# 10 Statistics and probability

Mastery Professional Development

## 10.2 Statistical representations and analysis

Guidance document | Key Stage 4



*Click the heading to move to that page. Please note that these materials are principally for professional development purposes. Unlike a textbook scheme they are not designed to be directly lifted and used as teaching materials. The materials can support teachers to develop their subject and pedagogical knowledge and so help to improve mathematics teaching in combination with other high-quality resources, such as textbooks.*

© Crown copyright 2026

## Making connections

Building on the Key Stage 3 mastery professional development materials, the NCETM has identified a set of five 'mathematical themes' within Key Stage 4 mathematics that bring together a group of 'core concepts'.

The fourth of the Key Stage 4 themes (the tenth of the themes in the suite of Secondary Mastery Materials) is *Statistics and probability*, which covers the following interconnected core concepts:

10.1 Statistical measures and analysis

**10.2 Statistical representations and analysis**

10.3 Probability

This guidance document breaks down core concept *10.2 Statistical representations and analysis* into three statements of **knowledge, skills and understanding**:

10.2 Statistical representations and analysis

- 10.2.1 Accurately construct statistical representations for univariate data
- 10.2.2 Accurately construct statistical representations for bivariate data
- 10.2.3 Interpret statistical measures and representations for bivariate data

Then, for each of these statements of knowledge, skills and understanding we offer a set of **key ideas** to help guide teacher planning:

10.2.1 Accurately construct statistical representations for univariate data

- 10.2.1.1 Understand that graphical representations of data can support comparison between data sets and identification of trends over time
- 10.2.1.2 Construct box plots for given data sets
- 10.2.1.3 Interpret features of data from a box plot and use them to make comparisons between data sets
- 10.2.1.4 Understand the idea of accumulation and why it is useful
- 10.2.1.5 Construct cumulative frequency graphs and use to estimate information about the data
- 10.2.1.6 Interpret and use features of data from a cumulative frequency graph
- 10.2.1.7 Construct histograms for a given data set
- 10.2.1.8 Interpret and use features of data from a histogram

10.2.2 Accurately construct statistical representations for bivariate data

- 10.2.2.1 Understand paired data as data points which are in relation to each other
- 10.2.2.2 Accurately construct a scatter graph

10.2.2.3 Interpret features of data from a scatter graph

10.2.3 Interpret statistical measures and representations for bivariate data

10.2.3.1 Recognise types of correlation

10.2.3.2 Understand that correlation alone does not indicate causation

10.2.3.3 Construct an estimated line of best fit with an understanding that this represents the trend of the data

10.2.3.4 Recognise outliers and make informed decisions about their relationship with the general trend of the data

10.2.3.5 Interpolate and extrapolate apparent trends in data sets

10.2.3.6 Understand the limitations of interpolation and extrapolation for data sets

## Overview

**This core concept builds on previously constructed representations for univariate data at Key Stage 3, with the introduction of further statistical representations that support the interpretation and comparison of features of data sets. Students' familiaritywith scatter graphs as a way of representing bivariate data is developed to focus more extensively on the identification of data trends, how to use this information to identify outliers and how to interpret them in relation to the general trend.**

Students will be familiar with producing graphical representations of statistical data from their work in earlier key stages. This is developed in Key Stage 4 with the introduction of histograms, box plots and cumulative frequency graphs. A key understanding that needs to be established and developed is **how** a representation communicates frequency. In particular, students will learn to recognise the difference between a bar chart, where the height of the bar represents the frequency of that particular category, and a histogram, where the area of a bar gives the frequency. The introduction of the cumulative frequency graph at Key Stage 4 further challenges students as they need to understand how the shape of a cumulative frequency curve – often an 's' shape – relates to the distribution of the data.

A key focus at Key Stage 4 is exploring representations of summary statistics. Understanding how the shape of the cumulative frequency curve reflects the characteristics of the data helps students to appreciate what the graph can tell us about how the data represented is distributed within the range. A cumulative frequency graph also provides an efficient way of estimating the median and interquartile range for large data sets, with a box plot being a fairly straightforward visual representation of the same information. It is important that the five summary values required to produce a box plot are determined correctly. However, students need to go beyond determining these values and creating the plots, to understanding them as a graphical display of the variance of a data set. This includes knowing what the features of a box plot represent; how they can be used to analyse a data set; and how they can be used to compare multiple data sets. While the exact distribution of data cannot be identified from a box plot, the position of the median and the comparable length of the whiskers provide insight into the extent to which the data are skewed.

Students also deepen their understanding of representations of bivariate data at Key Stage 4. It is important that they can interpret a scatter graph, by looking for trends in the data going from left to right, so that they can consider the extent to which two variables are related. Students develop their ability to recognise different types and strengths of correlation and, in making summary statements about the data, are encouraged to distinguish between the presence of correlation and causation. The introduction of lines of best fit to represent the trend of the data provides an opportunity to identify outliers and how they relate to the general trend of the data. A line of best fit can be used to predict values both within the given data

and outside the plotted points, by way of extrapolation. Students must recognise that interpolation will often provide a valid estimate of an unknown value, but it may not be precise. Extrapolation is even less reliable, as we are assuming that the observed trend of the data continues for values outside of the range used to form the line of best fit approximation model.

