10 Statistics and probability

Mastery Professional Development

Solutions to exemplified key ideas

Click the heading to move to that page. Please note that these materials are principally for professional development purposes; solutions are provided to support this aim.

10.1.1.4 Estimate the mean and range for a data set represented in a grouped frequency table

Interpret a grouped frequency table

Example 1:

a) False
b) True
c) Not enough information
d) Not enough information
e) False
f) True
g) Not enough information

Example 2:

a) Table B and Table D

b) Responses may vary but should demonstrate an understanding that while including all possible options is important, Wisal needs to be aware that she would not be able to estimate a single value to represent an open category (whereas, with fixed groups, it is possible to calculate using the lower/upper bounds or mean average). Also, it isn't clear if '20+ inches' includes 20 inches.

Example 3:

Responses may vary but should demonstrate an understanding that Miss Cauchy grouped without using inequalities; which is appropriate in this instance as it involves discrete data (assuming no fractional marks were awarded in the test). While Mr Schwarz's use of inequalities is appropriate for most of his groups, a student who scored 0 on the test cannot be recorded in Mr Schwarz's table because the first class interval is $0 < s \leq 10$. This could be overcome by changing the first group to $0 \leq s \leq 10$, or by using inequalities of the form $0 \leq s < 10$ (with the final group being $40 \leq s \leq 50$ to prevent the same issue happening for students with full marks).

Explore the effects of grouping data in different ways, including that some information is lost

Example 4:

Height (h)	*Frequency*
$1.0 \leq h \leq 1.2$	*2*
$1.2 < h \leq 1.4$	*2*
$1.4 < h \leq 1.6$	*3*
$1.6 < h \leq 1.8$	*4*
$1.8 < h \leq 2.0$	*7*
$2.0 < h \leq 2.2$	*6*
$2.2 < h \leq 2.4$	*4*
$2.4 < h \leq 2.6$	*3*
$2.6 < h \leq 2.8$	*2*
$2.8 < h \leq 3.0$	*2*

Example 5:

a) Students' responses will vary but might include:

Same	Different
• Holly and Noel have both grouped the data. • The minimum possible age recorded (one) is the same in each table. • Both tables represent the ages of the same 20 people. • The tally and frequency column totals are the same for the highest group in both tables, showing no-one is actually over 90 years old.	• Holly's table is in groups of 20, whereas Noel's is in groups of 10. • The maximum possible age recorded is different in each table. • The tally and frequency columns are different for most of the groups. • Noel's table doesn't allow for the under-ones.

b) Holly's table: 100 – 0 = 100 years. Noel's table: 90 – 1 = 89 years.

c) 89 – 2 = 87 years.

d) Responses may vary but should demonstrate an understanding that when data is recorded in grouped categories, we no longer know the exact values. In part b we calculated maximum possible ranges based on the maximum and minimum values in each category. In part c we calculated the exact range.

Know why we multiply by the midpoint when calculating the mean from a grouped frequency table

Example 6:

a) Anoushka is not correct. Explanations may vary but should demonstrate an understanding that the two tables are not comparable as they show different information, and that it is potentially a coincidence that the final row in both tables has the same frequency. We do not have enough information to know if the one person who visited all five stalls is the same person who stayed for 41-50 minutes.

b) Anoushka is not correct. Students' explanations may vary but should demonstrate an understanding that summing the five possible numbers of stalls, and then dividing by five rows, is a meaningless calculation. To find the mean number of stalls visited, Anoushka needs to sum the total frequency of stalls visited and then divide by the total frequency of people. This means recognising that six people visited one stall (so six total visits); eight people visited three stalls (so 24 total visits); five people visited four stalls (so 20 total visits); and one person visited five stalls (so 5 total visits). Summing 6, 24, 20 and 5 gives a total number of stalls visited as 55. Dividing by the total number of people (20) gives a mean of 2.75 stalls visited per person.

c) This estimated mean is likely to be significantly lower than the actual mean as Faris is using the smallest value in each group for each length.

d) A more representative value would be to use the midpoint of each group. Students may argue the case for other values, such as the upper bound of each group; this provides a teaching opportunity to clarify why the midpoint is a suitable value to choose.

e) Students' responses may vary but should demonstrate an understanding of what would constitute useful information for deciding if an event was successful. As this is a fundraising event, it could be argued that they need to gather information about the amount of money raised. It is also difficult to make judgements without a point of comparison – for example, a target amount of money raised, or the outcomes from a previous event.

Example 7:

a) (i) 1.25 (ii) 1.75 (iii) 2.25 (iv) 2.75

Responses may vary but should demonstrate an understanding that:

b) The answers to part a are the same as the midpoints in Carly's table.

c) Once in the grouped frequency table, Carly no longer knew the exact value of each piece of data. Carly used the midpoint as an estimated value for each, as it is the mean of the upper and lower class limits. She then multiplied this by the frequency to find an estimate for the sum of each piece of data in each row.

d) We are assuming that the mean of the upper and lower class limit is a representative value for the heights of the apple trees; i.e., that the heights of the trees are distributed relatively evenly across the class interval, and that the mean tree height in each class interval is equal to the mean of the upper and lower class limit.

Find the mean from a grouped frequency table

Example 8:

Responses may vary but should demonstrate an understanding that:

a) Dominic has divided by the total number of class intervals instead of the total frequency of apple trees, leading to an incorrect calculation.

b) Dominic first needs to 'sense check' and decide whether 14 m *could* be representative of trees that are between 1 m and 3 m in height. needs to understand that calculating the mean requires dividing by the total quantity being measured. In this case, we need to find the sum of the frequencies rather than the number of class intervals.

Recognise that calculations from grouped frequency tables will be estimates

Example 9:

Responses may vary, but they should demonstrate an understanding that Amy used the raw data to calculate the actual mean. In contrast, Ben and Callie relied on grouped frequency tables, resulting in an estimated mean based on assumed average values within each class interval. Additionally, Ben's data has a class width of 10 , whereas Callie's data has a class width of 20, which affected the midpoint used and therefore the accuracy of their estimates.

Understand that the accuracy of the estimate depends upon the distribution of data within the class interval

Example 10:

Monday: The midpoint is close to the mean of the ages, and so it is an accurate value to use. Three of the values are clustered near the midpoint, while one is significantly farther away. The combined distance of the three close values from the midpoint is slightly smaller than the distance of the outlier.

Tuesday: The midpoint is close to the mean of the ages, and so it is a relatively accurate value to use. Two of the values are equidistant from the midpoint but on opposite sides, while the other two are nearly the same distance away, also on opposite sides. However, the furthest value slightly skews the mean in its direction.

Wednesday: The midpoint is not close to the estimated mean of the ages, so it will not be an accurate value to use. Two values are equidistant from the midpoint on opposite sides, which helps balance their impact. However, the remaining two values are clustered on the same side and farther from the midpoint, which pulls the mean in that direction.

Thursday: The midpoint is close to the mean of the ages, and so it is an accurate value to use. Two of the values are equidistant from the midpoint but on opposite sides, while the other two are nearly the

same distance away, also on opposite sides. However, the furthest value slightly skews the mean in its direction.

Friday: The midpoint is exactly equal to the mean of the ages; it will have perfect accuracy. There are two pairs of values, each equidistant from the midpoint but positioned on opposite sides.

Example 11:

Responses may vary but should demonstrate an understanding that the most accurate mean came from when the data was in five groups as it is closest to the mean for the raw data. However, the differences here are quite small, meaning all estimates are reasonably close to the actual mean.

Example 12:

Responses may vary but should demonstrate an understanding that:

a) Since Salma uses individual values, she can determine the exact highest and lowest attendance figures for each class, allowing her to calculate the true range. By contrast, Ren relies on a grouped frequency table, meaning he does not have access to the precise values. Instead, he assumes the maximum possible value as the largest and the minimum possible value as the smallest when calculating the mean. Grouped data can obscure subtle variations in the dataset.

b) Valid data sets will consist of 30 values that could be distributed into Ren's grouped frequency table. The largest value for Class A will be 30 and the smallest 0. For Class B, the maximum and minimum values can vary, but the difference between the largest and smallest values must be 27.

c) Additional observations may be made based on students' individual data sets, but all data sets will contain the following points of comparison:

Same	Different
• Thirty individual data points. • The distribution of values is the same as in Ren's table (i.e. 20 values between 26 and 30 days, five between 21 and 25 days etc). • The maximum and minimum values for Class A are always 30 and 0 respectively.	• The maximum and minimum values are likely to be different for Class B. Possible answers are 30 and 3, 29 and 2, 28 and 1 or 27 and 0.

Example 13:

a) 2.95 – 1.01 = 1.94

b) Jonny is incorrect. Explanations may vary but should demonstrate an understanding that the range of a grouped frequency table depends on how the data is grouped, and that the group boundaries are largely an arbitrary decision. In the tables given, it has consistently been assumed the smallest value in the dataset is 1.0 metres and the largest is 3.0 metres, which results in a range of 2.0 metres. Different groupings could set different interval boundaries, affecting the estimated range. This could involve using either equally or unequally sized groups.

c) Accept student responses where the difference between the lower and upper bounds is greater than 2.0. The upper bound of the lowest group must not be lower than 1.0 and the lower bound of the highest group must not exceed 3.0. Using seven equally-sized groups, for example, would give the smallest group as $1.0 \leq h \leq 1.3$ and the largest group as $2.8 < h \leq 3.1$, so the range would be estimated to be $3.1 - 13.1 - 1.0 = 2.1$ metres.

d) Students' responses may vary but should recognise that we do not have enough information to answer this question. We do not have the raw data, and so we do not know if we can change the groups in a way that reduces the range but still includes the height of every tree. It would only be

possible if the actual minimum height was greater than 1.0 metres and/or the actual maximum height was less than 3.0 metres.

e) Responses may vary but should demonstrate an understanding that considerations include:

- The choice of class boundaries in the grouped frequency table.
- Whether the grouping included all values within the dataset or artificially extended beyond them.
- How interval selection impacts the estimated range.

