

Cambridge OL

Mathematics

CODE: (4024)

Chapter 44 and chapter 45

*PROBABILITY AND CATEGORICAL,
NUMERICAL AND GROUPED DATA*



The probability of a single event

When you toss a coin, there are two possible outcomes, heads or tails. If the coin is fair, these outcomes are equally likely.

The probability of getting a head is $\frac{1}{2}$ or 0.5.

The probability of getting a tail is $\frac{1}{2}$ or 0.5.

The sum of these two probabilities is 1, because there are no other possible outcomes.

If there is a set of equally likely outcomes then the probability of each **event** is given by this formula.

$$\text{Probability of an event} = \frac{\text{number of outcomes which give this event}}{\text{total number of outcomes}}$$

If it is impossible that the event will happen, the probability is 0.

If it is certain that the event will happen, the probability is 1.

Probabilities can be written as either fractions or decimals.

The probability of an event not occurring

There are ten counters in a bag.

Seven of the counters are black.

A counter is taken at random from the bag.

The probability that the counter is black is $\frac{7}{10}$.

You can write this probability as $P(\text{black}) = \frac{7}{10}$.

To find the probability that the counter is not black, you need to think about how many outcomes there are where the counter is not black.

As there are ten counters in the bag, there are three counters that are not black.

The probability that the counter is not black is $\frac{3}{10}$.

You can write this as $P(\text{not black}) = \frac{3}{10}$

Note that $P(\text{black}) + P(\text{not black}) = \frac{7}{10} + \frac{3}{10} = 1$.

This is because there are no other possible outcomes for this experiment.

In probability notation, if A is an event, then you can write A' for the event not occurring. A' is the event 'not A '.

The probability of an event not occurring = 1 – probability of the event occurring.

So $P(A') = 1 - P(A)$.

This rule can also be used when there are more than two possible events and you know the probability of all but one of them occurring.

Example 44.1

Question

A bag contains six red counters, three blue counters and one green counter.

A counter is taken at random from the bag.

Find the probability that the counter is

- a** red **b** green **c** black.

Solution

a There are $6 + 3 + 1 = 10$ counters in the bag, so there are ten equally likely outcomes.

Six of the counters are red, so six outcomes give the required event.

The probability that the counter is red is $\frac{6}{10}$ which can be simplified to $\frac{3}{5}$.

b One of the counters is green, so only one outcome gives the required event.

The probability that the counter is green is $\frac{1}{10}$. This fraction is in its simplest form.

c There are no black counters in the bag, so this event is impossible.

The probability that the counter is black is 0.

Note

When you write a probability as a fraction, write your final answer as a fraction in its simplest form.

Example 44.2

Question

- a** The probability that Nayim's bus is late is $\frac{1}{8}$.
What is the probability that Nayim's bus is not late?
- b** Salma can travel to school by bus, car or bike.
The probability that she travels by bus is 0.5.
The probability that she travels by car is 0.2.
What is the probability that she travels by bike?

Solution

$$\begin{aligned}\mathbf{a} \quad P(\text{not late}) &= 1 - P(\text{late}) \\ &= 1 - \frac{1}{8} \\ &= \frac{7}{8}\end{aligned}$$

$$\begin{aligned}\mathbf{b} \quad P(\text{bike}) &= 1 - P(\text{bus}) - P(\text{car}) \\ &= 1 - 0.5 - 0.2 \\ &= 0.3\end{aligned}$$

The only possibilities are bus, car and bike, so these three probabilities sum to 1.

Estimating from a population

If you know the probability of an event happening, then you can use this probability to make an estimate of the number of times this event will occur.

Example 44.3

Question

Saira goes to work by bus each morning.
The probability that the bus is late is 0.15.
Estimate the number of times out of 40 mornings that the bus is expected to be late.

Solution

Number of times late = $0.15 \times 40 = 6$
The bus is expected to be late on approximately six out of the 40 mornings.

Note

This does not mean that the bus will definitely be late six times.

It is an estimate based on the probability that has been given.

Relative frequency and probability

It is not always possible to find probabilities by looking at equally likely outcomes.

For example, supposing you have to work out the probability of scoring a 6 with a spinner that is numbered 1 to 6 that may be biased, the probability of a driver having a car accident or the probability that a person will visit a certain shop.

For this type of event, you need to set up an experiment, carry out a survey or use past results. Take the example of scoring a 6 with a spinner numbered 1 to 6 that may be biased.

For a fair (unbiased) spinner, the probability of getting a 6 is $\frac{1}{6} = 0.166\ldots$
 $= 0.17$ approximately.

You can record the number of 6s you get when you spin the spinner a number of times, and use the results to decide whether the spinner may be biased.

The proportion of times a 6 occurs is known as the relative frequency

$$\text{Relative frequency} = \frac{\text{number of times an outcome occurs}}{\text{total number of trials}}$$

The table shows the results of an experiment.