The knowledge, skills and understanding that students develop within this core concept are essential to prepare them to think critically about any data that they engage with. This goes beyond the mathematics classroom and the school environment: students will see representations of data throughout their lives, such as through traditional and social media, and this will include potentially misleading uses of statistics. It is important that students are able to interpret such representations effectively, critiquing and challenging where necessary, so that they can draw informed conclusions about the statistics they encounter.

## Prior learning

Students will have collected data about their environment, and been encouraged to organise and display it, from their earliest experiences of mathematics. More formal representations for data, such as bar charts, pie charts and pictograms, are introduced at Key Stage 2. Students should have been encouraged to think about when a particular representation should be used and what each one communicates about the data. For example, they should have been supported to understand the difference between representations that show the proportion of the whole data set in each category (such as pie charts), and representations that show the actual frequency in each category (such as bar charts). This is an understanding that should be checked and challenged, particularly if students' experience of statistics has been relatively spread out across the Key Stage 3 curriculum.

As well as working with univariate data, students will have previously explored bivariate data and, at Key Stage 3, scatter graphs will have been introduced as a representation. It is important to note that they are likely to have used scatter graphs to connect variables in other subjects such as science and geography, and so they may bring knowledge that has been taught outside the mathematics classroom. It is advisable to discuss this with colleagues from other subject teams, to gain a sense of how scatter graphs are introduced and used in different curriculum contexts, and the rationale behind any differences.

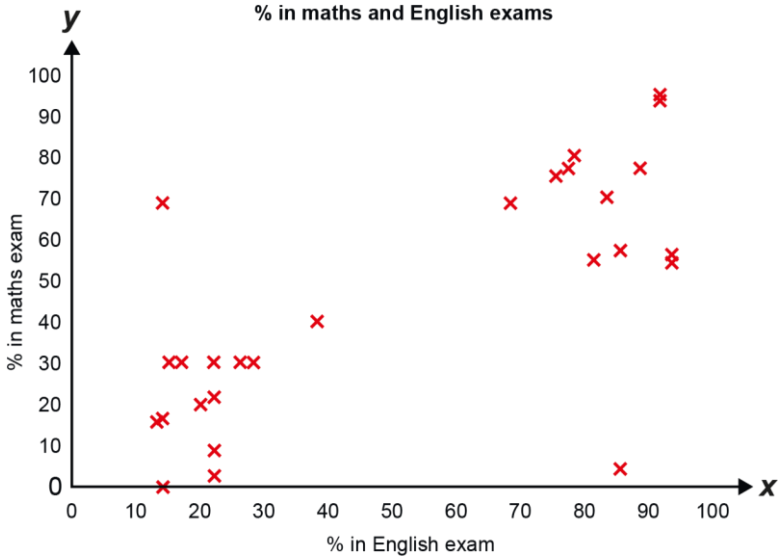
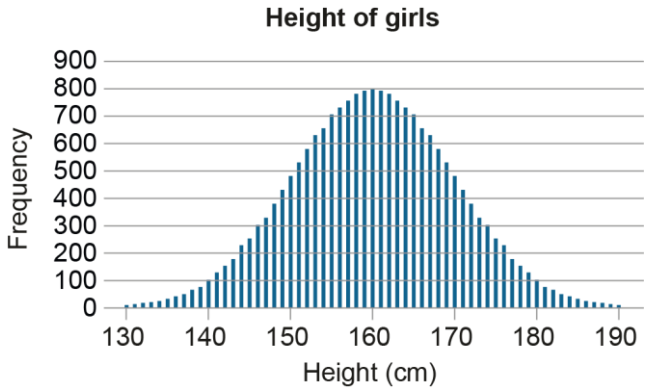
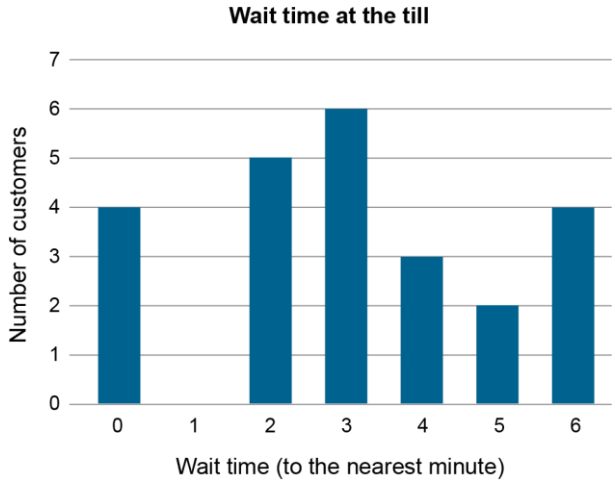
At Key Stage 3, students began to develop their ability to make informed choices about appropriate statistical tools to use. The introduction of box plots, cumulative frequency graphs and histograms at Key Stage 4 extends the choice of statistical representation to include ways of representing discrete, continuous and grouped data. Students should be encouraged to explain, justify and critique their choice of representations and interpret the data in terms of statistical measures and analysis, allowing for comparison of data sets and the identification of statistical trends.

The first two core concept documents in theme 5, '*5.1 Statistical representations and measures'* and '*5.2 Statistical analysis'* from the Key Stage 3 PD materials explore in depth the prior knowledge required for this core concept.