10.1.2.4 Describe the properties of a population using appropriate summary measures

Appreciate how different measures of central tendency may be more representative depending on the data they are summarising

Example 1:

Responses may vary but should demonstrate an understanding that:

a) The median is the most representative average of Fi's results. Fi's scores are similar in four out of five subjects, ranging between 54% and 62%, but in geography she has scored much more highly. This has resulted in the mean value of 65% being the least representative, as it is distorted by the 97% geography score, and all her other scores are actually below the mean of 65%. The mode is also representative in Fi's case, as it is the same value as the median, but this might not necessarily be the case as, in a context such as this, there might not always be a modal score.

b) Fi may choose to share the mean value for her test results, as this is the highest value. It suggests that her test results overall are higher than they actually are.

c) An advantage of using the mean is that it uses all values within the dataset. However, a significant disadvantage is that it may not provide an accurate measure of central tendency when there are outliers or when the data are skewed.

An advantage of using the median is that it is not affected by outliers. The converse of this is that it does not take account of all values.

Understand how data characteristics affect measures of central tendency and spread

Example 2:

Responses may vary but should demonstrate an understanding that:

a) The **range** is the difference between the highest and lowest values in a dataset. Since the given range is £12 399, this confirms that the difference between highest earner's salary the lowest earner's salary.

The **mode** is the most frequently occurring value in a dataset. Since the mode is £21 674, this means that at least two cleaners have this salary; if it appeared only once, it wouldn't be classified as the mode.

The **median** represents the middle value in an ordered dataset. Since there are 15 cleaners, the median salary is the eighth value when sorted in order. This means that seven cleaners earn below the median and seven cleaners earn above it.

The **final statement** suggests that the upper half of the salaries is more spread out compared to the lower half. Given that the mean (£18 024) is higher than the median (£16 682), this indicates a right-skewed distribution where the higher salaries pull the average up. If the upper earners were closer to the median, the mean would be lower and closer to the median.

b) We can infer that Chris's salary is much higher than that of the cleaners, as the mean increases significantly when including his wage. As the range is also much larger, his salary must be substantially higher than the previously highest-earning cleaner.

c) The **median** remains at £16 682, which means the middle value in the ordered dataset is unchanged. Since there's an additional piece of data, the position of the middle value shifts slightly (the 8.5th value, rather than the eighth). As the median is unchanged despite this, there must be multiple cleaners earning this exact amount (i.e., at least the eighth and ninth values, when ordered, are £16 682).

As we now know that at least two cleaners earn the median amount, but that the **mode** is a different value, this must mean that more than two (i.e., at least three) cleaners earned £21 674.

The **range** increased by £22 702 once Chris included his wage and the mean increased, showing that Chris's salary is £22 702 higher than the next highest earner.

The mean for the 15 cleaners was £18 024, which means the total of all their wages can be calculated as $15 \times$ £18 024 $=$ £270 360. The mean including Chris is £19 671, which means the total of all their wages including his can be calculated as $16 \times$ £19 671$=$ £314 736. The difference between these two totals, once Chris's wage was added, is £44 376.

As Chris's wage is £44 376 and the range is now £35 101, the **lowest** salary in the company is £44 376 $-$ £35 101 $=$ £9 275.

Example 3:

a) Responses may vary but should demonstrate an understanding that:

- Table A represents the student population. We know this because the range is 5 years (16−11). Also, all measures of central tendency are teenage years.
- Table B represents the staff population. We know this because all measures of central tendency are adult years. It is also feasible for there to be a range of 39, as the youngest member of staff could be 22 and the oldest 61.
- Table C represents the whole school population. We know this because the median and mode have remained the same at 14 and 15 respectively, showing there are a lot of teenagers in the population. However, the significant increase in mean and range shows that there are also adults being included.

b) Teachers should check students' responses and accept accurate responses based on the students' own age relative to their year group. For a student of age n, this should include:

- Median and mode being equal to n and/or $n \pm 1$.
- Mean being between n and $n \pm 1$.
- Range being 1.

Example 4:

Responses may vary but should demonstrate an understanding that:

- 4.5 must be the size of three or more shoes with no other size appearing as many times.
- Of the pairs of shoes listed, size 1 must be the smallest size and size 9 must be the largest.
- The total of all of the different shoe sizes should be 100.

Example 5:

Statement	***Subject***	***Explanations may vary but should indicate students understand:***
A	French	The smaller range and interquartile range (IQR) indicate that the marks in the class were more consistent, with less variation between students.
B	Geography	The larger IQR shows that, even within the middle 50% of the data, the marks were widely spread. The presence of modes at 20 and 90 suggests clusters at both the lower and higher ends of the marks, reinforcing the idea that some students were informed about the test while others were not.
C	RE	All three measures of central tendency were lower, indicating that many students found it particularly difficult. The high range suggests that this was true for the majority of the class, though not for all students.
D	Mathematics	The low mean and median indicate that the majority of students did not perform particularly well on this test. The relatively low IQR suggests consistency within the middle 50% of results. However, a mode of 80 highlights that more than one student significantly outperformed the majority.
E	Science	As the mean and median were the same, we know that the results were evenly distributed. The large range tells us that they were evenly distributed across a large spread of scores.

10.2.1.1 Understand that graphical representations of data can support comparison between data sets and identification of trends over time

Example 1:

a) Responses may vary but might include:

Same	Different
• Similar trend after the fourth hour on both days: a rise in temperature, which peaks in the afternoon and drops into the evening. • On both days, the final temperature is higher than the initial temperature. • Warmest part of the day is around 5pm – 6pm.	• The temperatures are consistently higher on one day compared to the other. • The temperature on the warmer day drops initially, whereas the colder day stays constant before increasing. • The range in temperatures is greater on day 1.

b) Responses may vary but should demonstrate an understanding that day 1 was warmer, so was likely to have been in summer; day 2 was colder, so was likely to have been in winter.

Example 2:

a) January. Explanations may vary but should demonstrate an understanding that more gas was used and, as gas is used to fuel many heating systems, the amount of gas used is likely to be higher when outside temperatures are colder.

b) Responses may vary but might include:

Same	Different
• The median and upper quartile appear to be the same. • The maximum value is similar.	• The lower quartile and minimum value are higher for April. • The range and interquartile range are greater for July.

Explanations may vary, but the similarities suggest that the top 50% of the data is comparable, so for half of each month a similar amount of gas was used. This suggests that the temperatures were similar for half of each of the months of April and July. The differences suggest that there was a greater range in temperature in the month of July, compared with the month of April, since less gas was used for around half of the days in July.

c) Responses may vary, but should demonstrate an understanding that the weather was likely to have been colder in January 2001 than in 2000, since more gas was used.

Example 3:

a) Red curve: weekday; blue curve: weekend.

b) Responses may vary but should demonstrate an understanding that the red curve rises steeply in the first two hours, indicating that people visit the shop early on weekdays, perhaps before starting work. By contrast, the blue curve starts more gradually, probably reflecting later activity on weekends. It then shows a steep rise later in the day, which could be explained by people visiting the shop for evening refreshments, which does not happen on a weekday.

c) Students might suggest the owner considers closing earlier on a weekday, since so few customers visit during the final hour; or staying open later at the weekend, since the line did not plateau by the time the shop closed.

10.2.1.3 Interpret features of data from a box plot and use them to make comparisons between data sets

Interpret the key values in a box plot

Example 1:

a) (i) lowest age: 4, highest age: 52; (ii) 9 years old; (iii) 10 years old; (iv) 6.5 years old.

b) Scale from 0-60 with marks at 4, 6.5, 9, 10 and 52. Note that, at this stage of students' learning, these marks may take any form and not necessarily be consistent with a formal box plot.

c) (i) minimum age: 5 years, maximum age: 52 years; (ii) 6 years old; (iii) 7 years old; (iv) 6 years old.

d) Scale from 0-60 with marks at 5, 6 (duplicate), 7 and 52. Note that, at this stage of students' learning, these marks may take any form and not necessarily be consistent with a formal box plot.

e) Students' responses will vary but could include:

Same	Different
• Five data points have been recorded from each set of data, and marked on the same scale. • The maximum value for both sets of data is 52. • The overall spread of both sets of data is similar – the range of the first set is only one more than the second set.	• The minimum value for the first data set is 4, but for the second is 5. • The median value for the first data set is 9, but for the second is 7. • The halfway points for the upper and lower half of each set of data are different: 10 and 6.5 respectively for the first set, and 7 and 6 for the second set. • The second set of data is less varied than the first set of data.

Example 2:

a) 50 b) 100 c) 150 d) 50 e) 100

Example 3:

a) Responses may vary but should demonstrate an understanding that:

Bar chart A: the parts of the bars shaded dark grey are the values in the 'box'. The parts of the bars that relate to values in the whiskers are dotted.

Bar chart B: this follows the same principles but differentiates four different groups to demonstrate the values that lie between each of the quartiles.

b) Responses may vary, but should demonstrate the understanding outlined below:

	Box plot	Bar chart A	Bar chart B
(i) the quartiles	The three verticals of the 'box'.	Two of the quartiles are shown by the change from spotted shading to dark grey shading, but the median (or 'second quartile') is not identified.	The change between the four different types of shading.

	Box plot	Bar chart A	Bar chart B
(ii) the middle 50% of the data	The 'box'.	The parts of the bars shaded dark grey.	The parts of the bars shaded grey and dark grey.
(iii) the top 25% of the data	The line between the rightmost vertical line of the 'box' to the end of the final whisker.	The second set of bars that are spotted to the right of the chart.	The striped parts of the bars.
(iv) the median of the data	The vertical line inside the 'box'.	Not represented.	The change from the dark-grey shading to the striped shading.

c) Students' answers will vary, as this question asks for their opinions and reasoning. Teachers should be prepared to accept and compare answers, to support them to assess whether students have an understanding of the distribution of the data from the representations.