Number of trials	Number of 6s	Relative frequency
10	4	$\frac{4}{10} = 0.4$
50	18	$\frac{18}{50} = 0.36$
100	35	$\frac{35}{100} = 0.35$
500	180	$\frac{180}{500} = 0.36$

The relative frequency gives an estimate of the probability of scoring a 6. The relative frequency changes depending on the number of trials. The greater the number of trials, the better the estimate of the probability will be. The results of this experiment suggest that the spinner is biased.

An estimate for the probability of scoring a 6 with this spinner is 0.36.

Example 44.4

Question

Farid carries out a survey on the colours of the cars passing his school. The table shows his results.

Colour	Black	Red	Blue	White	Green	Other	Total
Number of cars	51	85	64	55	71	90	416

Use his results to estimate the probability that the next car that passes will be

- a** red **b** not red.

Solution

Relative frequency can be used as an estimate of probability.

a Relative frequency of red cars = $\frac{\text{number of red cars}}{\text{total number of cars}} = \frac{85}{416}$

Estimate of probability = $\frac{85}{416} = 0.204$ to 3 decimal places.

b $P(\text{not red}) = 1 - P(\text{red})$

Estimate of $P(\text{not red}) = 1 - \frac{85}{416} = \frac{331}{416} = 0.796$ to 3 decimal places.

Note

The relative frequency can be written as either a fraction or a decimal.

The probability of combined events

To find the probability of more than one event happening, you need to make sure that you consider all of the possible outcomes.

One way of finding all of the outcomes is to use a sample space diagram. If two fair coins are tossed, the first coin can show either a head or a tail and the second coin can show either a head or a tail.

The sample space diagram below shows all of the possible outcomes.

		1st coin	
		H	T
2nd coin	H	X	X
	T	X	X

This shows that there are four possible outcomes: HH, HT, TH, TT.

These are all equally likely.

The probability of each event can be found from the sample space diagram.

$$P(2 \text{ heads}) = \frac{1}{4}$$

$$P(1 \text{ head, 1 tail}) = \frac{2}{4} = \frac{1}{2}$$

$$P(2 \text{ tails}) = \frac{1}{4}$$

When the events are numerical – such as when two fair spinners numbered 1 to 6 are spun – it can be helpful to show the numerical outcomes on the sample space diagram rather than using a cross.

The sample space diagram here shows the outcomes for spinning two fair spinners numbered 1 to 6 when the two scores are added together.

This shows that there are 36 possible outcomes.

Each outcome is equally likely.

The diagram can be used to find the probability of different events.

There are five different outcomes that give a score of 8, so $P(8) = \frac{5}{36}$.

There are two different outcomes that give a score of 11, so $P(11) = \frac{2}{36}$.

Suppose that you want to find the probability of scoring 8 **or** 11.

From the sample space diagram, you can see that there are a total of seven outcomes that give a score of 8 or 11, so $P(8 \text{ or } 11) = \frac{7}{36}$.

You can also see that $P(8) + P(11) = \frac{5}{36} + \frac{2}{36} = \frac{7}{36}$.

		1st spinner					
		1	2	3	4	5	6
2nd spinner	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Score of 8 Score of 11

$$\text{So } P(8 \text{ or } 11) = P(8) + P(11).$$

When two events cannot happen at the same time they are known as **mutually exclusive events**.

The events 'scoring 8' and 'scoring 11' are mutually exclusive because they cannot both happen at the same time.

If events A and B are mutually exclusive, then $P(A \text{ or } B) = P(A) + P(B)$.

This rule does not apply if the events can happen at the same time, for example 'scoring 8' and 'scoring a 4 on the first spinner'.

$$P(8) = \frac{5}{36}$$

$$P(4 \text{ on first spinner}) = \frac{6}{36}$$

From the sample space diagram, you can see that there are 10 outcomes that give a score of 8 **or** score a 4 on the first spinner, so

$$P(8 \text{ or } 4 \text{ on first spinner}) = \frac{10}{36}.$$

$$P(8 \text{ or } 4 \text{ on first spinner}) \neq P(8) + P(4 \text{ on first spinner}).$$

This is because scoring a double 4 fits both events.

These two events are not mutually exclusive, so the probabilities cannot be added.

		1st spinner					
		1	2	3	4	5	6
2nd spinner	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Score of 8 4 on 1st spinner

Example 44.5

Question

Ravi spins two fair spinners numbered 1 to 6 and adds the scores.

Find the probability that he scores

- a 12
- b less than 6
- c 10 or 11.

Solution

The outcomes can be seen in the sample space diagram shown above.

- a There is only one way to score 12, so $P(12) = \frac{1}{36}$.
- b It can be seen from the diagram that there are 10 outcomes with a score of less than 6.
 $P(\text{less than } 6) = \frac{10}{36} = \frac{5}{18}$ Write the fraction in its simplest form.
- c It can be seen from the diagram that there are five outcomes with a score of 10 or 11.

$$P(10 \text{ or } 11) = \frac{5}{36}$$

$$\text{Note that } P(10) = \frac{3}{36} \text{ and } P(11) = \frac{2}{36}.$$

Scoring 10 and 11 are mutually exclusive,
so $P(10 \text{ or } 11) = P(10) + P(11)$

$$= \frac{3}{36} + \frac{2}{36} = \frac{5}{36}.$$

Independent events

When you toss a coin twice, the outcome of the second toss is not affected by the outcome of the first toss. Two events are said to be independent events when the outcome of the second event is not affected by the outcome of the first.