## Checking prior learning

The following activities from the NCETM secondary assessment materials, Checkpoints and/or Key Stage 3 PD materials offer a sample of useful ideas for assessment, which you can use in your classes to check understanding of prior learning.

| Reference | Activity |
|---|---|
| Secondary assessment materials page 52 | Justine is exploring the hypothesis 'students who are good at maths are also good at English.'<br>She gathers data and draws a scatter graph (shown below).<br>What does this graph show? |

| | |
|---|---|
| | Does it help Justine to see if her hypothesis is correct?<br>Are there any results that you'd like to explore further or any other questions that the scatter graph raises for you?<br>  |
| Key Stage 3 PD materials document '*5.2 Statistical analysis'*, Key idea 5.2.2.2, Example 3 | What would be the best average to choose to represent the height of a 'typical' girl?<br>  |
| Key Stage 3 PD materials document '*5.1 Statistical representations and measures'*, Key idea 5.1.1.1, Example 9 | This bar chart shows the length of time people spent waiting at a self-service till (to the nearest minute). What is the mean wait time?<br>  |

# Key vocabulary

**Key terms used in Key Stage 3 materials**

- Bar chart
- Bivariate
- Continuous data
- Discrete data
- Correlation
- Outlier
- Scatter graph

The NCETM's mathematics glossary for teachers in Key Stages 1 to 3 can be found here.

**Key terms introduced in the Key Stage 4 materials**

| Term | Explanation |
|---|---|
| box plot | A graphical display of the median, quartiles and extremes of a data set, on a number line, to show the distribution of the data. |
| cumulative frequency graph | A graph for displaying cumulative frequency by plotting the running total of frequencies against the upper class limit of the respective group.<br>At a given point on the horizontal axis, the sum of the frequencies of all the values up to that point is represented by a point whose vertical coordinate is proportional to the sum. |
| frequency density | The frequency per unit for the data within a given group. The frequency density is calculated by dividing the frequency of a class by the class width. |
| histogram | A particular form of representation of grouped data. Segments along the x- axis are proportional to the class interval. Rectangles are drawn with the line segments as bases. The area of the rectangle is proportional to the frequency in the class.<br>Where the class intervals are not equal, the height of each rectangle is called the frequency density of the class. |
| univariate data | Data that consists of just one variable. |

## Key ideas

In the following list of the key ideas for this core concept, selected key ideas are marked with a  . These key ideas are expanded and exemplified in the next section – click the symbol to be taken direct to the relevant exemplifications. Within these exemplifications, we explain some of the common difficulties and misconceptions, provide examples of possible pupil tasks and teaching approaches and offer prompts to support professional development and collaborative planning.

### 10.2.1 Accurately construct statistical representations for univariate data

A key focus of statistical representations at Key Stage 3 is the construction of graphs for categorical data. While students have had some experience of representing ungrouped and grouped data graphically, new statistical representations for univariate data are introduced at Key Stage 4.

The similarities between histograms and bar charts provide a familiarity for students, but can equally cause confusion, as students mix up the properties of the two statistical representations. The height of a bar chart represents the frequency, but the height of a histogram represents the frequency density. Students need to understand this difference, and it should not be assumed that they will be able to infer the meaning of 'density' in this context. Time should therefore be spent reinforcing that the **area** of a histogram represents the frequency. Histograms are most commonly used to represent continuous data that have been grouped, and these groupings can be unequal if it is appropriate for the data set. Students, with their prior experience of bar charts, may instinctively see this as 'wrong', and so need to understand why data grouped into unequal class widths can be more suitable than data in equal classes. It can be helpful to model 'what it isn't' in this case, by showing students the disproportionate heights of bar charts with unequal class widths.

Box plots are an entirely new representation introduced at Key Stage 4, and it is likely to be the first time students have represented summary statistics rather than the whole data set. Representing data in a box plot can help students to have a visual sense of the range and more easily distinguish it from a measure of central tendency. While a box plot features the median value, the other measures of central tendency are not identifiable. Students should recognise that box plots provide a graphical representation that can be used to identify the variation in a data set, including signs of skew and potential outliers.

Students might begin by constructing box plots from given values but must also learn to generate these values from other representations, such as cumulative frequency tables and graphs, which are also new learning at Key Stage 4. When using a cumulative frequency table to identify the five key values needed for a box plot, students need a secure understanding of accumulation. This is where the sum of all the previous frequencies up to the current point are identified, which helps us to obtain insight into how much of the data are below, or greater than, a particular value. When we consider the data set as a whole, and not just as varying frequencies of distinct values, we are also able to identify the $\frac{(n+1)}{4}$th, $\frac{(n+1)}{2}$th and $3\frac{(n+1)}{4}$th terms more easily. These values provide an insight into both measures of central tendency and spread.

Plotting accumulated data on a graph provides a way of displaying this cumulative information graphically, and therefore of comparing two or more data sets. The general shape of the cumulative frequency curves can then easily be compared for different data sets, as well as values of particular interest being identified and compared. There is no limit to the number of values that can be estimated from the curve, but it is common to read off the values required to construct a box plot. Cumulative frequency curves provide an accessible way of determining these key data points when presented with large, grouped data sets. They offer opportunities to make comparisons between different statistical representations of the same data and understand how they are related.