Example 4:

a) Students' answers will vary, but observations about similarities and differences might include:

Same	Different
• There is a frequency in each of the possible grades, showing that each grade was awarded. • In all the graphs, the percentage of students achieving U was 3% or less. • In both Engineering and Maths, the majority of students were awarded the middle grades, and fewer students were awarded the 'extreme' grades of U, 1, 2, 7 8 and 9.	• The lower quartile and minimum value are higher for April. • The range and interquartile range are greater for July. • The bar chart for Italian is negatively skewed.

Students' conclusions about the grade distribution will also vary, but they should demonstrate an understanding that the profile of grades in Italian was very different to the other two subjects, with the most-awarded grade being the top grade of 9 and almost half of the cohort achieving this. There is a similarity in the grade distribution of Maths and Engineering, in that fewer students were awarded the grades at either end of the spectrum (for example, U-2 or 7-9). However, broadly the same number of 4, 5 and 6 grades were awarded in Engineering, with 3 being the most commonly awarded grade. This differs from Maths where the most commonly awarded grade was 5, closely followed by 4, with 3 and 6 being awarded much less often.

b) Responses may vary, but might include the following features:

- The shape of the bar chart for Engineering shows a peak of data towards the left side, where the lower grades are concentrated; this may suggest that students generally found this subject more challenging than Maths.
- The shape of the bar chart for Italian shows a significant peak for the highest grade, showing that the majority students excelled at this subject; this could be indicative of a high proportion of native speakers taking the exam.

c) Sketches based on student predictions. Teachers should primarily attend to the values that students have chosen, to check their understanding of the distribution of the data. Teachers should also check the features of the sketches, to ensure students have included all relevant information for box plots.

d) Two box plots representing the summary statistics are provided in the table.

e) Students' reflections will vary, based upon their predictions in part c and the accuracy of their box plots in part d.

Example 5:

a) Students should notice that there are 35 data points, and so they need to identify the fixed values from the box plot and then the number of values which need to be between these fixed values:

- The 18th value is the median and must be 70.
- The 9th value is the lower quartile value and is 62.
- The 27th value is the upper quartile value and has to be 79.

There are 8 values between each of these – and above and below - dividing the data into four equal parts. The maximum and minimum values are also known.

	Min		Lower IQ		Median		Upper IQ		Max
Value	43		62		70		79		98
Number of scores		1	7	1	8	1	8	1	7	1	

Score on maths test	Frequency
0-10	0
11-20	0
21-30	0
31-40	0
41-50	2
51-60	6
61-70	12
71-80	10
81-90	4
91-100	1

Hence, accept student responses that adhere to the following conditions:

- Frequency of 0 in the first four class intervals
- Less than 25% of the data (8 scores or less) represented in the 41-50 and 51-60 class intervals
- More than 25% but less than 50% (between 9 and 17 scores) of the data represented in the 61-70 class
- More than 25% of the data represented in the 71-80 class
- Less than 25% of the data represented in the 81-90 and 91-100 class intervals
- A value greater than 0 in the final class

An example of a valid table is given above right, which is obtained through recognising that the distributions for the two subjects are the same (and therefore a possibility is to use the same frequencies).

b) Responses may vary but should demonstrate an understanding that yes, because there can be some variation in the frequency amounts with the minimum, lower quartile, median, upper quartile and maximum scores being unaffected.

Understand that comparing the interquartile range provides a way of comparing dispersion

Example 6:

a) Table A: history; Table B: science

b) Students' explanations will vary, but should demonstrate an understanding that, while the tests for both subjects resulted in scores across the possible range of marks, the majority of test scores for science lie more consistently around the median score. In contrast, there is more variability in the test scores for history.

Example 7:

Responses may vary but should demonstrate an understanding that:

a) The two box plots have the same lower quartile, median, and upper quartile. Since the length of the box is identical, the distribution of the middle 50% of scores is the same for French and science. However, the minimum and maximum scores differ, indicating a much greater range in science.

b) The median is the same for both subjects, and the minimum and maximum values are also identical, indicating the same overall range. However, the lower and upper quartiles differ significantly, resulting in different inter quartile ranges (IQRs). This demonstrates a much greater spread in history scores compared to science.

c) In both parts a and b, the given measure of central tendency (median) was the same, one measure of spread was the same, and one measure of spread was different. In part a, the identical IQRs indicated a similarity in the spread of the middle 50% of the data, but a vast difference in range indicated a very different overall spread of test scores. In part b, the identical range indicated that the extremes of marks were similar (for example, the very highest and lowest marks), but that the middle 50% of the scores were spread very differently among the cohort. Without looking at IQR and range for both, it would not be possible to compare the spread of data accurately.

Understand how to identify skew in a data distribution

Example 8:

Box plot C. Explanations may vary but should demonstrate an understanding that almost half the class achieved scores within the lowest two groups (i.e., the data are positively skewed) so the median is closer to the lower end of the box. Students may also identify that the 18^{th} value will be the median value, and from the table this means that it is in the 21-30 interval.

Know how to determine whether a data value is an outlier or not

Example 9:

Responses may vary but should demonstrate an understanding that Zoe should consider whether the maximum score of 95 is an outlier, and therefore whether to exclude it from the box plot. See the guidance column for further suggestions of how to frame this discussion with students.

Appreciate the information that can and cannot be determined from a box plot

Example 10:

Responses may vary but should demonstrate an understanding that the two data sets produce the same summary data, despite the raw values being different. The relevant information for the box plots (minimum value, lower quartile, median, upper quartile, maximum value) are the same.

Example 11:

A: True B: Cannot tell from the box plot C: Cannot tell from the box plot D: True

Example 12:

Responses may vary but should demonstrate an understanding that:

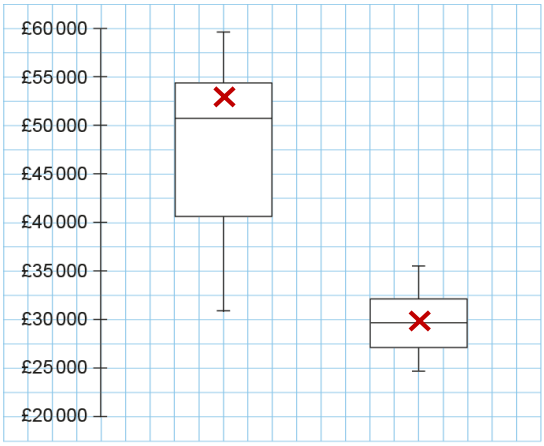
a) The median annual net income for the left-hand box plot is just over £50 000, with an upper quartile of £55 000 and a lower quartile slightly above £40 000. This indicates that the area is affluent, but that there is a wide range of annual net incomes. In contrast, the median annual net income for the right-hand box plot is just under £30 000, with an interquartile range of approximately £5 000 and a relatively small overall range. This suggests that the area is less affluent, and that most people have similar earnings.

b) Eluned's comments suggest that the left-hand box plot represents Kensington and Chelsea, and the right-hand box plot represents Flintshire. This is because most of the summary statistics for the right-hand box plot are lower than the left-hand box plot.

c) Eluned's statement can be argued to be accurate. The much higher median in the first box plot indicates that people in Kensington and Chelsea earned more on average than those in Flintshire. Additionally, the lowest net income in Kensington and Chelsea was higher than the median net income in Flintshire. However, while the majority of people in Kensington and Chelsea earn more than the majority of people in Flintshire, this is not true for everyone. Some of the lowest earners in Kensington and Chelsea earn less than the highest earners in Flintshire.

d) The wider IQR in Kensington and Chelsea suggests that earnings for the middle 50% of people there were less consistent and more spread out, compared with a smaller IQR showing the middle 50% of Flintshire residents earn a comparable amount.

e) Approximate positions indicated below:



f) In Kensington and Chelsea, the mean is smaller than the median, indicating that the data are negatively skewed (i.e., that the top half of the data is more condensed, and the bottom half of the data is more spread out). In Flintshire, the median and mean are almost the same, suggesting that the data are spread evenly on both sides.

Example 13:

Responses may vary but should demonstrate an understanding that:

a) In Class 1, the lowest value, lower quartile, and median are all zero, indicating that at least half of the students received zero detentions. The highest value was two, showing that no student received more than two detentions.

In Class 2, the minimum was also zero, showing that some students still did not get detentions. However, the median was two, the highest value was five, and the IQR was three, indicating that

there were more detentions given out on average, and that at least one student received five detentions.

b) Students' motivation for choosing a class may vary, but they are likely to select Class 1. There are far fewer detentions on average, suggesting that the behaviour is better and/or detentions are less likely to be given.

c) The bar chart tells us the frequency of each number of detentions (i.e., how many students were issued zero, one, two, three, four or five detentions).

d) The box plot tells us the value of key summary statistics, including the quartiles and median.

e) Teachers should allow students to draw their own conclusions, listening carefully to their justifications.

10.2.1.4 Understand the idea of accumulation and why it is useful

Use the idea of accumulation to find the median for a data set presented in a frequency table

Example 1:

Responses may vary but should demonstrate an understanding that Jake could use the running total column to locate the 8th value.

Example 2:

Responses may vary but should demonstrate an understanding that Simon could use the running total column locate both the 8th and 9th values and then calculate the mean of these.

Know how to determine the lower and upper quartiles for data sets presented in frequency tables

Example 3:

a) 52 kg – 49 kg = 3 kg.

b) 53 kg – 47.5 kg = 5.5 kg.

c) Responses may vary, but should demonstrate an understanding that there are 15 ewes and 16 rams recorded, and that this affects whether the quartiles are actually values in the data set. There is an odd number of ewes, so it is easier to locate the $\frac{(n+1)}{4}$th and $3\frac{(n+1)}{4}$th position (4th and 12th data values). There is an even number of rams: for the lower quartile, the 4th and 5th data values are not the same, and similarly for the upper quartile, the 12th and 13th data values are different. Therefore, you are using values not within the data set and a calculation (the mean of each pair of values) is needed to establish the upper and lower quartiles.

