^2$.

Q3. Use your answers to find the variance and standard deviation of the attendance figures.

Q4. The figures for the number of TVs (a) in the households of 20 students are shown in the table.

<table>
<thead>
<tr>
<th>x</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency</td>
<td>7</td>
<td>8</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

The data set is summarised as follows:

- $n = 10$, $\sum x = 29$ and $\sum x^2 = 95.03$

a) Find the standard deviation for the data.

b) The lowest value in the data set is 0.35. Use the fence (x̄ - 2 standard deviations) to determine whether or not this value is an outlier.

Example: Find the variance and standard deviation of the following data set:

1. Find the mean:
   $$\bar{x} = \frac{\sum x}{n} = \frac{2 + 3 + 4 + 4 + 6 + 11 + 12}{7} = 6$$

2. Find $\sum x^2$:
   $$\sum x^2 = 4^2 + 3^2 + 2^2 + 4^2 + 6^2 + 11^2 + 12^2 = 346$$

3. Use the variance formula:
   $$\text{variance} = \frac{\sum x^2}{n} - \bar{x}^2 = \frac{346}{7} - 6^2 = 13.428... = 13.4$$

4. Take the square root:
   $$\text{standard deviation} = \sqrt{13.428...} = 3.66$$

b) Use the fences (x̄ ± 2 standard deviations) to check the data for outliers.

Find the fences:

- $\bar{x} + 2 \times \text{s.d.} = 6 + 2 \times 3.66 = 13.3$ (1 d.p.)
- $\bar{x} - 2 \times \text{s.d.} = 6 - 2 \times 3.66 = -1.3$ (1 d.p.)

All of the data is between these fences, so there are no outliers.

For data in a frequency table, the variance formula can be written:

$$\text{variance} = \frac{\sum f_x^2}{\sum f} - \bar{x}^2$$

where $\bar{x} = \frac{\sum f_x}{\sum f}$

---

Chapter 11: Sampling, Data Presentation and Interpretation
Exercise 4.5

Q1 A student asks a sample of 21 people in his year group how many contacts they have on their mobile phone. The 21 responses can be summarised by \(\sum x = 946\) and \(\sum x^2 = 50,290\), where \(x\) is the number of contacts. Calculate the mean and sample standard deviation of \(x\).

Q2 Calculate the sample variance and sample standard deviation for the data in the table.

Exercise 4.6

Q1 The yields \((w, \text{in kg})\) of potatoes from a number of allotments is shown in the grouped frequency table on the right.

a) Estimate the variance and standard deviation for this data.

b) Explain why your answers to a) are estimates.

Q2 Estimate the standard deviation for the data in the following table:

<table>
<thead>
<tr>
<th>(p)</th>
<th>0.5 &lt; (p &lt; 0.7)</th>
<th>0.7 &lt; (p &lt; 0.8)</th>
<th>0.8 &lt; (p &lt; 0.9)</th>
<th>0.9 &lt; (p &lt; 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>67</td>
<td>132</td>
<td>97</td>
<td>22</td>
</tr>
</tbody>
</table>

Comparing measures of dispersion

The range, interquartile range, variance and standard deviation have advantages and disadvantages:

**Range**
- The range is the **easiest** measure of dispersion to calculate.
- It's heavily affected by even a **single** extreme value / outlier, and it depends on only **two** data values — it doesn't tell you anything about how spread out the rest of the values are.

**Interquartile range**
- It's not affected by **extreme values** — so if the data contains outliers, then the interquartile range is a good measure of dispersion to use.
- It can be fairly **difficult** to calculate, particularly for grouped data.

**Variance**
- The variance depends on **all** the data values, but it's **difficult** to calculate, and is affected by outliers.
- It's also expressed in **different units** from the actual data values, so it can be difficult to interpret.

**Standard deviation**
- Like the variance, the standard deviation depends on **all** the data values, but is **difficult** to calculate, and is affected by extreme values.
- It has the same units as the data values so it is easier to interpret.

Exercise 4.7

Q1 Catherine is examining data on the number of court cases that have occurred each year for the past 20 years. Give one advantage of using the interquartile range as a measure of dispersion.

Q2 A TV channel wants to evaluate data on the amount of time that people spend watching TV per week. They anticipate that there will be a number of extreme values in the data. Which would be a better measure of dispersion: the standard deviation or the interquartile range?