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

10.2.1.2 Construct box plots for given data sets

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

10.2.1.4 Understand the idea of accumulation and why it is useful

10.2.1.5 Construct cumulative frequency graphs and use to estimate information about the data

10.2.1.6 Interpret and use features of data from a cumulative frequency graph

10.2.1.7 Construct histograms for a given data set

10.2.1.8 Interpret and use features of data from a histogram

### 10.2.2 Accurately construct statistical representations for bivariate data

Students should be aware that scatter graphs are a way of representing bivariate data, and have had experience at Key Stage 3 of illustrating data involving two variables using this type of graphical representation. When plotting bivariate data, it is important that students understand the difference between independent and dependent variables. This will ensure that the independent variable (the variable that isn't affected by something else) is plotted on the $x$-axis and the dependent variable (that is affected by the independent variable) is plotted on the $y$-axis. As well as mastering the accurate construction of scatter graphs, students must develop their ability to use scatter graphs to investigate the relationship between two variables. They may already be familiar with drawing conclusions from scatter graphs about the way in which two variables are related. For example, identifying what happens to the dependent variable as the independent variable increases. At Key Stage 4, students' interpretation of scatter graphs is developed further, to describing more precisely the features of the data that the scatter graph represents.

10.2.2.1 Understand paired data as data points which are in relation to each other

10.2.2.2 Accurately construct a scatter graph

10.2.2.3 Interpret features of data from a scatter graph

### 10.2.3 Interpret statistical measures and representations for bivariate data

Students may have explored the patterns between two variables when represented on a scatter graph at Key Stage 3. They may have some understanding of the difference between positive and negative correlation, and how a lack of correlation is identifiable from the plotted points not following a pattern and looking haphazard. Students often assume that when correlation between variables is present, the change in one variable is the cause of the change in the values of the other variable. It is important that they understand the difference between correlation and causation. Correlation is where there is an association between two variables. Causation is where one event is the result of the occurrence of the other event. Correlation does not necessarily imply causation.

When correlation occurs, it is possible to construct an estimated line of best fit that represents the trend of the data. Students often assume that the line of best fit must go through at least one of the plotted points. This may have its basis in work with bivariate data in science; with experimental data, there is an assumption that readings are correct and the line should therefore be drawn to pass through as many points of possible. This does not apply in many other situations, such as height and mass, as there is not a 'correct' pair of values. Students should be clear that different curriculum areas will ask them to draw a 'line of best fit' differently, and be supported to understand the rationale behind each approach.

This misconception can be especially present when working with positive correlation, as students often assume that the line of best fit must pass through the origin. They need to recognise that the line of best fit should represent the directional trend of the data, by minimising the distances between all the points and the line; and that the closer the points are to the line of best fit, the stronger the correlation between the two variables is. When constructing a line of best fit, the presence of outliers can become evident, if one or more data points do not fit the pattern of the rest of the data and lie much further away from the line than the other plotted points. Identifying a potential outlier, and how it relates to the general trend of the data, is key to recognising how predictions made from the line may be less accurate due to its presence, as well as understanding the extent to which such predictions might be affected.

The line of best fit can be used to make predictions for values that were not included in the data. The accuracy and reliability of these predictions will depend on the strength of the correlation between the two variables and the presence of any outliers. The line of best fit can be used for values that are within the plotted points (interpolating) or outside the plotted points (extrapolating). It is important for students to recognise that while we have a greater likelihood of obtaining a valid prediction from interpolation, unless the data points lie in a straight line, the predictions will have limited accuracy, unless we can be sure there is an exact linear relationship between the variables (e.g., speed and time taken to travel a given distance). With extrapolation, the limitations in terms of precision are even greater, as we are having to assume that the trends and patterns in the data points extend beyond the known data set.

10.2.3.1 Recognise types of correlation

10.2.3.2 Understand that correlation alone does not indicate causation

10.2.3.3 Construct an estimated line of best fit with an understanding that this represents the trend of the data

10.2.3.4 Recognise outliers and make informed decisions about their relationship with the general trend of the data

10.2.3.5 Interpolate and extrapolate apparent trends in data sets

10.2.3.6 Understand the limitations of interpolation and extrapolation for data sets

## Exemplified key ideas

In this section, we exemplify the common difficulties and misconceptions that students might have and include elements of what teaching for mastery may look like. We provide examples of possible student tasks and teaching approaches (in italics in the left column), together with ideas and prompts to support professional development and collaborative planning (in the right column).

The thinking behind each example is made explicit, with particular attention drawn to:

| | |
|---|---|
| **Deepening** | How this example might be used for **deepening** all students' understanding of the structure of the mathematics. |
| **Language** | Suggestions for how considered use of **language** can help students to understand the structure of the mathematics. |
| **Representations** | Suggestions for key **representation(s)** that support students in developing conceptual understanding as well as procedural fluency. |
| **Variation** | How **variation** in an example draws students' attention to the key ideas, helping them to appreciate the important mathematical structures and relationships. |

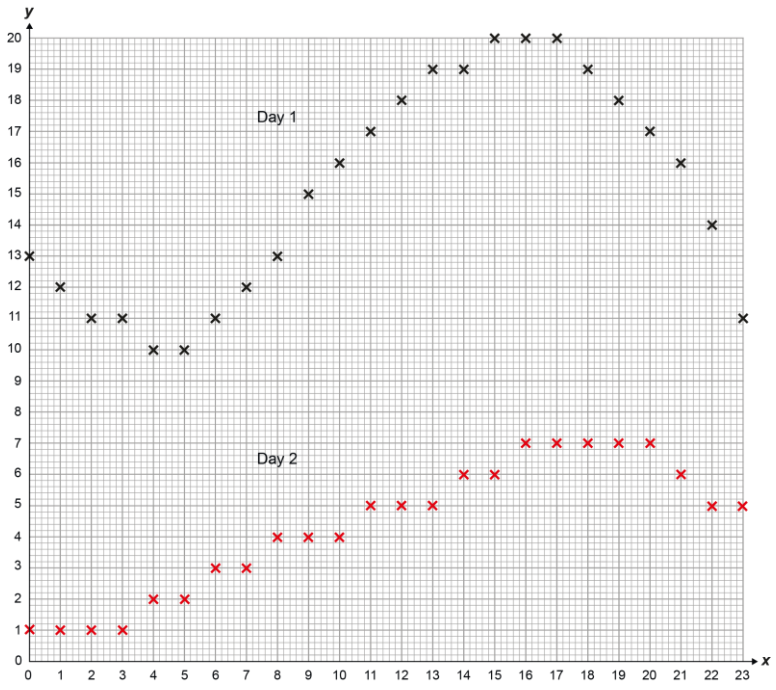
In addition, questions and prompts that may be used to support a professional development session are included for some examples within each exemplified key idea.