Begin to appreciate how cumulative frequencies can be represented on a graph

Example 4:

Explanations may vary but should demonstrate the understanding described after the true/false answer.

a) True. Cumulative frequency curves always start at 0 on the y -axis, indicating that the cumulative frequency begins at zero. However, the x -axis can start at a value greater than 0 depending on the data being represented.

b) False. Two cumulative frequency curves will only start and finish at the same place if they cover the same data range on the x-axis and have the same total frequency

c) False. The median is found halfway up the y-axis and then read across to the x-axis. It doesn't have to be in the middle of the x-axis. It will depend on the shape and spread of the data.

10.2.1.7 Construct histograms for a given data set

Know that frequency density rather than frequency is plotted on the y -axis of a histogram

Example 1:

1 cm (× 1 cm) 0.5 cm (× 2 cm) 5 cm (× 0.2 cm) 0.1 cm (× 10 cm)

Example 2:

Histogram B. Students' reasoning may vary, but should indicate that they understand the different class widths make it challenging to identify the frequencies; it looks as though the majority of dogs are in the second group, as this covers the greatest area.

Example 3:

Graph B. As with *Example 2*, students' reasoning will vary, but should demonstrate an understanding that it is difficult to ascertain the frequency for the newly-merged first group, as it looks disproportionately large compared with the original two groups in the first histogram.

Example 4:

Responses may vary but should demonstrate an understanding that:

a) Although the first bar is taller, the larger group is the second group of those who arrived more than five minutes late. This is indicated by the larger area on the histogram (including more faces).

b) The two bars in the histogram each still have the same area, and therefore the same frequency, as in the first histogram. As the areas have not changed, we know that no-one arrived between five and ten minutes late.

Example 5:

a) Responses will vary but may include:

- Between four and five was the most common number of months that students had to wait.
- No students were waiting between 10 and 11 months for their birthday.
- There were 19 students in the class.
- There was the same number of students waiting 5 to 6, 8 to 9 and 9 to 10 months.
- There was the same number of students waiting 0 to 1, 2 to 3, 3 to 4, 6 to 7, 7 to 8 and 11 to 12 months.

b) Responses may vary but should demonstrate an understanding that:

Same	Different
• The histograms still represent the same 19 students, and the same number of months. • The sticky notes still cover the same area.	• The number of months has been grouped differently each time. • In the first chart, the sticky notes are arranged into groups of 2 months. • In the second, it is in groups of 3 months.

c) First histogram:

- Zero to 4 months covers an area of 6 sticky notes (1.5 high and 4 wide).
- Four to 8 months covers an area of 8 sticky notes (2 high and 4 wide).

- Eight to 12 months covers an area of 5 sticky notes (1.25 high and 4 wide).

Second histogram:

- Zero to 1 months covers an area of 1 sticky note (1 high and 1 wide).
- One to 7 months covers an area of 12 sticky notes (2 high and 6 wide).
- Seven to 9 months covers an area of 3 sticky notes (1.5 high and 2 wide).
- Nine to 12 months covers an area of 3 sticky notes (1 high and 3 wide).

Example 6:

(0 to 5 years), 5 to 55 years, 55 to 80 years, 80 to 82.5 years.

Example 7:

Graph C. Explanations may vary but should demonstrate an understanding that, in Graph A, frequency is plotted rather than frequency density. In Graph B, although the y-axis is labelled correctly, the frequency density hasn't been calculated. Graph C is accurate, as all frequencies have correctly been divided by the class width to give the frequency density for each bar.

Understand the relationship between frequency, frequency density and class width

Example 8:

B: 0.5. Explanations may vary, but should demonstrate an understanding that for the area of the bar to represent a frequency of 10, we need to divide 10 by the class width of 20.

Example 9:

C: 20. Explanations may vary, but should demonstrate an understanding that if a frequency density of 0.5 is representing a frequency of 10, then the area of the bar needs to be 10. This means that the class width needs to be 20 (since $10 \div 0.5 = 20$).

Example 10:

C: 10. Explanations may vary, but should demonstrate an understanding that the area of the bar of height 0.5 and class width 20 is $0.5 \times 20 = 10$.

Example 11:

a)

Weight (g)	*Class width*	*Frequency*	*Frequency density*
$0 < g \leq 20$	20	*80*	4
$20 < g \leq 30$	*10*	20	2
$30 < g \leq 60$	*30*	*90*	3
$60 < g \leq 70$	10	20	*2*
$70 < g \leq 100$	*30*	*30*	1

b) Accurate histogram. Teachers should take care to check that:

- The y-axis is labelled 'Frequency density' with scale from 0-4.
- The $x-$axis is labelled 'Weight (g)' with scale from 0-100.

- The first bar has an area of 80 (starting at 0 and ending at 20, with a height of 4).
- The second bar has an area of 20 (starting at 20 and ending at 30, with a height of 2).
- The third bar has an area of 90 (starting at 30 and ending at 60, with a height of 3).
- The fourth bar has an area of 20 (starting at 60 and ending at 70, with a height of 2).
- The fifth bar has an area of 30 (starting at 70 and ending at 100, with a height of 1).
- There are no gaps between consecutive bars.

Appreciate the importance of using a continuous scale on the x-axis when constructing histograms

Example 12:

a) Ben is correct. Explanations may vary but should indicate students know there should be no gaps between the bars as the histogram is representing continuous data.

b) Ben's first bar has a class width of 11 (not 10); the height is incorrect; $12 \div 11 = 1.09$ (to 2 d.p.). He's used a class width of 10, to give a frequency density of 1.2.

c) Students' responses may vary, but should demonstrate an understanding that, in the context of age, values are truncated so all ages up to 11 would be in the first group. In the context of weight being rounded to the nearest kilogram, any value from 10.5 up to 11 would now be included in the second group, not the first. The bars should therefore be shifted left by 0.5 squares, with the first bar on the histogram now going from 0 to 10.5, the second from 10.5 to 30.5 and so on. The class width has only changed for the first bar, so the heights of all but the first bar would remain the same. The first bar would now have a frequency density of $12 \div 10.5 = 1.14$ (to 2 d.p.).

d) Now that inequalities are being used, the bars will shift again so that the first goes from 0 to 10, the second from 10 to 30 and so on. As in part c, the class width has only changed for the first bar – it now has the frequency density that Ben originally calculated, as $12 \div 10 = 1.2$.

10.2.1.8 Interpret and use features of data from a histogram

Understand what the heights of bars in a histogram represent

Example 1:

C: 15. This is because frequency is represented by the area of the bar in a histogram, and so we need to multiply the class width (15) by the frequency density (1).

Example 2:

a) 30

b) This is because the sum of the areas of the bars is 30 $(10 \times 0.5 + 16 \times 0.5 + 24 \times 0.25 + 18 \times 0.5 + 8 \times 0.25) = (5 + 8 + 6 + 9 + 2)$.

Recognise which measures of central tendency can be found from a histogram

Example 3:

Responses may vary, but should demonstrate an understanding that the modal class is $30 < x \leq 40$. The histogram looks fairly symmetrical, so the mean and median values must be similar and lie within the modal class.

Example 4:

B: The mean is greater than the median.

Students' responses may vary, but should demonstrate an understanding that the frequencies are greater for the lower ranges of scores, meaning that there are more values below the median than above the median. This has the effect of 'dragging down' the mean.

Appreciate the effect of the mean on a histogram

Example 5:

70

Students' explanations may vary, but should demonstrate an understanding that the two distributions are identical, and so the profile of marks for 10W4 has just been shifted along the x-axis by 40. The mean is therefore greater than 30 by 40 marks.

Recognise the insight a histogram provides on the distribution of a data set

Example 6:

Responses may vary, but should demonstrate an understanding that the mean summarises the data as a single numerical statistic but tells us nothing about the shape of the distribution: two very different profiles of scores can 'average' to the same single value. The histograms, by contrast, show the spread of the data.

10.2.3.2 Understand that correlation alone does not indicate causation

Understand the difference between correlation and causation

Example 1:

B: Correlation but not causation.

Explanations may vary, but should demonstrate an understanding that the pattern of points suggests there is a positive correlation: as a baby's height increases, so too does its weight. For causation to be present, an increase in a baby's height would result in an increase in its weight. However, there are many factors that can influence someone's weight, so it does not imply causation.

Example 2:

a) Positive correlation, but no causation. Both the number of ice creams and the number of air-conditioning units sold are likely to increase as the temperature increases, but one does not cause the other.

b) No correlation or causation.

c) This depends on whether the person is paid per hour or a fixed salary. If they are paid an hourly rate, there would be positive correlation and causation. If they are paid a fixed salary, there would be no correlation or causation.

d) Negative correlation and causation. The faster the train, the less time it will take over the distance.

e) This depends on whether a discount is offered for bulk purchases. As a general rule, there is not necessarily correlation or causation. However, there is the possibility of negative correlation and causation if purchasing in bulk reduces the cost per item.

Understand that a causal variable may be present and be able to identify it

Example 3:

Responses may vary, but should demonstrate an understanding that Graph B is the only graph which suggests correlation only and not causation.

10.3.1.1 Understand that relative frequencies tend towards theoretical probabilities as sample size increases

Appreciate the difference between theoretical and experimental probabilities

Example 1:

a) Baz

b) Responses may vary but should demonstrate an understanding that Ali assumed the results of an experiment will mirror theoretical probabilities.

Understand that relative frequency can be used as an estimate of probability when theoretical probability is unknown

Example 2:

a) P(even number) $= \frac{51}{100} = 51\%$
Responses may vary, but should demonstrate an understanding that, given they want the chance of an even number to be at least 50%, they should use the weighted dice.

b) Responses may vary, but should demonstrate an understanding that by rolling the dice 200 times, they have increased the reliability of the probability estimation.