**Coding**

Coding is a way of making the numbers in a data set **easier** to work with, by doing one or both of:
- **Adding** a number to (or **subtracting** a number from) all your readings,
- **Multiplying** (or **dividing**) all your readings by a number.

You have to change your original variable, \(x\), to a different one, such as \(y\). An original data value \(x\) will be related to a **coded** data value \(y\) by an equation of the form \(y = \frac{a\times x}{b}\) where \(a\) and \(b\) are numbers you choose.
The mean and standard deviation of the original data values will then be related to the mean and standard deviation of the coded data values by the following equations:

\[ y = \frac{x-\bar{x}}{b} \]

where \( x \) and \( y \) are the means of variables \( x \) and \( y \), and \( b \) is the standard deviation of \( x \).

**Example:**

Find the mean and standard deviation of:

1. All the values start with 1862..., so subtract 1 862 000:
   \[ x - 1 862 000 = (20, 40, 10, 50) \]
2. All of these results can be divided by 10:
   \[ x/10 = (2, 4, 1, 5) \]
3. Define your coded variable:
   \[ y = \frac{x - 1 862 000}{10} \]
4. Find the mean and s.d. of the coded data:
   \[ \bar{y} = \frac{2 + 4 + 1 + 5}{4} = \frac{12}{4} = 3 \]
   \[ \text{s.d. of } y = \sqrt{\frac{(2 - 3)^2 + (4 - 3)^2 + (1 - 3)^2 + (5 - 3)^2}{4}} \]
   \[ = \sqrt{\frac{1 + 1 + 4 + 4}{4}} = \sqrt{\frac{10}{4}} = \frac{\sqrt{10}}{2} \]
5. Use the formula and rearrange for \( x \):
   \[ y = \frac{x - 1 862 000}{10} \Rightarrow x = \frac{10y + 1862000}{10} \]
   \[ \text{s.d. of } y = \frac{\text{s.d. of } x}{b} = \frac{\text{s.d. of } x - \text{b.s.d. of } y}{10} \]
   \[ = 10 \times 1.581 = 15.8 \text{ (3 s.f.)} \]

With grouped data, you assume that all of the data values are equal to their class midpoint, so this is the x-value that you use with the coding equation \( y = \frac{x - \bar{x}}{b} \).

**Example**

Estimate the mean and standard deviation of this data using \( y = \frac{x-\bar{x}}{10} \).

<table>
<thead>
<tr>
<th>Number of beans</th>
<th>11-20</th>
<th>21-30</th>
<th>31-40</th>
<th>41-50</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency, ( f )</td>
<td>17</td>
<td>21</td>
<td>27</td>
<td>27</td>
<td>80</td>
</tr>
</tbody>
</table>

1. Add to the table.
2. Find the mean and s.d. of \( y \):
   \[ \bar{y} = \frac{17 	imes 20 + 21 	imes 30 + 27 	imes 40 + 27 	imes 50}{80} = 30 = 1.5 \]
   \[ \text{s.d. of } y = \sqrt{\frac{(17 - 30)^2 + (21 - 30)^2 + (27 - 30)^2 + (27 - 30)^2}{80}} \]
   \[ = \sqrt{\frac{9 + 9 + 9 + 9}{80}} = \sqrt{\frac{36}{80}} = 0.50 \]
3. Uncode the data:
   \[ \bar{y} = \frac{10y + 1862000}{10} \Rightarrow x = 10y + 15.5 = 30.5 \text{ beans} \]
   \[ \text{s.d. of } y = \frac{\text{s.d. of } x}{b} = \frac{\text{s.d. of } x}{10} \]
   \[ = 10 \times 0.50 = 5 \text{ (2 s.f.)} \]

**Exercise 4.8**

Q1 A set of data values \( x \) are coded using \( y = \frac{x - 20000}{15} \).

The mean of the coded data \( (\bar{y}) \) is 12.4, and the standard deviation of the coded data is 1.34.

Find the mean and standard deviation of the original data set.

Q2 The widths (in cm) of 10 sunflower seeds in a packet are given below:

0.61, 0.67, 0.63, 0.66, 0.65, 0.64, 0.68, 0.64, 0.62

a) Code the data \( (x) \) to form a new set consisting of integer values \( (y) \) between 1 and 10.

b) Find the mean and standard deviation of the original values \( (x) \).

Q3 The table below shows the weight \( x \) of 12 items on a production line.

<table>
<thead>
<tr>
<th>Weight (to nearest g)</th>
<th>100-104</th>
<th>105-109</th>
<th>110-114</th>
<th>115-119</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

Use the coding \( y = x - 102 \) to estimate the mean and standard deviation of the items' weights.