These are indicated by this symbol.

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

| Common difficulties and misconceptions |
|---|
| When describing features of graphical representations, students often focus solely on a comparison of statistical measures, sometimes failing to consider what these measures mean in terms of the context the data represent. This 'within the context' interpretation is key to recognising the usefulness of statistical measures and representations in understanding data trends. It also provides opportunities to explore of possible explanations as to why a data set might have the characteristics described. Exploring changes in data sets over time (for example, different times of day, week, year, etc.) is a useful way of providing a context that can be easily grasped and offers an opportunity for students to relate what the data show to the real world. |

| Students need to | Guidance, discussion points and prompts |
|---|---|
| **Interpret graphical representations in real-world contexts**<br><br>*Example 1:*<br><br>*At a UK weather station, the outside temperatures are recorded hourly on two different days of the year.*<br><br>*The data are represented on the scatter graph presented below this example.*<br><br>a) *What is the same and what is different about the temperatures on the two days?*<br><br>b) *At what time of year might the two temperature recordings have taken place? Explain how you can tell.* | *Example 1* focuses on trends of temperature in different seasons of the year and provides an opportunity for students to make comparisons between two data sets within the same **representation**. Students should be able to comment on the difference in the range of temperatures for the two days, as well as recognise that the temperatures are a lot warmer on one day than the other. (This is indicative of the two days being recorded in spring/summer and autumn/winter). The trend of the data for both days is similar, in that there is a rise in temperature, which peaks in the afternoon and drops into the evening. However, the temperatures on the warmer day drop initially, before rising to a maximum temperature.<br><br>As well as interpreting what the data actually tells us, there is potential for **deepening** students' thinking by asking about the implications of the data:<br><br>• *'How are the temperatures at the end of the day different to the temperatures at the start of the day for the two days? What might this suggest about the temperatures on the days that followed?'*<br><br>• *'Can we tell anything about the temperatures the day before the two days represented? Why or why not?'*<br><br>  Students must become comfortable with providing possible explanations for trends in data that cannot necessarily be proven, but that are sensible within the context of the data set. Discuss how you might build students' confidence and willingness to suggest possible implications from the data, when it may not be possible to prove them as being right or wrong. |



*Example 2:*

*A household's daily gas usage is summarised for the months of January, April, July and October, presented below this example.*

a) *What was the coldest month in 2000? Explain how you know.*

b) *What is the same and what is different about the gas usage in April and July? Explain why this might be.*

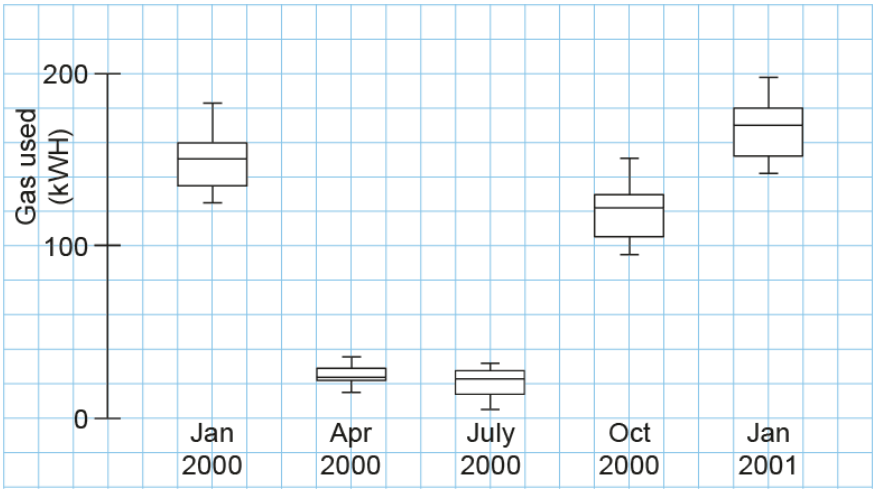
c) *What can you say about the weather in January 2001 compared to January 2000?*

In *Example 2*, students consider the effects of seasonal weather on a household's gas consumption. It is important that students work with real-world data, but it is important that teachers are aware that **language** that is familiar for adults may be a barrier for students. Ensure they understand that the main gas usage is to fuel the house's heating system and so the amount of gas used is likely to be higher when outside temperatures are colder.

The summary information used to construct the box plot **representations** is not easily identifiable for some of the plots. Instead of relying on the accurate comparison of measures of central tendency and spread, encourage students to think about the overall trend of the data and explain what this reveals in the given context. For example, while the top five per cent of the data for April and July is comparable (so it is likely there were some days where the temperatures were also comparable) the box plots suggest that there were days in July when the amount of gas used was lower than any of the daily gas amounts recorded in April.