If events A and B are independent, $P(A \text{ and } B) = P(A) \times P(B)$.

The probability of getting a head when you toss a fair coin is $\frac{1}{2}$.

You have seen from the sample space diagram that the probability of getting two heads when you toss a coin twice is $\frac{1}{4}$.

$$P(\text{head and head}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

The result for independent events can be extended to more than two events. The probability of all of the events happening is found by multiplying together the probabilities of each individual event

Example 44.6

Question

- a** The probability that Imran's bus to school is late is 0.2.
Assuming that the events are independent, find the probability that the bus is late
- two mornings in a row
 - three mornings in a row.
- b** There are four green balls and one blue ball in a bag.
Anna selects a ball, notes its colour and replaces it.
She then selects another ball.
What is the probability that the first ball is green and the second ball is blue?

Solution

- a** The events are independent, so the probabilities are multiplied.
- $P(\text{late two mornings}) = P(\text{late}) \times P(\text{late}) = 0.2 \times 0.2 = 0.04$
 - $P(\text{late three mornings}) = P(\text{late}) \times P(\text{late}) \times P(\text{late}) = 0.2 \times 0.2 \times 0.2 = 0.008$
- b** The first ball is replaced in the bag before the second is taken, so the events are independent.
- $$P(\text{green}) = \frac{4}{5}$$
- $$P(\text{blue}) = \frac{1}{5}$$
- $$P(\text{first green and second blue}) = P(\text{green}) \times P(\text{blue}) = \frac{4}{5} \times \frac{1}{5} = \frac{4}{25}$$

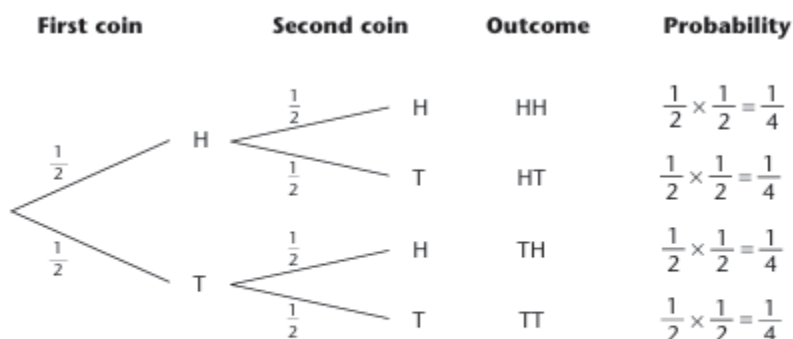
Note

The larger the number of repeats of the event, the smaller the probability that it occurs.

Tree diagrams for combined events

You have seen that sample space diagrams are one way of showing all of the outcomes for two combined events. A tree diagram can be used to show the outcomes, and their probabilities, for two or more events.

The tree diagram for tossing two coins looks like this.



		1st coin	
		H	T
2nd coin	H	x	x
	T	x	x

To find the probability of one outcome, *multiply* the probabilities on the branches.

To find the probability of more than one outcome, *add* the probabilities of each outcome.

Notice that on each pair of branches the probabilities add up to 1. Also the probabilities of all the four possible outcomes add up to 1. If there are more than two combined events, further sets of branches can be added to the tree.

Example 44.7

Question

There are six red balls and four black balls in a bag.

Gita selects a ball, notes its colour and replaces it.

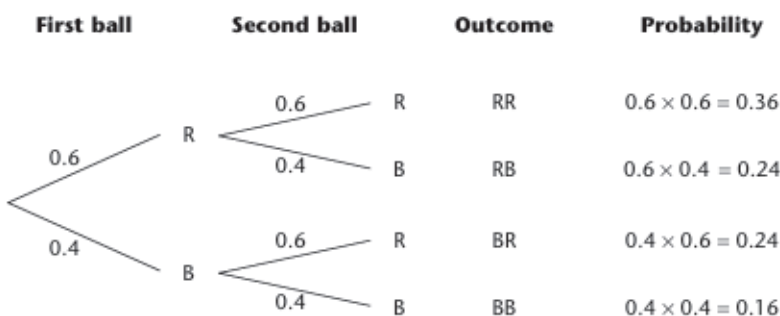
She then selects another ball.

What is the probability that Gita selects

- a two red balls
- b one of each colour?

Solution

A tree diagram can be drawn to show this information.



- a $P(RR) = 0.6 \times 0.6 = 0.36$
- b For one of each colour, the outcome is either RB or BR.

$$P(\text{one of each colour}) = P(RB) + P(BR)$$

$$= (0.6 \times 0.4) + (0.4 \times 0.6)$$

$$= 0.24 + 0.24 = 0.48$$

Note

Some questions may not ask you to draw a tree diagram. However, drawing a tree diagram will help you to make sure that you have considered all of the possible outcomes. You may be able to answer some probability questions without a tree diagram.