Q4 Twenty pieces of data \( (x) \) have been summarised as follows:

\[ \sum x = 7 \quad \text{and} \quad \sum x + 2 = 80 \]

Calculate the mean and standard deviation of the data.

**Comparing distributions**

You can compare two distributions by comparing measures of location (mean, median, midrange, mode) or measures of dispersion (range, interquartile range, standard deviation, etc.). It's important to relate these back to the context of the data in order to draw conclusions about the distributions.

**Example**

The table summarises the marks obtained by a group of students in Maths 'Calculator' and 'Non-calculator' papers.

<table>
<thead>
<tr>
<th>Calculator paper</th>
<th>Median, ( Q_2 )</th>
<th>Non-calculator paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>58</td>
<td>42</td>
<td>55</td>
</tr>
<tr>
<td>39</td>
<td>Interquartile range 21</td>
<td>Mean 46</td>
</tr>
<tr>
<td>21</td>
<td></td>
<td>21.2 Standard deviation 17.8</td>
</tr>
</tbody>
</table>

| Location | Midrange of calculator papers is higher, so scores are generally higher on the calculator paper. |
| Dispersion | The IQR and standard deviation are higher for the calculator paper, so scores on the calculator paper are more spread out than those for the non-calculator paper. |

**Example**

The box plots below show how the masses (in g) of the tomatoes in two harvests were distributed. Compare the distributions of the two harvests.

| Location | Midrange of the 'Harvest 2' is higher, so the tomatoes in Harvest 2 were generally heavier. |
| Dispersion | The interquartile range (IQR) on the range for Harvest 1 are higher than those for Harvest 2, so the masses of the tomatoes in Harvest 1 were more varied than the masses of the tomatoes in Harvest 2. |
Exercise 4.9

Q1 The box plots show the prices of shoes (in £) from two different shops. Compare the location and dispersion of the two shops' prices.

Q2 10 men and 10 women were asked how many hours of sleep they got on a typical night.

The results are as follows: Men: 6, 7, 9, 8, 8, 6, 7, 10, 5
Women: 9, 5, 7, 8, 5, 11, 10, 8, 10, 8

Compare the location and dispersion of the two data sets.

5. Correlation and Regression

Scatter diagrams and correlation
Data made up of pairs of values (x, y) is called bivariate data. You can plot bivariate data on a scatter diagram — where each variable is plotted along one of the axes. Scatter diagrams are helpful for recognising when data is correlated — the closer the data points are to forming a straight line, the stronger the correlation is. You may have to identify outliers — those could be values that don’t fit the pattern, or measurement errors.

Examples
Describe the correlation shown in each scatter diagram:

a) Most of the data points lie very close to a straight line with positive gradient, so the diagram shows strong positive correlation.

b) The data does not seem to form any line, so there is no correlation between the variables.

c) Overall, the data seems to trend towards a line with positive gradient, so the diagram shows positive correlation overall. However, the data forms two clusters, which both show only weak positive correlation.

Although two variables might be correlated, this does not mean that a change in one causes a change in the other. They could be linked by another factor — for example, sales of barbecues and sales of ice cream might be correlated, but they’re both affected by another factor (temperature).

Exercise 5.1

Q1 8 runners participating in a 1 km race were asked how many hours of exercise they did per week. Their responses, and their race times in minutes, are shown in the table below:

<table>
<thead>
<tr>
<th>Exercise, hours</th>
<th>15</th>
<th>10</th>
<th>5</th>
<th>2</th>
<th>7</th>
<th>5</th>
<th>2</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race time, min</td>
<td>6.7</td>
<td>6.8</td>
<td>9.5</td>
<td>6.1</td>
<td>11.5</td>
<td>6.3</td>
<td>8</td>
<td>9.9</td>
</tr>
</tbody>
</table>

a) Plot a scatter diagram to show this data.

b) Describe any correlation shown, and identify any results that appear to be outliers.

Q2 Zoe lives near a beach. She records the number of seagulls and the number of people flying kites she sees each day, over the course of one month. She finds negative correlation in the data, and concludes that seagulls must be scared off by kites. Explain whether the data supports this conclusion.

Q3 This table shows the average length and the average circumference of eggs for several species of bird, measured in cm.