Students should be encouraged to have confidence to make sensible assertions within the context, without the need to prove their accuracy. They might assert, for example, that 'January 2001 was a colder month than January 2000', while being aware that other factors might have affected gas usage, meaning that this cannot be determined explicitly from the representation.

The summary information for the box plots is given beneath. This is not so that it can be revealed to students: the focus is to evaluate the overall shape of the data, rather than make a detailed comparison of statistical measures. It is important, however, that students can interpret box plots in relation to the summary statistics and relate this back to the context. Discuss how you might use the actual summary measures to support them, using prompts and questions that move beyond direct comparisons and towards trends over time.



Summary information for box plots (for teachers' information):

| | Min | LQ | Med | UQ | Max |
|---|---|---|---|---|---|
| **Jan 2000** | 125 | 135 | 152 | 160 | 183 |
| **Apr 2000** | 15 | 22 | 23 | 29 | 36 |
| **Jul 2000** | 5 | 14 | 23 | 28 | 32 |
| **Oct 2000** | 95 | 105 | 122 | 130 | 153 |
| **Jan 2001** | 142 | 152 | 170 | 180 | 198 |

*Example 3:*

*A local shop is open between 8am and 8pm, seven days a week. The owner records the number of till transactions made during each opening hour on one weekday and one day over the weekend, to compare the flow of customers.*

*The results are plotted on the cumulative frequency graph presented below this example.*

a) *Determine which curve represents a weekday and which represents a weekend day.*

b) *Explain how you know, referring to features of the cumulative frequency curves.*

*The owner recorded this information to help him decide if he needed to change his opening hours on either weekdays or weekends.*

c) *What would you recommend the owner does? Make reference to the cumulative frequency curves in your answer.*

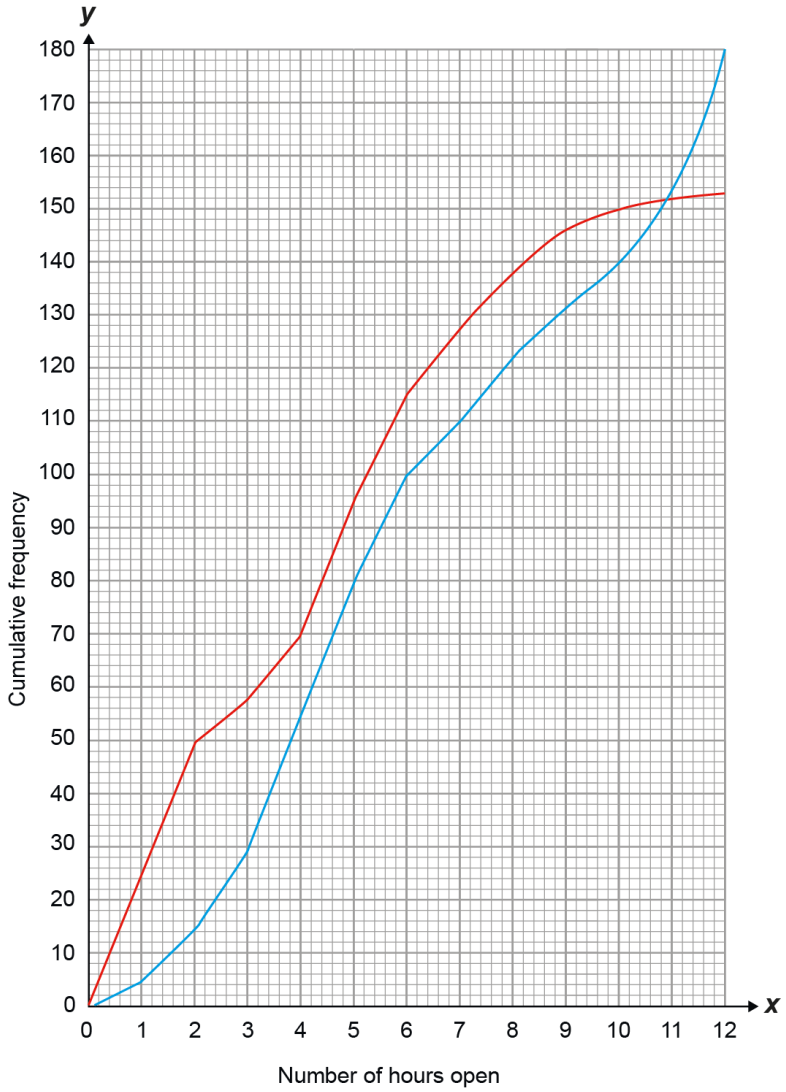
*Example 3* focuses on the comparison of two cumulative frequency curves. Students consider the effect the day of the week may have on the flow of customers through a corner shop. If students have already explored box plots, it is likely that they might initially identify and compare the median and quartiles. While these values provide some insight into the distribution of customers throughout the day, it is important that this example is used for **deepening** understanding of how to interpret the shapes of the curves and recognise the information they provide about the busiest and quietest times for the two days.

Students need to recognise that the data in this **representation** may not be indicative of the trends in transactions for different times of the week and evaluate these accordingly. Hypothesising about what might be expected, then determining whether the graphs confirm this, is an important part of the statistical analysis process. Encourage students to provide a possible reason why the data may not represent the expected trends. They should recognise that there is often a number of factors that affect the data, and the implications that these will have.