Example 3:

a) 15

b) 104

c) Responses may vary, but should demonstrate an understanding that 96.9% were on time here. However, a sample of a single month is not sufficient to draw conclusions about reliability.

d) Yes. Responses may vary but should demonstrate an understanding that these results produce a 99% reliability rate when rounded to the nearest per cent, based on a reasonable sample.

Example 4:

Responses may vary but should demonstrate an understanding that Em's conclusion is not reliable. If the coin were fair, it would theoretically land on heads 50% of the time. However, Em's sample size of just 10 trials is too small to draw a reliable conclusion. Larger sample sizes provide results that are more representative of the true probability: Fi flipped the coin 20 times and observed 35% heads, Gerry observed 34% heads after 50 flips, and Hafizah observed 29% heads after 200 flips. It is therefore likely that the coin is biased, given that the probabilities tend to a percentage that is significantly below 50% as the number of trials increases.

Example 5:

a) Responses may vary but should demonstrate an understanding that, although 80% of Dean's pack germinated, a sample size of only five is too small to draw a reliable conclusion.

b) 90% (or equivalent)

c) The answer from part b, because the sample size is greater.

d) 230

e) 1112

Example 6:

a) Five, if it was a fair coin.

b) Responses may vary, but should demonstrate an understanding that 10 trials are too small a sample size to make a reliable conclusion.

c) Responses may vary, but should demonstrate an understanding that this would be more convincing as evidence of bias because 100 trials is significantly more than 10, making it more reliable.

10.3.2.2 Understand how to choose and use a representation appropriate to the given situation

Understand the mathematical structure of a Venn diagram

Example 1:

a) Responses may vary but should demonstrate an understanding that both Venn diagrams are sorting the shapes from the pattern. Abigail's diagram sorts the actual shapes, while Boki's sorts the quantities of each shape.

b) Responses may vary but should demonstrate an understanding that Boki's Venn diagram is more useful. The two-way table has 'total' rows and columns, so it requires numerical input. Abigail's Venn diagram displays the types of shapes in each section but doesn't indicate the quantity of each.

	Right angles	***No right angles***	***Total***
Is a rhombus	9	6	15
Is not a rhombus	4	2	6
Total	13	8	21

Example 2:

a) Responses will vary, given the open nature of the question. Students with a more superficial understanding may just compare the values in the corresponding parts of the Venn diagram. For example, noticing that the number of AI systems built just by academics decreased from 18 to three between 2014 and 2024, while the numbers in every other part of the Venn increased. Students with a more developed understanding may not only comment on the raw quantities, but also begin to notice the relative proportions that are explored in parts c and d. For example, the overall number of AI systems being built has increased from 34 to 84, but in 2014 over half of all artificial intelligence systems were built by individuals working in academia, and by 2024 the vast majority were developed by those working in industry.

b) Including those who also worked in academia, 13 AI systems in 2014 were built with involvement from someone in industry. By 2024, this number had risen to 76, which is an increase of 63.

c) In 2014, the proportion of AI systems built solely by academics was $\frac{18}{32} = \frac{9}{16}$ (or equivalent). By 2024 the proportion had reduced to $\frac{3}{84} = \frac{1}{28}$ (or equivalent).

d) Responses may vary, but should demonstrate an understanding that while Rana is correct, her statement does not fully capture how AI design has changed as she is only comparing raw quantities, not proportions, in relation to the total number of AI systems. Although there was only one additional AI system built in collaboration, there has been a significant overall increase in the

total number of systems being developed, and so the proportion of systems built in collaboration between academia and industry has, in fact, significantly reduced.

Example 3:

a) Responses may vary, but should demonstrate an understanding that, despite their different structures, both diagrams represent the same information about the herd. Eleanor's (Euler) diagram uses all four possible descriptions of the cows (both possible breeds, and both possible numbers of calves born) to arrange the four values in her diagram into four separate sets. Arthur's (Venn) diagram uses just two categories for the sets (one breed, and one number of calves born), so that the same four values are arranged according to whether they are or are not in those two sets.

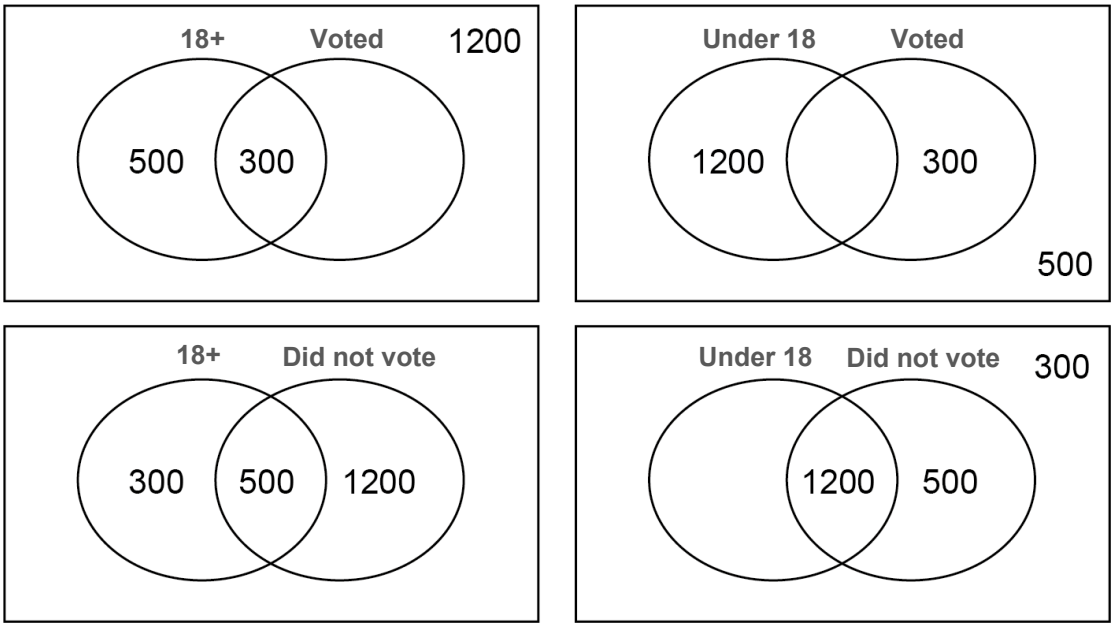
b) Either 'Single calf' and 'Twin calves' **or** 'Friesian' and 'Jersey'. Neither of these pairs of labels would produce a true Venn diagram.

c) Responses may vary, but should demonstrate an understanding that Arthur could have labelled the sets as 'Twin calves' and 'Jersey'. This would change the location of the numbers, and once again create a true Venn diagram.

d) Responses may vary, but should demonstrate an understanding that mutually-exclusive events cannot both be true at the same time, as explored in part b. 'Single calf' and 'Twin calves' are mutually exclusive, since an animal cannot be in both categories and therefore there would be a zero in their intersection. In contrast, being a twin calf and a Jersey breed are not mutually exclusive. These events can occur together and, according to the two-way table, they did on two occasions, so there would therefore be a '2' in their intersection.

Example 4:

a)



b) Responses may vary, but should demonstrate an understanding that all the diagrams show the same information in different ways. Since under-18s cannot vote, there is an empty set in three of the four diagrams and so, arguably, the bottom-left Venn diagram is the most useful representation of this data. Neither set nor the intersection of the sets is empty, as it is, the events are mutually-inclusive events.

Example 5:

a) Responses may vary, but should demonstrate an understanding that the three diagrams show different ways to sort the students in Year 10 into categories. The first diagram represents the number of students in Year 10, and the intersection shows how many of those are in Form 10B. The second diagram shows the number of students aged 14 and 15 in the year, with an empty intersection as students cannot be both ages at once. The final Venn diagram represents how many Year 10 students are studying Spanish GCSE, Music GCSE, both subjects or neither subject.

b) 175

c) Any responses that could satisfy the three Venn diagrams. For example:

- First Venn diagram: any two categories where the entirety of set B is included in set A. For example, 'Is a student at Vicci's school' and 'Is in Year 8'.
- Second Venn diagram: any two categories that are mutually exclusive. For example, 'Passed the exam' and 'Failed the exam'.
- Third Venn diagram: any two categories that are not mutually exclusive, and that there may be an exception to. For example, 'Plays in the school orchestra' and 'Goes to netball club'.

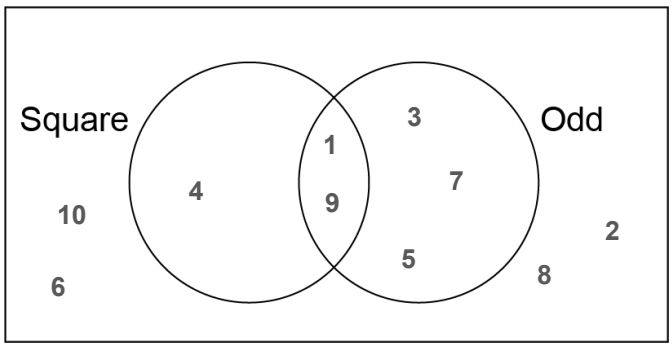
d) The six in the intersection between 'Music GCSE' and 'Age 14' is incorrect. It should be seven, as the values in the Music GCSE set need to sum to 25, and the numbers in the 'Age 14' set should total 102.

e) Questions (i) and (iv) can be now answered.

Begin to use set notation for Venn diagrams

Example 6:

a)



b) Responses may vary depending on the stage students are at in their learning around set notation.