<table>
<thead>
<tr>
<th>Species</th>
<th>Length</th>
<th>Circumference</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5.9</td>
<td>19.6</td>
</tr>
<tr>
<td>B</td>
<td>2.1</td>
<td>16.7</td>
</tr>
<tr>
<td>C</td>
<td>3.4</td>
<td>21.3</td>
</tr>
<tr>
<td>D</td>
<td>5.1</td>
<td>18.1</td>
</tr>
<tr>
<td>E</td>
<td>8.9</td>
<td>20.2</td>
</tr>
<tr>
<td>F</td>
<td>6.6</td>
<td>19.8</td>
</tr>
<tr>
<td>G</td>
<td>7.2</td>
<td>19.5</td>
</tr>
<tr>
<td>H</td>
<td>4.5</td>
<td>18.7</td>
</tr>
<tr>
<td>I</td>
<td>6.8</td>
<td>20.3</td>
</tr>
</tbody>
</table>

a) Plot a scatter diagram to show this data.

b) Describe any trends in the data.

c) One of the measurements was recorded incorrectly. Use your scatter diagram to determine which.

Linear regression

If two variables are correlated, you can draw a line of best fit through the data on the scatter diagram. Linear regression is a process that is used to find the equation of this line, called the regression line.

In order to interpret a regression line, you need to decide which is the explanatory variable and which is the response variable.

- The explanatory variable (or independent variable) is the variable you can directly control, or the one that you think is affecting the other — it is always drawn along the horizontal axis.
- The response variable (or dependent variable) is the variable you think is being affected — it is always drawn along the vertical axis.

Examples
For each situation below, explain which quantity would be the explanatory variable, and which would be the response variable.

a) A scientist is investigating the relationship between the amount of fertiliser applied to a tomato plant and the eventual yield.

1. The scientist can directly control the amount of fertiliser they give each plant, so 'amount of fertiliser' is the explanatory variable.
2. The scientist measures the effect this has on the plant's yield, so 'yield' is the response variable.

b) A researcher is examining how a town's latitude and the number of days when the temperature rose above 10 °C are linked.

Although the researcher can't control the latitude of towns, it would be the difference in latitude that leads to a difference in temperature, and not the other way around. So 'town's latitude' is the explanatory variable, and 'number of days when the temperature rose above 10 °C' is the response variable.
The regression line of $y$ on $x$ is a line of the form: $y = a + bx$ where $a$ and $b$ are constants.

The 'of $y$ on $x$' part means that $x$ is the explanatory variable, and $y$ is the response variable.

**Example**

A company is collecting data on the fuel efficiency of a type of lorry. They compare the load on a lorry, $x$ (in tonnes), with the fuel efficiency, $y$ (in km per litre), and calculate the regression line of $y$ on $x$: $y = 12.5 - 0.6x$. Interpret the values of $a$ and $b$ in this context.

1. The value of $a$ tells you that a load of 0 tonnes corresponds to a fuel efficiency of 12.5 km per litre — this is the fixed fuel efficiency of the lorry before you have even loaded anything on it.
2. The value of $b$ tells you that for every extra tonne carried, you'd expect the lorry's fuel efficiency to fall by 0.6 km per litre (since when $x$ increases by 1, $y$ falls by 0.6).

You can use a regression line to predict values of your response variable. There are two forms of this — interpolation and extrapolation.

- **Interpolation** is when you use the regression line to predict values of the response variable for values of the explanatory variable that are within the range of your collected data. This is usually reliable since the observed data supports the prediction.

- **Extrapolation** is when you predict values of the response variable for values of the explanatory variable that are outside the range of the collected data. This is a lot more unreliable, as the data only provides evidence that the regression line is accurate within the observed range. Outside of this range, the relationship could change in a way that the data does not show.

**Examples**

The length of a spring $(x, \text{ in cm})$ when loaded with different masses $(m, \text{ in g})$ has the regression line of $y$ on $x$: $y = 7.8 + 0.01043x$.

a) Estimate the length of the spring when loaded with a mass of (i) 370 g (ii) 670 g.

(i) $m = 370$, so: $y = 7.8 + 0.01043 \times 370 = 11.7$ cm (1 d.p.)

(ii) $m = 670$, so: $y = 7.8 + 0.01043 \times 670 = 14.8$ cm (1 d.p.)

b) The smallest value of $m$ used to find the regression line was 200 g, and the largest value of $m$ was 500 g. Comment on the reliability of the estimates in part a).

1. $m = 370$ falls within the range of the original data for $m$, so this is an interpolation. This means the result should be fairly reliable.