The probability of dependent events

In the last section, all of the events were independent. The outcome of the second event was not affected by the outcome of the first.

In some situations, the outcome of the second event is affected by the first event. In this situation, the probability of the second event depends on what happened in the first event. In this case the events are called **dependent events**.

You can solve problems involving dependent events using tree diagrams.

When events are dependent, the probabilities on the second sets of branches are not the same as the probabilities on the first set

Example 44.8

Question

There are seven blue balls and three red balls in a bag.

A ball is selected at random and **not** replaced.

A second ball is then selected.

Find the probability that

- a two blue balls are chosen
- b two balls of the same colour are chosen.

Solution

A tree diagram can be used to find the probabilities of the outcomes.

The probabilities for the second ball depend on the colour of the first ball.

If the first ball was blue there are now six blue balls and three red balls left in the bag.

The probabilities are therefore $\frac{6}{9}$ and $\frac{3}{9}$, respectively.

If the first ball was red, there are now seven blue balls and two red balls left in the bag.

The probabilities are therefore $\frac{7}{9}$ and $\frac{2}{9}$, respectively.

The tree diagram looks like this.

First ball	Second ball	Outcome	Probability
$\frac{7}{10}$ B	$\frac{6}{9}$ B	BB	$\frac{7}{10} \times \frac{6}{9} = \frac{42}{90}$
	$\frac{3}{9}$ R	BR	$\frac{7}{10} \times \frac{3}{9} = \frac{21}{90}$
$\frac{3}{10}$ R	$\frac{7}{9}$ B	RB	$\frac{3}{10} \times \frac{7}{9} = \frac{21}{90}$
	$\frac{2}{9}$ R	RR	$\frac{3}{10} \times \frac{2}{9} = \frac{6}{90}$

Note

You should give your final answer as a fraction in its simplest form, but it is usually better not to simplify the fractions in the tree diagram. This is because you may need to add the probabilities and you will need them with a common denominator.

$$\begin{aligned} \text{a } P(\text{BB}) &= \frac{7}{10} \times \frac{6}{9} \\ &= \frac{42}{90} \\ &= \frac{7}{15} \end{aligned}$$

$$\begin{aligned} \text{b } P(\text{BB or RR}) &= \frac{7}{10} \times \frac{6}{9} + \frac{3}{10} \times \frac{2}{9} \\ &= \frac{42}{90} + \frac{6}{90} \\ &= \frac{48}{90} \\ &= \frac{8}{15} \end{aligned}$$

You use the tree diagram in the same way as you would for independent events.

The multiplication and addition rules still apply.

It is the probabilities that are different from the independent case.

Key points

- The probability of an event = $\frac{\text{number of outcomes that give the event}}{\text{total number of outcomes}}$.
- Probability can have a value from 0 to 1 or from 0% to 100%.
- Probability can be written as a fraction, a decimal or a percentage only.
- Probability of an event not occurring = $1 - \text{probability of event occurring}$ or $P(A') = 1 - P(A)$.
- Expected frequency = probability of an event occurring \times total number of observations of an event.
- Relative frequency is an estimate of probability:
Relative frequency = $\frac{\text{number of times an event occurs}}{\text{total number of trials}}$
- A sample space diagram is a two-way table listing all possible outcomes of two separate events occurring.
- For mutually exclusive events A and B
 $P(A \text{ or } B) = P(A) + P(B)$.
- For independent events A and B
 $P(A \text{ and } B) = P(A) \times P(B)$.
- A probability tree diagram shows the outcomes and probabilities of two or more events.

CHAPTER 45 - CATEGORICAL, NUMERICAL AND GROUPED DATA

Collecting and grouping data You may have seen headlines like these in newspapers or on television.

THE BEATLES ARE THE MOST POPULAR BAND or
TEENAGERS VOTE DIET COKE AS THE BEST SOFT DRINK

You may not agree with either headline, but how is such information found out? One way is to ask people questions – that is, to carry out a survey using a questionnaire and collect the answers on a data collection sheet. The answers are then analysed to give the results.

Example 45.1

Question

Jamila is asked to find out how many people there are in cars that pass the school gates each day between 8.30 a.m. and 9.00 a.m.


She writes down the number of people in the first 40 cars that pass on the first Monday.

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

Make a frequency table to show her results.





Solution

You make a tally chart in the usual way.

Remember to group the tally marks into fives, which are shown like this, , for easier counting.

A frequency table is a tally chart with an extra column for the totals.

The frequencies are the total of the tallies.

Number of people	Tally	Frequency
1		18
2		12
3		6
4		4

Surveys

When you are trying to find out some statistical information, it is often impossible to find this out from everyone concerned. Instead, you ask a smaller section of the population. This selection is called a sample.

Choosing a sample

The size of your sample is important. If your sample is too small, the results may not be very reliable. In general, the sample size needs to be at least 30. If your sample is too large, the data may take a long time to collect and analyse.