The correct answers are:

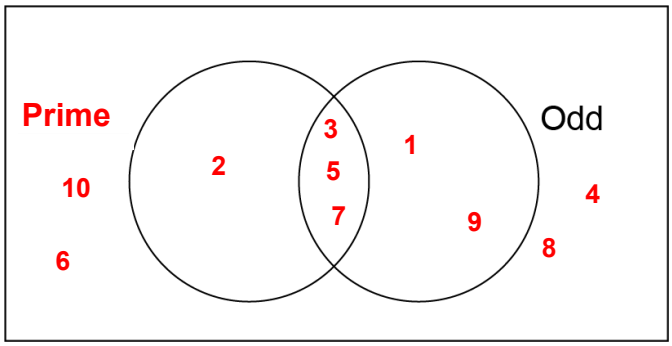
$A \cup B$: **union** of Sets A and B, i.e., in Set A **or** Set B **or** both

$A \cap B$: **intersection** of Sets A and B, i.e., in Set A **and** Set B

A': **complement** of Set A, i.e., **not** in Set A; B': **complement** of Set B, i.e., **not** in Set B

c) (i) $(\mathrm{A} \cap \mathrm{B})' = \{2, 3, 4, 5, 6, 7, 8, 10\}$ (ii) $(A \cup B)' = \{2, 6, 8, 10\}$ (iii) $\mathrm{A}' \cap \mathrm{B} = \{3, 5, 7\}$

d)



e) (i) $A \cup B = \{1, 2, 3, 5, 7, 9\}$ (ii) $A \cap B = \{3, 5, 7\}$ (iii) $A' = \{1, 4, 6, 8, 9, 10\}$

(iv) $B' = \{2, 4, 6, 8, 10\}$ (v) $(A \cap B)' = \{1, 2, 4, 6, 8, 9, 10\}$ (vi) $(A \cup B)' = \{4, 6, 8, 10\}$

(vii) $A' \cap B = \{1, 9\}$

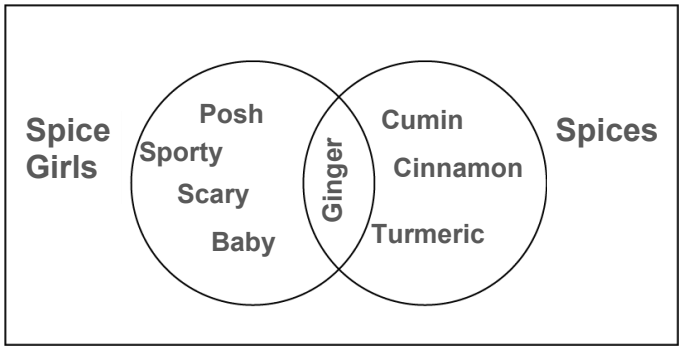
Example 7:

a) $A \cap B = \{Ginger\}$ and *'Spice Girls who are also spices'*

b) Responses may vary, but should demonstrate an understanding of the notation. Example statements are given in the right-hand column of the table below.

Statement	Possible matching statement
Cumin $\in A'$	'Cumin is not a Spice Girl'
Set A = 'members of Spice Girls'	$A =$*{*Baby, Ginger, Scary, Sporty, Posh*}*
$A \cup B$ = *{Baby, Ginger, Scary, Sporty, Posh, Turmeric, Cumin, Cinnamon}*	"Spice Girls and/or spices"
Posh $\in A$	'Posh is a member of Spice Girls'
$B' =$*{Baby, Scary, Posh, Sporty}*	'Spice Girls who aren't spices'
'Turmeric is a member of the spices'	Turmeric $\in B$

c)



d) Students may suggest a variety of responses, and so teachers should accept any correct statements that have not already been given. For example:

Baby $\in B'$

Baby is not a spice

$A \cup B'$ = {Baby, Ginger, Scary, Sporty, Posh}

Understand the relationship between frequencies and probabilities on tree diagrams

Example 8:

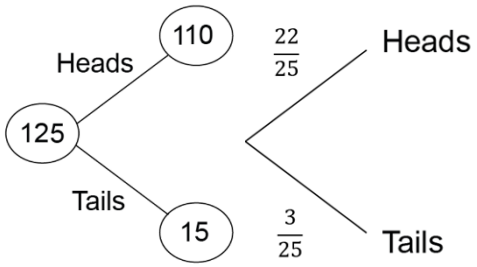
a) 120. Explanations may vary, but should demonstrate an understanding that the starting value in a frequency tree diagram represents the total frequency.

b) Responses may vary, but should demonstrate an understanding that the fractions on the probability tree have been derived from the frequencies on the frequency tree. The frequency tree shows the coin landed on heads 105 times, having been flipped a total of 120 times. These values can be used as numerator and denominator, giving a relative frequency of $\frac{105}{120} = \frac{7}{8}$. The coin landed on tails 15 times, giving a relative frequency of $\frac{15}{120} = \frac{1}{8}$.

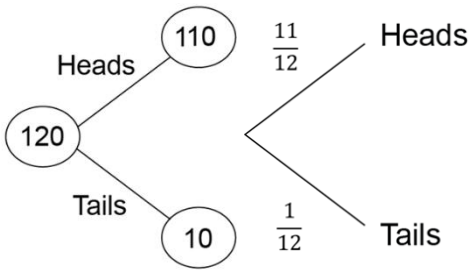
c) Responses may vary, but should demonstrate an understanding that the two fractions on each set of branches in a probability tree diagram sum to 1. This is because they represent all possible outcomes at that stage, and the total probability must equal 1.

d) There are two valid approaches to this question, which are explained in more detail both in the guidance and the diagrams below. Regardless of the approach taken, students should demonstrate an understanding of the difference between frequency tree diagrams and probability tree diagrams. In a frequency tree diagram, the total can change as the values on the branches change. For the probability tree diagram, while the fractions on the branches might change, the total will not (i.e., the branches should still sum to one).

The approach of adding the extra five heads to the total frequency should lead the following tree diagrams (or equivalent):



The approach of reducing the number of tails by five to accommodate the extra five heads within the total frequency should lead the following tree diagrams (or equivalent):



Example 9:

Yes, there is enough information to complete the frequency tree. Explanations for how this is known may vary, as students may use the information given in a different order. An example of a valid approach is:

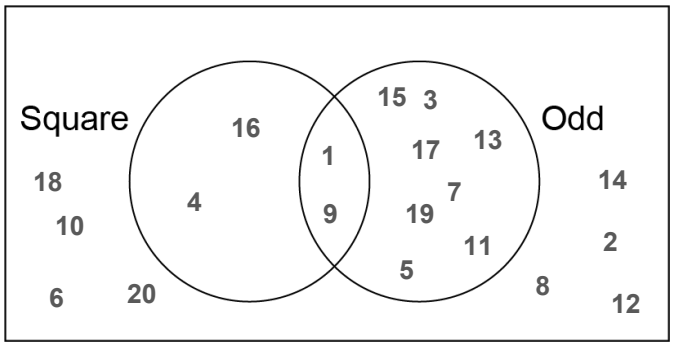
- The starting value is 250, and since 215 people eat meat, we can subtract 215 from 250 to find that there are 35 people who do not eat meat.
- Of those, three are dairy free, so subtracting three from 35 gives us 32 people who don't eat meat but do eat dairy.
- We are told there are 38 people in total who are dairy free, so subtracting three from 38 reveals that 35 of the dairy-free guests must eat meat.
- Since 215 people eat meat, and 35 of them are dairy free, the subtracting 35 from the remaining 215 tells us that 180 people eat both meat and dairy.

Use different representations to calculate probabilities of independent events

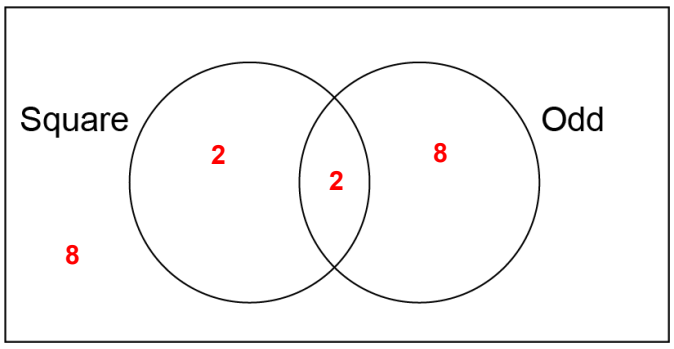
Example 10:

$\frac{2}{20}\left(=\frac{1}{10}\right)$ or equivalent.

Students may have arranged the numbers 1 to 20 in the Venn diagram in order to support them to find this probability, as shown below:

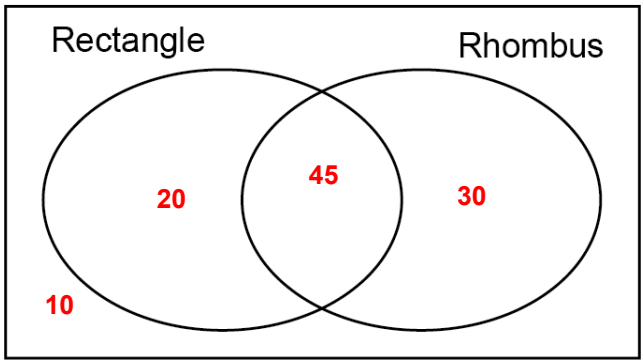


Alternatively, students may have used the Venn diagram to record the frequencies of each type of number, as shown below:



Example 11:

a)



b) (i) $\frac{75}{105}\left(=\frac{5}{7}\right)$ (ii) $\frac{65}{105}\left(=\frac{13}{21}\right)$ (iii) $\frac{45}{105}\left(=\frac{3}{7}\right)$ (iv) $\frac{30}{105}\left(=\frac{2}{7}\right)$ (v) 1

(vi) 0 (vii) 1 (viii) $\frac{30}{105}\left(=\frac{2}{7}\right)$ (ix) $\frac{10}{105}\left(=\frac{2}{21}\right)$

Example 12:

a) Responses may vary but should demonstrate an understanding that:

- Kalil incorrectly calculated the total by summing all the values in his table to get 42. In doing so, he double counted the five people who like both tea and coffee. To correct this, he should subtract five from both the tea (15) and coffee (13) totals before adding them.
- Lee correctly represented the five people who like both tea and coffee in the intersection of the Venn diagram. He correctly calculated the total number surveyed by adding the union (those

who like tea or coffee or both) to the complement (those who like neither), giving a correct total of 32.

b) The total number of students that completed the survey.