2. But $m = 670$ falls outside the range of the original data for $m$, so this is an extrapolation. This means the regression line may not be valid, and the estimate of $y$ should be treated with caution.

**Exercise 5.2**

Q1 For each pair of variables, state which is the explanatory variable, and which is the response variable:

a) 'time spent practising the piano in a week' and 'number of mistakes made in a test at the end of the week'

b) 'age of a second-hand car' and 'value of the car'

c) 'number of phone calls made in a town in a week' and 'population of the town'

d) 'growth rate of a plant in an experiment' and 'amount of sunlight falling on the plant'

Q2 The equation of the regression line of $y$ on $x$ is $y = 1.67 + 0.107x$.

a) Which variable is the response variable?

b) Find the predicted value of $y$ corresponding to: (i) $x = 5$ (ii) $x = 20$

Q3 A volunteer counted the number of spots (s) on an area of skin after $d$ days of acne treatment, where $d$ had values 2, 6, 10, 14, 18 and 22. The equation of the regression line of $s$ on $d$ is $s = 58.8 - 2.47d$.

a) Estimate the number of spots the volunteer had on day 7.

b) Comment on the reliability of your answer.

c) She forgot to count how many spots she had before starting to use the product. Estimate this number. Comment on your answer.

c) The volunteer claims that the regression equation must be wrong, because it predicts that after 30 days she should have a negative number of spots. Comment on this claim.

**Non-linear regression**

When the relationship between two variables is non-linear (i.e. they don't form a straight line), the 'best fit' of the data can be described by a non-linear function, such as a quadratic, trigonometric, logarithmic or exponential function.

**Examples**

A concert venue collects data on the number of tickets sold for an event, $y$, and the amount they spend on advertising for the event, $x$. They calculate that the regression curve of $y$ on $x$ is: $y = 0.002x^2 + 55$

a) The venue has budgeted to spend £500 on advertising for a particular event.

How many tickets do they expect to sell?

Set $x = 500$: $y = 0.002 \times 500^2 + 55 = 555$ tickets

b) The venue has 1500 seats. One of the staff members claims that, if they want to sell out the event, they should spend £850 on advertising. Comment on the validity of this claim.

1. This claim is trying to predict the amount they should spend based on the number of tickets they want to sell — i.e. predict the explanatory variable from the response variable. The regression curve should only be used to predict the value of the response variable, so this claim is not reliable.

2. If 1500 is the maximum number of tickets the venue can sell, then they may not have a lot of data to support the claim and it could be extrapolated from their collected data — this is another reason why it might be unreliable.

**Exercise 5.3**

Q1 The scatter diagram shows a bivariate data set $(x, y)$.

Which of the following types of function would be the most appropriate as a best fit model for this data?

- linear
- quadratic
- cubic
- sine
- exponential

A bivariate data set $(x, y)$ is summarised as follows:

- number of pairs of data values: 5
- range of $x$: 1.6 to 7.9
- regression curve of $r$ on $x$: $r = 2e^{0.2x}$

Predict the value of $r$ when $x = 2$, and comment on the validity of this estimate.

Chapter 11 Sampling, Data Presentation and Interpretation
**Review Exercise — Chapter 11**

**Q1** The manager of a tennis club wants to know if members are happy with the facilities provided.
   a) Identify the population the manager is interested in.
   b) Is this population finite or infinite?

**Q2** A teacher is investigating whether a student's ability to run a set of letters is related to their ability to spell. He plans to test students from his school, which has 1200 pupils, to do a standard spelling test and then to memorise a set of 20 letters.
   a) What is the population?
   b) Give two reasons why he should use a sample survey rather than carry out a census.

**Q3** The houses on Park Road are numbered from 1 to 173. Forty households are to be chosen to take part in a council survey. Describe a method for choosing an unbiased sample.

**Q4** A school uniform manufacturer wants to collect data on the heights of teenagers at the schools in their county. Describe a suitable sampling method for this situation.

**Q5** For each of the following situations, name the sampling method used and give one disadvantage of using this sampling method:
   a) A tea company is investigating tea-drinking habits of its customers. The interviewer is asked to sample exactly 60 women and 40 men using a non-random sampling method.
   b) After a concert, a band is looking for feedback from their fans. Using the ticket numbers, they select every 100th fan to complete a survey.
   c) A student is researching shopping habits in the UK. He records how many people enter his local shopping centre between 9 am and 5 pm on a Monday.