You need to decide what represents a reasonable sample size for the hypothesis you are investigating. You also need to eliminate bias. A biased sample is unreliable because it means that some results are more likely than others. It is often a good idea to use a random sample, where every person or piece of data has an equal chance of being selected.

You may want, however, to make sure that your sample has certain characteristics. For example, when investigating the hypothesis 'boys are taller than girls', random sampling within the whole school could mean that all the boys selected happen to be in Year 7 and all the girls in Year 11: this would be a biased sample, as older children tend to be taller. So you may instead want to use random sampling to select five girls and five boys from within each year group.

Example 45.2

Question

Candace is doing a survey about school meals.
She asks every tenth person going into lunch.
Why may this not be a good method of sampling?

Solution

She will not get the opinions of those who dislike school meals and have stopped having them.

Data collection sheets

When you collect large amounts of data, you may need to group it in order to analyse it or to present it clearly.

It is usually best to use equal class widths for this.

Tally charts are a good way of obtaining a frequency table, or you can use a spreadsheet or another statistics program to help you.

Before you collect your data, make sure you design a suitable data collection sheet or spreadsheet.

Think first how you will show the results, as this may influence the way you collect the data.

You could quite easily collect data from every student in your class on the number of children in their family.

One way to collect this data would be to ask each person individually and make a list like the following one.

Class 10G

1 2 1 1 2 3 2 1 2 1 1 2 4 2 1
5 2 3 1 1 4 10 3 2 5 1 2 1 1 2

The data for Class 10G is shown in the completed frequency table.

Using the same method to collect data for all the students in your year could get very messy. One way to make the collection of data easier is to use a data collection sheet. Designing a table like the one here can make collecting the data easy and quick.

Number of children	Tally	Frequency
1		12
2		10
3		3
4		2
5		2
6		0
7		0
8		0
9		0
10		1

Before you design a data collection sheet it is useful to know what the answers might be, although this is not always possible. For example, what would happen if you were using the tally chart shown above and someone said there were 13 children in their family?

One way to deal with this problem is to have an extra line at the end of the table to record all other responses. The table here shows how this might look.

Number of children	Tally	Frequency
1		12
2		10
3		3
4		2
5		2
More than 5		1

Adding the 'More than 5' line allows all possible responses to be recorded and gets rid of some of the lines where there are no responses.

Designing a questionnaire

A questionnaire is often a good way of collecting data. You need to think carefully about what information you need and how you will analyse the answers to each question. This will help you get the data in the form you need. Here are some points to bear in mind when you design a questionnaire.

- Make the questions short, clear and relevant to your task.
- Only ask one thing at a time.
- Make sure your questions are not 'leading' questions. Leading questions show bias. They 'lead' the person answering them towards a particular answer. For example: 'Because of global warming, don't you think that families should be banned from having more than one car?'
- If you give a choice of answers, make sure there are neither too few nor too many.

Example 45.3

Question

Suggest a sensible way of asking an adult their age.

Solution

Please tick your age group:

- ☐ 18–25 years ☐ 26–30 years ☐ 31–40 years
☐ 41–50 years ☐ 51–60 years ☐ Over 60 years

This means that the person does not have to tell you their exact age.

Note

When you are using groups, make sure that all the possibilities are covered and that there are no overlaps.

When you have written your questionnaire, test it out on a few people. This is called doing a pilot survey. Try also to analyse the data from the pilot survey, so that you can check whether it is possible.

You may then wish to reword one or two questions, regroup your data or change your method of sampling before you do the proper survey.

Any practical problems encountered in collecting the data would be described in a report of the survey

Two – way tables

Sometimes the data collected involve two factors. For example, data can be collected about the colour of a car and where it was made. You can show both of these factors in a two-way table

Example 45.4

Question

Peter has collected data about cars in a car park.

He has recorded the colour of each car and where it was made.

The two-way table below shows some of his data.

Complete the table.

	Made in Europe	Made in Asia	Made in the USA	Total
Red	15	4	2	
Not red	83			154
Total		73		

Solution

	Made in Europe	Made in Asia	Made in the USA	Total
Red	15	4	2	21
Not red	83	69	2	154
Total	98	73	4	175

73 cars are made in Asia so $73 - 4$ are not red.

154 cars are not red so $154 - 83 - 69$ are not red and are made in the USA.

All the totals can now be completed by adding across the rows or down the columns.

Note

A useful check is to calculate the grand total (the number in the bottom right corner of the table) twice.

The number you get by adding down the last column should be the same as the number you get by adding across the bottom row.

Averages and range

The mode and the median

The **mode** and the **median** are two different types of 'average'.

The mode is the most common number in a set of numbers.

A set of data may have more than one mode or it may have no mode.

The median is the middle number of a set of numbers arranged in order.

In a list of n numbers, the middle number is $\frac{1}{2}(n + 1)$.

Example 45.5**Question**

Here is a list of the weights of people in an exercise class.

73 kg, 58 kg, 61 kg, 43 kg, 81 kg, 53 kg, 73 kg, 70 kg, 73 kg, 62 kg, 60 kg, 85 kg

The instructor wants to know the average weight for the group.

Solution

One way is to find the most common weight, the mode.