Discuss possible prompts to support students with making comparisons. When comparing the two curves, it is helpful to recognise that when the number of customers is greater than in the previous hour, the curve will concave upwards. When the curve concaves downwards, the number of customers in the shop was less than during the previous hour. Teachers may use the prompts below:

- *'How can we tell that the shop was busier at the start of the day during the week compared with at the weekend? Why might this be?'*
- *'Why might the shop be busier at the end of the day at the weekend compared with a weekday? How can you tell this from the graph?'*
- *'What do you notice about the total number of customers who visited the shop on the two days? Is this what you would expect? Explain your answer.'*

- *'Is the number of customers sufficient information to decide about whether it is worth keeping the shop open? What other data might the shopkeeper usefully collect?'*




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

**Common difficulties and misconceptions**

Box plots provide a visual representation of five specific statistical details and are a useful way of comparing two or more data sets, but students often misunderstand them. As the mean is a more commonly used measure of central tendency than the median, students often assume that the middle line represents the mean rather than the median. The box plot only directly shows the median. However, it is possible to estimate whether the mean is less than or greater than the median, by looking at whether the box plot has positive or negative skew. It is important that students appreciate this as part of their broader understanding of how a box plot describes the underlying distribution of a data set. Other common misconceptions when interpreting box plots are that the size of the box is representative

of the number of data values, rather than the distribution of the data, and that the whiskers do not represent any data values other than the minimum and maximum. Providing opportunities for students to explore features of box plots alongside the (raw) data can help to address these misconceptions, supporting students to deeply appreciate the ways in which box plots describe a data set's distribution.

| Students need to | Guidance, discussion points and prompts |
|---|---|

**Interpret the key values in a box plot**

*Example 1:*

*There are 90 pupils and 10 staff at a primary school. Their ages are shown, in order, in the hundred grid below. The lower half of the data is shaded grey.*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 25 | 28 | 34 | 39 | 39 | 41 | 42 | 46 | 50 | 52 |
| 10 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 |
| 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| 9 | 9 | 9 | 9 | 10 | 10 | 10 | 10 | 10 | 10 |
| 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 |
| 8 | 8 | 8 | 8 | 8 | 8 | 9 | 9 | 9 | 9 |
| 7 | 7 | 7 | 7 | 7 | 7 | 7 | 8 | 8 | 8 |
| 6 | 6 | 6 | 6 | 6 | 7 | 7 | 7 | 7 | 7 |
| 5 | 5 | 5 | 5 | 5 | 5 | 6 | 6 | 6 | 6 |
| 4 | 4 | 4 | 4 | 4 | 5 | 5 | 5 | 5 | 5 |

a) *What is the:*

   *(i) lowest and highest age*

   *(ii) value that is half-way*

   *(iii) half-way value in the upper half of the data*

   *(iv) half-way value in the lower half of the data?*

b) *Draw a scale from 0-60 and mark on it the five data points you have found.*

*The grid is changed so that only the adults and children from one class are shown:*

| | | | | |
|---|---|---|---|---|
| 7 | 7 | 7 | 34 | 52 |
| 7 | 7 | 7 | 7 | 7 |
| 6 | 7 | 7 | 7 | 7 |
| 6 | 6 | 6 | 6 | 6 |
| 5 | 5 | 6 | 6 | 6 |
| 5 | 5 | 5 | 5 | 5 |

The first few examples in this key idea are designed to support with conceptual understanding of box plots, so that students are not just mechanically finding the five key values. Here, the familiar **representation** of a hundred grid is used to help students visualise the people behind a data set, and to imagine them divided into two equal-sized groups. Teachers might like to first explore the representation, with prompts such as:

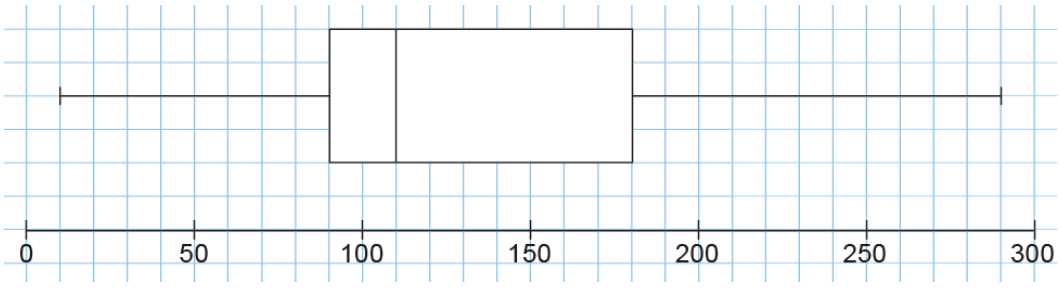
- *'Where can you see the ages of the adults?'*
- *'How big do you think each class might be? Are the classes all the same size?'*
- '*Is there a single value that is half-way?'*

Compare and contrast the **language** of parts a and c. Identical structures have been used, but with colloquial language in part a and specific mathematical terms in part c. This is designed to support students to develop a deep understanding of concept of quartiles, by highlighting the connections between other terminology that they might encounter or have already used. The terms lower and upper quartile are not yet used, but students are asked to think about the median of the two halves of the data set, reinforcing the idea that the quartiles split the data into equally sized groups.