**Q6** Twenty phone calls were made by a householder one evening. The lengths of the calls (in minutes to the nearest minute) are recorded in the table below:

<table>
<thead>
<tr>
<th>Length of calls</th>
<th>0-2</th>
<th>3-5</th>
<th>6-8</th>
<th>9-15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of calls</td>
<td>10</td>
<td>6</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

Show this data as: a) a frequency polygon, b) a histogram.

**Q7** The histogram represents the number of people working at a company, grouped by age.
   a) Find the percentage of people aged between 26 and 29.
   b) Find the probability that a randomly-chosen person is aged 36 or over.
   c) Estimate the percentage of people aged 35 or over.
   d) Explain why your answer in part (c) is only an estimate.

**Q8** The stem and leaf diagram on the right represents the lengths (in cm) of 15 bananas. Write down the original data as a list.

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>13</td>
<td>2 5</td>
</tr>
<tr>
<td>14</td>
<td>3 6 8</td>
</tr>
<tr>
<td>15</td>
<td>2 9</td>
</tr>
<tr>
<td>16</td>
<td>1 2 3</td>
</tr>
<tr>
<td>17</td>
<td>0 2</td>
</tr>
</tbody>
</table>

Key: 12|8 means 12.8 cm

---

**Q9** Describe the distribution of the data shown by each of the three curves on the diagram, stating any skewness and whether the data appears to be unimodal or bimodal.

**Q10** Calculate the mean, median and mode of the data in the table.

<table>
<thead>
<tr>
<th>x</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

**Q11** The speeds of 60 cars travelling in a 40 mph speed limit area are measured to the nearest mph. The data is summarised in the table. Estimate the mean and median, and state the modal class.

<table>
<thead>
<tr>
<th>Speed (mph)</th>
<th>30-34</th>
<th>35-39</th>
<th>40-44</th>
<th>45-50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>12</td>
<td>37</td>
<td>9</td>
<td>2</td>
</tr>
</tbody>
</table>

**Q12** Two data sets, A and B, are given below:

A: 16, 41, 28, 23, 7, 11, 37, 16, 9, 21, 26, 18, 14, 31, 8
B: 33, 36, 25, 15, 42, 13, 6, 24, 30, 15, 19, 15, 40, 36, 24

Calculate:
   a) the range of B
   b) the interquartile range of A
   c) the 20% to 80% interpercentile range of B

**Q13** a) Draw a cumulative frequency diagram for the following data:

<table>
<thead>
<tr>
<th>r</th>
<th>0 ≤ r ≤ 2</th>
<th>2 ≤ r ≤ 4</th>
<th>4 ≤ r ≤ 6</th>
<th>6 ≤ r ≤ 8</th>
<th>8 ≤ r ≤ 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>2</td>
<td>6</td>
<td>7</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

b) Use your cumulative frequency diagram to estimate the number of values that are:
   (i) less than 3
   (ii) more than 5
   (iii) between 3.5 and 7

c) Use your cumulative frequency diagram to estimate:
   (i) the interquartile range
   (ii) the 10% to 90% interpercentile range

**Q14** A company is testing a new video game. In the game, each level lasts for 200 seconds, after which the next level starts immediately. 100 people played the game, and the level they reached before losing is shown in the table.

<table>
<thead>
<tr>
<th>Level reached</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency, f</td>
<td>6</td>
<td>11</td>
<td>24</td>
<td>22</td>
<td>10</td>
<td>7</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

a) Draw a grouped frequency table with 6 classes showing how long the games lasted, in seconds.
   b) Draw a cumulative frequency diagram for the data in your new table.
   c) Use your cumulative frequency diagram to estimate the number of games that lasted:
      (i) less than 500 seconds,
      (ii) more than 1500 seconds.
   d) Use linear interpolation to estimate the interquartile range of the games’ durations.

**Q15** Two workers iron 10 items of clothing and record the time, to the nearest minute, that each takes:

Worker A: 3 5 2 7 10 4 5 5 4 12
Worker B: 4 4 8 6 7 8 9 10 11 9

a) For worker A, find:
   (i) the median
   (ii) the lower and upper quartiles
   (iii) whether there are any outliers, using the fences Q1 - 1.5 × IQR and Q3 + 1.5 × IQR.

b) Draw two box plots, using the same scale, to represent the times of each worker.

c) Make one statement comparing the two sets of data.

d) Which worker would be better to employ? Give a reason for your answer.