The mode is 73 kg.

Another way is to put the weights in order and find the middle weight, the median.

There are 12 numbers.

$$\frac{1}{2}(n + 1) = \frac{1}{2}(12 + 1) = 6\frac{1}{2}$$

The middle numbers are the sixth and seventh.

The median is halfway between the two.

It is found by adding the values in the two middle positions and dividing by 2.

43 kg, 53 kg, 58 kg, 60 kg, 61 kg, 62 kg, 70 kg, 73 kg, 73 kg, 81 kg, 85 kg

$$\text{The median} = \frac{62 + 70}{2} = 66 \text{ kg.}$$

Mean and range

The mean is another type of average.

$$\text{Mean} = \frac{\text{sum of data values}}{\text{number of data values}}$$

The **range** is a measure of how spread out the data is.

$$\text{Range} = \text{largest data value} - \text{smallest data value}$$

Example 45.6**Question**

Madea and Parveen play cricket.

These are the number of runs they score in five innings.

Madea 7 8 5 4 7

Parveen 10 10 2 1 6

- Calculate the mean number of runs per innings for each player.
- Who is the more consistent batter?

Solution

$$\text{a Mean} = \frac{\text{total number of runs}}{\text{number of innings}}$$

$$\text{Mean for Madea} = 31 \div 5 = 6.2$$

$$\text{Mean for Parveen} = 29 \div 5 = 5.8$$

- Use the range to determine who has more consistent scores.

$$\text{Range} = \text{largest data value} - \text{smallest data value}$$

$$\text{Range for Madea} = 8 - 4 = 4$$

$$\text{Range for Parveen} = 10 - 1 = 9$$

Madea has a smaller range so she is the more consistent batter.

Which average to use when comparing data

When you are comparing sets of data, you need to compare the size of the data values and how spread out the data values are. You compare the size of the data values using an average. You have now met three types of average: the mode, the median and the mean.

You can use any of the three to compare sets of data, but in some circumstances one is better than the others. The mean has the advantage that it uses every data value but it can be misleading if there are one or two very high or very low data values.

The median is not affected by a few high or low values. The mode has the advantage that it is often quick and easy to find, but can be unreliable, particularly with a small sample of numbers. Also, some sets of data will not have a mode and some will have more than one

The range is a measure of how spread out the data values are. The greater the range, the more spread out the data values are. A greater range tells you that the data values are less consistent. The smaller the range, the less spread out the data values are. A smaller range tells you that the data values are more consistent. You will meet another measure of spread in Chapter 46.

Note

Always state your answer in the context of the question when comparing two sets of data.

So, in Example 45.7, don't just state that the potatoes from the market had a higher mean, state that this means that they are heavier on average.

Also don't just state that the range is smaller for the potatoes from the market, state that this means they are more consistently the same mass.

Example 45.7

Question

These are the masses, in grams, of 12 potatoes in a bag bought from a supermarket.

200 410 300 250 280 290 420 380 310 280 210 320

These are the masses, in grams, of 12 potatoes in a bag bought from a local market.

230 400 350 360 270 390 410 370 360 380 410 380

Compare the masses.

Solution

Arrange the data in order and calculate the averages and range.

Supermarket potatoes:

200 210 250 280 280 290 300 310 320 380 410 420

The mode is 280g.

The median is halfway between 290 and 300, so it is 295g.

Mean = $3650 \div 12 = 304\text{g}$ (to the nearest gram)

Range = $420 - 200 = 220\text{g}$

Market potatoes:

230 270 350 360 360 370 380 380 390 400 410 410

There are three modes, 360g, 380g and 410g.

The median is halfway between 370 and 380, so it is 375g.

Mean = $4310 \div 12 = 359\text{g}$ (to the nearest gram)

Range = $410 - 230 = 180\text{g}$

The mean shows that the potatoes from the market are heavier on average.

The range shows that the potatoes from the market are more consistently the same mass than the ones from the supermarket.

In Example 45.7, the mode was not useful. Either the median or the mean could be used to compare the masses, but there were no extreme values affecting the mean and the mean uses all the data. Also, the total quantity of the potatoes is important. The bags of potatoes from the market weigh more in total.

Example 45.8

Question

At the local shoe shop 15 pairs of ladies' shoes were sold.

These were the sizes sold. (They do not sell half sizes.)

3, 5, 6, 3, 3, 4, 4, 5, 5, 5, 5, 7, 3, 4 and 4

- Work out the mean, median and mode for these sizes.
- Which measure of average is the most useful in this case, and why?

Solution

- Arrange the data in order and calculate the averages and range.

3, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 6, 7

Mean = $66 \div 15 = 4.4$

The median is the 8th value.

Median = 4

Mode = 5

- The most useful measure in this case is the mode, because it can be used to find out which size of shoe to order most of.

Example 45.9

Question

The label on a matchbox states, 'Average contents 50 matches'.

A survey of the number of matches in ten matchboxes produced the following data.

45 51 48 47 46 47 49 50 52 47

- Find the mode, median, mean and range of the contents.
- What do these results tell you about the statement on the label?