Example 13:

a) Responses may vary, but should demonstrate an understanding that just because there are two possible overall outcomes, this does not mean that each outcome is equally likely. In this case, there are four different combinations that can create the two outcomes of 'odd' and 'even'. For the product of two numbers to be odd, both numbers must be odd. This means that there are three possible combinations that produce even numbers, and only one combination that produces an odd number.

b) Using the probability tree diagram: $P(odd \times odd) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} = 25\%$.

Using the sample space diagram: of the 36 results, 9 are odd, so $\frac{9}{36} = \frac{1}{4} = 25\%$.

Example 14:

a) The probability tree diagram just represents the data for girls, using the first column of the two-way table. The first pair of branches represents the probabilities of whether girls are gamers (58%) or non-gamers (42%); 58% is the total of non-problematic gaming girls (53%) and problematic gaming girls (5%). The second pair of branches represents the probabilities of whether the gamers are non-problematic gamers (91%) or at risk of problematic gaming (9%).

b) Responses may vary, but should demonstrate an understanding that the values on the second set of branches are a sub-set of the values from the table. The probabilities in the tree diagram are calculated from the 58% of girls who game, not the 100% of girls in the survey. As 53% of the girls in the survey were non-problematic gamers, they represent 91% of the girls who are gamers with 9% similarly being calculated from the girls at risk of problematic gaming as a proportion of gamers that are girls ($91\% = \frac{53\%}{58\%}$ and $9\% = \frac{5\%}{58\%}$).

c) We already know that the probability of a random 15-year-old girl being a gamer with unproblematic habits is 53% from the two-way table. To show this from the probability tree, we need to identify that the probability of being a gamer is represented by the top branch of the first set of branches, and having unproblematic habits is represented by the top branch on the second set of branches. Multiplying these probabilities gives: $58\% \times 91\% = 53\%$.

d) Responses may vary, but should demonstrate an understanding that Chiamaka has incorrectly added all the probabilities in the tree diagram. However, each set of branches represents a different event (the first, the probability of a girl being a gamer; and the second, the probability of a female gamer having unproblematic habits). This means that she has accidentally added the probabilities of all of the outcomes for two events. Only the probabilities on branches from the same point (i.e. the same event) should sum to 1.

e) There are multiple possible probability trees that can be generated from the two-way table. Teachers should attend to the order in which students place events, to check that probabilities are calculated correctly. In this example, a 50/50 split of girls and boys among the population of 15-year-olds has been used, but teachers may wish to clarify that this is an assumption.

A sample of the *labels* for three potential tree diagrams, and their relevant probabilities, is given on the next page. In these examples, we have continued to group the gaming population so that there are two branches, but students also could create tree diagrams with three branches per event (i.e., non-gamer, non-problematic gamer, and gamer at risk of problematic gaming).

Summary	*Outcomes for first event*	*Outcomes for second event*	*Outcomes for third event*
Using boys' data only, probability of being a gamer followed by probability of problematic gaming habits.	Gamer (89%) Non-gamer (11%)	Non-problematic gamer (83%) At risk of problematic gaming (17%)	n/a
Using all the data, probability of girl/boy, followed by probability of being a gamer, followed by probability of problematic gaming habits.	Girl (50%) Boy (50%)	Girl, gamer (58%) Girl, non-gamer (42%) Boy, gamer (89%) Boy, non-gamer (11%)	Girl, gamer, non-problematic (91%) Girl, gamer, problematic (9%) Boy, gamer, non-problematic (83%) Boy, gamer, problematic (17%)
Using all the data, the probability of being a gamer, followed by probability of girl/boy.	Gamer (74%) Non-gamer (26%)	Gamer, girl (39%) Gamer, boy (61%) Non-gamer, girl (79%) Non-gamer, boy (21%)	n/a

Example 15:

a) The first pair of branches should both be labelled 0.5, since there is a 50% chance of Pat having or not having the gene.

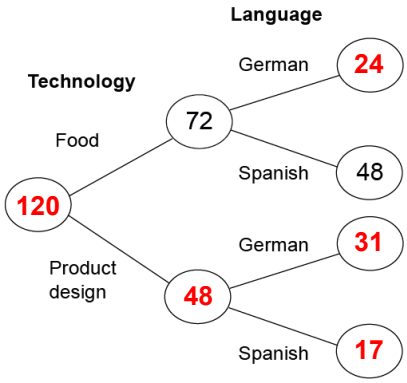
b) P(disease) = P(gene ∩ disease) + P(no gene ∩ disease) = (0.5 × 0.75) + (0.5 × 0.13) = 0.44 = 44%

c) The branch for 'does not have gene' should be labelled with 0.998 to represent the 99.8% chance of not inheriting the gene for the general population. The branch for 'has gene' should be labelled with 0.002 to represent the 0.2% chance of inheriting the gene for the general population.

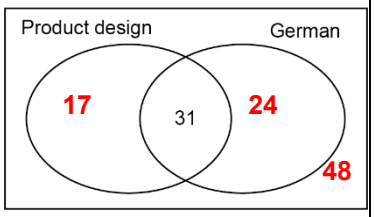
d) P(disease) = P(gene ∩ disease) + P(no gene ∩ disease) = (0.002 × 0.75) + (0.998 × 0.13) = 0.13124 = 13%

Example 16:

a)



	Food	PD	Total
Ger	24	31	55
Span	48	17	65
Total	72	48	120



b) (i) P(Food) = $\frac{72}{120}\left(=\frac{3}{5}\right)$ (ii) P(Spanish) = $\frac{65}{120}\left(=\frac{13}{24}\right)$ (iii) P(Spanish and Food) = $\frac{48}{120}\left(=\frac{3}{5}\right)$

c) (i) P(German student also studied PD) = $\frac{31}{55}$ (ii) P(PD student also studied German) = $\frac{31}{48}$

(iii) P(Spanish student also studied Food) = $\frac{48}{65}$ (iv) P(Food student also studied Spanish) = $\frac{48}{72}\left(=\frac{2}{3}\right)$

d) Individual responses will vary, but teachers should notice whether students have an awareness that some values that are visible from the two-way table will require further calculation or identification from the probability tree or Venn diagram. For example, the total number of Spanish students.

10.3.3.2 Understand how the calculation of probabilities of combined events is affected by dependence/independence

Know how to distinguish between independent and dependent events

Example 1:

A: Independent B: Dependent C: Independent D: Independent

E: Independent F: Independent G: Independent

The answers given above reflect answers that are appropriate for students at this stage in their understanding of determining the probabilities of real-world events. However, it could be argued that there is more complexity to some of these answers. For example, seasonal weather patterns affecting the answer to C; roadworks and traffic-light programming affecting the answers to E and F; or the cumulative psychological effect of missed penalties affecting the answer to G. Teacher may wish to explore this further with their students if they demonstrate that they are ready to consider these issues when discussing modelling probabilities of real-life situations.

Understand the effects on probabilities of replacement/non-replacement

Example 2:

a) Salote's probability tree diagram.

b) $\text{P(Hard-boiled then chewy)} = \frac{5}{9}\times\frac{4}{8}=\frac{20}{72}$ $\text{P(Chewy then hard-boiled)} = \frac{4}{9}\times\frac{5}{8}=\frac{20}{72}$

Example 3:

There is a variety of valid responses. Some suggestions include:

a) The probability of selecting a specific colour (e.g., red).

b) The probability of selecting a specific suit (e.g., diamond).

c) The probability of selecting a specific value (e.g., ace).

d) The probability of selecting two consecutive cards of the same specific value, when the first card is not replaced (e.g., selecting seven of hearts followed by seven of clubs).

e) The probability of selecting a specific suit, replacing the card, and then selecting a specific suit (e.g., selecting a spade, putting it back, then selecting a diamond. Note that the suit could be the also be the same, e.g., two spades).

f) The probability of selecting consecutive cards of a specific suit, without replacement (e.g., two clubs in a row).

g) The probability of selecting consecutive cards of the same specific colour, without replacement (e.g., two blacks in a row).

h) The probability of selecting a specific value, replacing it, and then selecting a DIFFERENT card of the same value (e.g., ace of spades, putting it back, and then selecting the ace of hearts).

i) The probability of selecting a specific card, replacing the card, and then selecting a specific card (e.g., selecting the two of hearts, putting it back, and then selecting the nine of hearts. Note that the card could also be the same, e.g., two of hearts twice.)

Example 4:

a) $\frac{1}{120}$

b) (i) $\frac{45}{120}\left(=\frac{3}{8}\right)$

(ii) $\frac{1}{45}$

c) (i) $\frac{15}{120}\left(=\frac{1}{8}\right)$

(ii) $\frac{1}{105}$

d) $\frac{50}{120}\left(=\frac{5}{12}\right)$

Calculate the probabilities of combined events

Example 5:

a) Responses may vary but should demonstrate an understanding that:

Same	***Different***
Both scenarios have the same probabilities for the first selection. In both, a second selection is possible regardless of the outcome of the first.	In scenario A, the probabilities remain the same for both selections, they are independent. In scenario B, the probabilities change after the first selection; the second selection is dependent on the outcome of the first.

b) Any two valid, mutually-exclusive categories where two of the numbers 1 to 5 fit in one category, and three of them fit into the other. Examples might include 'Odd' and 'Even; 'Prime and 'Square; 'Includes curves when written as a digit' and 'Does not include curves with written as a digit'; 'More than three letters long when written as a word' and 'Exactly three letters long when written as a word'.

c) Scenario A as P(prime, prime) = 0.6 × 0.6 = 0.36 (in Scenario B P(prime, prime) = 0.6 × 0.5 = 0.33)

d) Scenario B as P(prime, square) + P(square, prime) = (0.6 × 0.5) + (0.4 × 0.75) = 0.3 + 0.3 = 0.6 (in Scenario A P(prime, square) + P(square, prime) = (0.6 × 0.4) + (0.4 × 0.6) = 0.24 + 0.24 = 0.48)

e) No, the probabilities would not be the same. Explanations may vary, but should demonstrate an understanding that since there is only one prime and one square number between six and ten, the probability of selecting either of these types of number would be 20%. As there are three other numbers that do not fit into either category, if Bev wanted to use the probability tree to represent the situation, she would need to include another branch to ensure all possible outcomes are shown and the probabilities sum to 1.