Solution

- Arrange the data in order and calculate the averages and range.

45 46 47 47 47 48 49 50 51 52

Mode = 47

The median is halfway between the 5th and 6th values.

Median = $\frac{47 + 48}{2} = 47.5$

Mean = $\frac{482}{10} = 48.2$

Range = $52 - 45 = 7$

- The three measures of average – the mode, the median and the mean – are all below 50, so the statement on the label is untrue. The range shows that the number of matches varies greatly from one box to another.

Working with larger data sets

When working with larger data sets, it is often easier to see any patterns in the data if it is displayed in a frequency table.

Number of goals	Frequency
0	4
1	6
2	4
3	3
4	2
5	0
6	1

For example, this is a list of goals scored in 20 matches.

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

This is the same data displayed in a frequency table.

From the table, it is easy to identify the mode. It is the number of goals with the greatest frequency. Here the mode is 1 goal.

The median is the middle value. In this case, it is halfway between the 10th and 11th values.

Looking at the frequency column in the table, the 10th value is 1 and the 11th value is 2.

The median is, therefore, 1.5.

The table can also be used to calculate the mean.

There are: four matches with 0 goals $4 \times 0 = 0$ goals
 six matches with 1 goal $6 \times 1 = 6$ goals
 four matches with 2 goals $4 \times 2 = 8$ goals

and so on.

To find the total number of goals scored altogether, multiply each number of goals by its frequency and then add the results. You can add an extra column to your table to help you work out the values multiplied by their frequencies.

Number of goals	Frequency	Number of goals \times frequency
0	4	0
1	6	6
2	4	8
3	3	9
4	2	8
5	0	0
6	1	6
Totals	20	37

Then, dividing by the total number of matches (20) gives the mean.

Mean = $37 \div 20 = 1.85$ goals

Example 45.10

Question

Work out the mean, mode and range for the number of children in the houses in Berry Road, listed in this table.

Number of children (c)	Frequency (number of houses)	Total number of children (c \times frequency)
0	6	0
1	4	4
2	5	10
3	7	21
4	1	4
5	2	10
Totals	25	

Solution

Mean = $49 \div 25 = 1.96$ children

Mode = 3 children

Range = $5 - 0 = 5$ children

Working with grouped and continuous data

So far in this chapter, most of the data has been whole numbers only. The information is obtained through counting, not measuring. Data obtained by counting is called discrete data. Data obtained by measuring is called continuous data.

This is because you have the whole range of measurement between any two values.

For example, a length L given as 18 cm to the nearest centimetre means $17.5 \leq L < 18.5$.

Any length between these values will be recorded as 18 cm.

Often, however, the groups are larger, to make handling the data easier.

Discrete data can also be grouped.

For example, if you are looking at examination results in an examination marked out of 100, it is difficult to draw any conclusions from individual scores, so it is often better to group the marks.

For example, 1 to 10 marks, 11 to 20 marks, 21 to 30 marks, etc.

When you group data, some of the detail is lost.

For example, a frequency table may show that one student scored between 1 and 10 marks. You can no longer tell whether that student scored 1 mark, or 10 marks or any of the scores in between. For this reason, you cannot obtain exact values for the mode, median, mean or range.

The modal class may be found when data are given as a table or bar graph. It is the class with the highest frequency. Similarly, you cannot find the median from grouped data, but you can find the class which contains the median.

You can use the midpoints of the largest and smallest classes to estimate the range. Similarly, you can use the midpoint of the classes to calculate an estimate of the mean.

Example 45.11

Question

The frequency table shows some information about the heights of a year group in a school.

Height h (cm)	Frequency
$155 < h \leq 160$	2
$160 < h \leq 165$	6
$165 < h \leq 170$	18
$170 < h \leq 175$	25
$175 < h \leq 180$	9
$180 < h \leq 185$	4
$185 < h \leq 190$	1

Note

Add two columns to the frequency table to help you work out the mean: one column for the midpoint of each class and one for the midpoint multiplied by the frequency of the class.

- State the modal class.
- State the class which contains the median.
- Estimate the range.
- Calculate an estimate of the mean.

Solution

- The highest frequency is 25 so the modal class is $170 \text{ cm} < h \leq 175 \text{ cm}$.
- The total frequency is 65 so the median is the 33rd data value.
Adding the frequencies $2 + 6 = 8$, $8 + 18 = 26$, so 26 are less than 170 cm.
 $26 + 25 = 51$ so all of the 27th to 51st heights are in the $170 < h \leq 175$ class.
So the 33rd height is in the $170 \text{ cm} < h \leq 175 \text{ cm}$ class.
- Using the midpoints of the first and last classes,
estimate of the range = $187.5 - 157.5 = 30$

d

Height h (cm)	Frequency	Midpoint	Midpoint \times frequency
$155 < h \leq 160$	2	157.5	315
$160 < h \leq 165$	6	162.5	975
$165 < h \leq 170$	18	167.5	3015
$170 < h \leq 175$	25	172.5	4312.5
$175 < h \leq 180$	9	177.5	1597.5
$180 < h \leq 185$	4	182.5	730
$185 < h \leq 190$	1	187.5	187.5
Total	65		11 132.5

Estimate of the mean = $11\,132.5 \div 65$
 = 171.3 cm (to 1 decimal place)

Grouped discrete data

One method for finding the midpoint of a class is to add the end values of the class and divide by 2.

For example, if the class is $60 \leq s < 80$, then the midpoint is $\frac{60 + 80}{2} = 70$.

Strictly, this method is not correct. The class $60 \leq s < 80$ comprises the digits 60, 61, 62, 63, ..., 78, 79. The midpoint of these is $\frac{60 + 79}{2} = 69.5$.

However, using midpoints is an approximation and the error in using the simpler calculation is not significant.

Example 45.12

Question

The table shows the scores of 40 students in a history test.

Score (s)	Frequency
$0 < s \leq 20$	2
$20 < s \leq 40$	4
$40 < s \leq 60$	14
$60 < s \leq 80$	16
$80 < s \leq 100$	4
Total	40

Calculate an estimate of the mean score.

Solution

Score (s)	Frequency	Midpoint	Midpoint \times frequency
$0 < s \leq 20$	2	10	20
$20 < s \leq 40$	4	30	120
$40 < s \leq 60$	14	50	700
$60 < s \leq 80$	16	70	1120
$80 < s \leq 100$	4	90	360
Total	40		2320

Mean = $2320 \div 40 = 58$

Note

Don't forget to divide by the *total frequency*, not by the number of groups.

Key points

- The mode of a set of numbers is the most common value.
- The median of an ordered list of numbers is the middle value.
- Range = largest value – smallest value.
- The mean of a set of values = $\frac{\text{the sum of all values}}{\text{the number of values}}$.
- When comparing two sets of data, compare a measure of average and a measure of spread from each set. This is usually the mean and the range.
- For a discrete frequency table,
 Mean = $\frac{\text{total of (value} \times \text{frequency) products}}{\text{total of all frequencies}}$.
- Large sets of data are easier to deal with if they are grouped.
- For a grouped frequency table,
 Mean = $\frac{\text{total of (mid-group value} \times \text{frequency) products}}{\text{total of all frequencies}}$.
- The modal class of a grouped frequency distribution is the group with the largest frequency.

Revision questions

1.

The time taken for each of 120 students to complete a cooking challenge is shown in the table.

Time (t minutes)	$20 < t \leq 25$	$25 < t \leq 30$	$30 < t \leq 35$	$35 < t \leq 40$	$40 < t \leq 45$
Frequency	44	32	28	12	4

A student is chosen at random.

Find the probability that this student takes more than 40 minutes.

2.

The frequency table shows information about the time, m minutes, that each of 160 people spend in a library.

Time (m minutes)	$0 < m \leq 10$	$10 < m \leq 40$	$40 < m \leq 60$	$60 < m \leq 90$	$90 < m \leq 100$	$100 < m \leq 120$
Frequency	3	39	43	55	11	9

Find the probability that one of these people, chosen at random, spends more than 100 minutes in the library.

3.

A bag contains blue, red, yellow and green balls only.

A ball is taken from the bag at random.

The table shows some information about the probabilities.

Colour	Blue	Red	Yellow	Green
Probability	0.15	0.2		0.43

Complete the table.

4.

A group of 200 people were asked which city they would like to visit next.
The table shows the results.

City	London	Paris	New York	Tokyo
Number of people	50	48	56	46

A person from the group is chosen at random.

Write down the probability that this person would like to visit either Paris or Tokyo next.

5.

There are only red counters, blue counters, white counters and black counters in a bag.

The table shows the probability that a counter taken at random from the bag will be red or blue.

Colour	red	blue	white	black
Probability	0.2	0.5		

The number of white counters in the bag is the same as the number of black counters in the bag.

Tania takes at random a counter from the bag.

Work out the probability that Tania takes a white counter.

6.



Tanya plants some seeds.

When a seed produces flowers, the probability that the flowers are red is 0.6 and the probability that the flowers are yellow is 0.3.

Tanya has a seed that produces flowers.

Find the probability that the flowers are not red and not yellow.

7.



Suleika has six cards numbered 1 to 6.

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

She takes one card at random, records the number and replaces the card.

i) Write down the probability that the number is 5 or 6.

ii) Suleika does this 300 times.

Find how many times she expects the number 5 or 6.

8.



Morgan picks two of these letters, at random, **without** replacement.

Find the probability that he picks

- i) the letter Y first,
- ii) the letter B then the letter Y.

9.



The diagram shows 5 cards.

Donald chooses two of the five cards at random, without replacement.

He works out the total number of dots on these two cards.

- i) Find the probability that the total number of dots is 5.
- ii) Find the probability that the total number of dots is an odd number.

10.

Angelo has a bag containing 3 white counters and x black counters.
He takes two counters at random from the bag, without replacement.
Complete the following statement.

The probability that Angelo takes two black counters is $\frac{x}{x+3} \times \frac{\dots\dots\dots}{\dots\dots\dots}$