Example 6:

No, the game is not fair. Explanations may vary but should demonstrate an understanding that Nel has an advantage as there are more ways of getting two different types of fruit than there are of getting two of the same types of fruit. The guidance column has some suggestions of representations to demonstrate this.

Example 7:

a) The detail of students' labels for each may vary but should capture the meaning as detailed in the table below:

	Day 1 shirt choice		***Day 2 shirt choice***			
Probability	$\frac{n}{5}$	$1-\frac{n}{5}$	$\frac{n-1}{4}$	$1-\frac{n-1}{4}$	$\frac{n}{4}$	$1-\frac{n}{4}$
Label	Logo	No logo	Logo	No logo	Logo	No logo

b) Responses may vary, but should demonstrate an understanding that on the first day, the probability he picks a shirt with a logo is $\frac{n}{5}$. After choosing one logo shirt, there are now $n-1$ logo shirts left out of four remaining shirts. So, the probability that he picks another logo shirt the next day is $\frac{n-1}{4}$. Multiply these probabilities to get the probability of two consecutive logo shirts: $\frac{n}{5} \times \frac{n-1}{4} = \frac{n(n-1)}{5\times 4} = \frac{n^2-n}{20}$.

c) Responses may vary, but should demonstrate an understanding that as the shirt is replaced, the probability of choosing a shirt with a logo on is $\frac{n}{5}$ for both days, and so the probability of two consecutive shirts with logos increases. P(logo, logo) $= \frac{n}{5} \times \frac{n}{5} = \frac{n^2}{25}$. The exception is if $n = 5$; if all of Charlie's school shirts had logos, then the probability would not change.

d) Probability tree diagram with the same probabilities on all pairs of branches: $\frac{n}{5}$ for the 'Logo' branch and $1-\frac{n}{5}$ for the 'No logo' branch.

e) $\left(1-\frac{n}{5}\right)\left(1-\frac{n}{4}\right)=0.3$

$1^2-\frac{n}{5}-\frac{n}{4}+\frac{n^2}{20}=\frac{3}{10}$

$1-\frac{4n}{20}-\frac{5n}{20}+\frac{n^2}{20}=\frac{6}{20}$

$20-9n+n^2=6$

$n^2-9n+14=0$

$(n-7)(n-2)=0$

$\therefore n=7 \text{ or } n=2$

Since Charlie has a maximum of five shirts, the only viable solution is two.

So, **two** of Charlie's shirts do not have a logo, and **three** do.

Example 8:

a) Brown eyes = 79 Blue eyes = 21 Right-handed = 71 Left-handed = 29

b) $a+b=79$ $a+c=71$ $c+d=21$ $b+d=29$ $b+c=44$

$a=53$ $b=26$ $c=18$ $d=3$

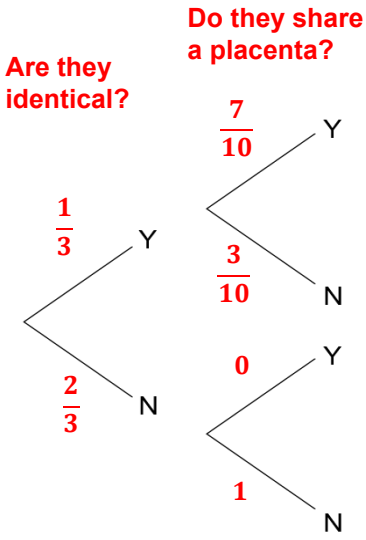
$P(A\cap B)=\frac{18}{100}$

10.3.3.4 Find and use expected frequencies from Venn diagrams, two-way tables and tree diagrams

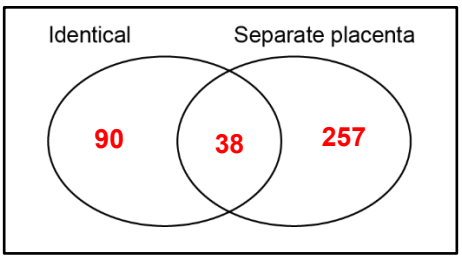
Understand the relationship between probability and expected frequency.

Example 1:

a)



b)



c) This (23) is more than you would expect. Explanations may vary but should demonstrate an understanding the probability of a randomly-selected twin set being identical with a shared placenta is $\frac{1}{3}\times\frac{7}{10}=\frac{7}{30}$, and $\frac{7}{30}\times 82\approx 19$.

Example 2:

a)

i) $\frac{19\,451 + 3\,329}{1\,389\,604}$

ii) $\frac{175\,649}{1\,389\,604}$

iii) $\frac{115\,316 + 3\,426 + 29\,187}{1\,389\,604}$

b)

i) $\frac{19\,451 + 3\,329}{19\,451 + 29\,187 + 3\,329 + 2\,310}$

ii) $\frac{2\,310}{19\,451 + 29\,187 + 3\,329 + 2\,310}$

iii) $\frac{29\,187}{19\,451 + 29\,187 + 3\,329 + 2\,310}$

c)

i) $\frac{19\,451 + 3\,329}{115\,316 + 19\,451 + 3\,329 + 2\,630}$

ii) $\frac{2\,630 + 3\,329}{115\,316 + 19\,451 + 3\,329 + 2\,630}$

iii) $\frac{115\,316}{115\,316 + 19\,451 + 3\,329 + 2\,630}$

d) Responses may vary, but should demonstrate an understanding that the numerator stays the same but the denominator changes depending on whose patients we're looking at, which affects the probability for each set of patients.

Use different representations to work with expected frequencies

Example 3:

a) Responses will vary depending on number of students:

i) $\mathbf{0.35} \times$ class size

ii) $\mathbf{0.35} \times$ number of students in year group

iii) $\mathbf{0.35} \times$ number of students in school

b) $\mathbf{0.4 \times 5}$

c) Responses may vary, but should demonstrate an understanding that Marta's estimate is based on random probability across the general population. Genetics introduces dependence, meaning family members are not independent samples. Therefore, expected frequency models are less reliable in families when traits are hereditary.

Example 4:

a) 10 or 11

b) Emilia is not correct. Explanations may vary, but should demonstrate an understanding that each birth is an independent event, so each child of a parent who carries the gene has a 50% chance of inheriting it. The outcome for one sibling does not affect the outcome for another.

c) (i) 69 000

(ii) 17 250

(iii) 4 527 780

d) Between 1 000 and 1 500.

Example 5:

Responses may vary but should demonstrate an understanding that:

a) Nessa is assuming the games are equally likely to result in a win just because they had the same number of winners in five minutes. However, she hasn't considered how many people played each game. If more people played one of the games, then the chance of winning per person might be lower for that game.

b) The tombola would be the recommended game. The probability of winning on the hook-a-duck is $\frac{6}{42}$. The probability of winning on the tombola is $\frac{6}{24}$. Since $\frac{6}{42} < \frac{6}{24}$, there is a greater probability of winning on the tombola.

c) P(winning hook-a-duck) = $\frac{6}{42} = \frac{1}{7}$

P(losing hook-a-duck) = $\mathbf{1} - \frac{1}{7} = \frac{6}{7}$

P(losing two) = $\frac{6}{7} \times \frac{6}{7} = \frac{36}{49}$

P(winning at least one) = $\mathbf{1} - \frac{36}{49} = \frac{13}{49} \approx \mathbf{0.265}$

This does change the answer to part b – just! The probability of winning on the tombola was 0.25. But, as the tombola is twice as expensive, Nessa would be able to play hook-a-duck twice. The probability of winning at least once from two games of hook-a-duck is 0.265 (3dp), which is greater than 0.25.

Example 6:

a) Responses may vary, but should demonstrate an understanding that the guests are not exactly representative but are reasonably close at 14%.

b) 15.2%

Work with conditional probability.

Example 7:

a) i) $\frac{13}{52}\left(=\frac{1}{4}\right)$ ii) $\frac{4}{52}\left(=\frac{1}{13}\right)$ iii) $\frac{1}{52}$ iv) $\frac{13}{52}\left(=\frac{1}{4}\right)$

b) i) $\frac{13}{26}\left(=\frac{1}{2}\right)$ ii) $\frac{2}{26}\left(=\frac{1}{13}\right)$ iii) $\frac{1}{26}$ iv) 0

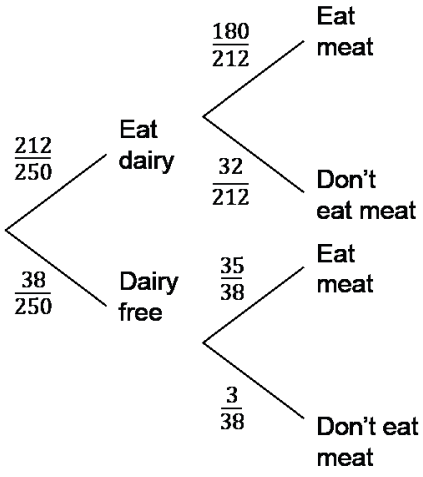
c) i) 1 ii) $\frac{1}{13}$ iii) $\frac{1}{13}$ iv) 0

d) Students' responses will vary, so teachers will need to check the question sets carefully and ensure that they have the same probabilities as parts b and c.

Example 8:

a) Petra is correct. Explanations may vary, but should demonstrate an understanding that we have already been told that it is a guest who doesn't eat meat, and so we are finding the probability from this smaller sub-group of the guest population. Olly has answered it based on finding the probability that someone doesn't eat meat and is dairy free from the whole population of guests.

b)



c) (i) The new tree diagram

(ii) The original tree diagram

(iii) The new tree diagram

(iv) The original tree diagram