Part a offers some opportunity for **deepening** understanding of defining 'half-way' in data sets with an even number of values. There is no single value in the 'middle' of either the whole data set, or each half of the data. For parts (ii) and (iii) this is unambiguous as the two values either side of the midpoint are identical. Students are likely to instinctively chose 6.5 for their answer for part (iv); check that they can articulate this as the midpoint and understand that they could also describe this as the mean of the two values either side.

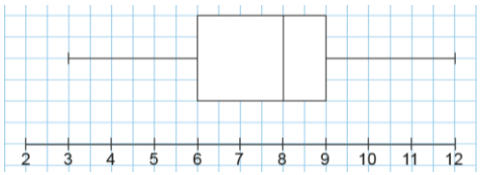
Parts b and d ask students to mark the five data points onto scale, but does not specify the way in which students do this. This ambiguity is deliberate to offer an opening for teachers to then introduce the box plot representation. Discuss with colleauges what the next steps might be in the classroom.

| | |
|---|---|
| c) *What is the:*<br>*(i) minimum and maximum age*<br>*(ii) median of the whole data set*<br>*(iii) median of the upper half of the data*<br>*(iv) median of the lower half of the data?*<br>d) *Draw a scale from 0-60 and mark on it the five data points you have found.*<br>e) *Compare your answers. What is the same and what is different?* | |
| *Example 2:*<br>*Annie is a researcher investigating the impact of smartphones on mental health. She wants to know how many times, on average, people pick up their phone in a day. She carries out a survey of 200 people and shows the results of her survey in the box plot presented below this example.*<br>*How many people answered that they picked up their phone:*<br>a) *between 10 and 90 times per day?*<br>b) *between 90 and 180 times per day?*<br>c) *between 10 and 180 times per day?*<br>d) *between 180 and 290 times per day?*<br>e) *up to 110 times per day?* | In *Example 2*, students are offered a box plot and asked to quantify each part of the **representation**. Students need to attend to the information about frequency in the question, and not confuse this with the visible scale (which refers to the number of times the phone has been picked up per day). The scale is deliberately unlabelled so that teachers can use the opportunity to discuss the different 'frequencies' referred to in this example – the 'frequency' of people alongside the 'frequency' of phone pick-ups.<br>The **variation** is designed to draw attention to what each of the key values represents, and how the data are split into equally-sized groups. If students are finding this challenging, vary this by asking the same questions but changing the number of people in the sample. To reinforce this point further, change the box plot but keep the number of people in the survey constant, to demonstrate that the frequency represented by each whisker and each section of the box will always be 25 per cent of the total frequency, regardless of the scale.<br>The context in this example is likely to be very familiar to students, but the data, while based on news reports, are made up. This offers opportunity for rich discussion both with colleagues and students. Do they think that the data set is realistic? How does it compare with their own smartphone usage? How might the data change depending on the age of those surveyed? |



*Example 3:*

*The conductor of a youth orchestra wants to know how long each player has been learning their instrument. She represents the data in a box plot:*




*Some of the younger players have never seen a box plot before, so she constructs the two bar charts shown below.*

a) *What is the significance of the different patterns and shades?*

b) *How does each representation show:*
   - *(i) the quartiles*
   - *(ii) the middle 50% of the data*
   - *(iii) the top 25% of the data*
   - *(iv) the median of the data?*

c) *Which representation do you find most useful for understanding the shape of the data? Why?*

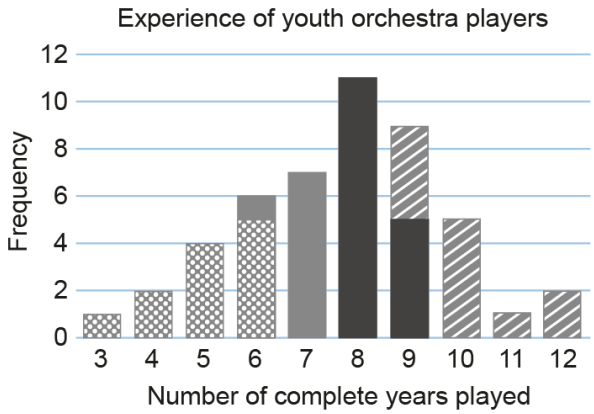
While useful as a **representation** for summary statistics, box plots lack a concrete sense of the frequencies between each quartile. *Example 3* provides another way for students to conceptualise how the data points are distributed in a box plot. In the first bar chart, the parts of the bars that can be attributed to the values in the 'box' are shaded dark grey and parts of the bars that relate to the values in the whiskers are dotted. The second bar chart follows the same principles, but differentiates four different groups to demonstrate the values that lie between each of the quartiles. Visualisations such as this can help students to make sense of how a box plot works and the actual data that sit behind it.

There are numerous ways that teachers could use this task as a springboard for further **deepening** students' understanding. For example, ask students to:

- Create other bar charts that would also be compatible with the same box plot.
- Explain how each representation would change if a particular data value changed.
- Explain how each representation would change if a partiular data value was added or removed.
- Investigate whether the median always lies within the modal group.

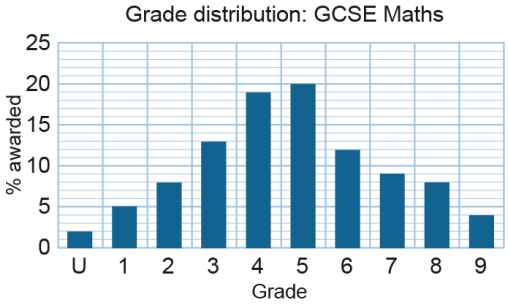
*Bar chart A:*

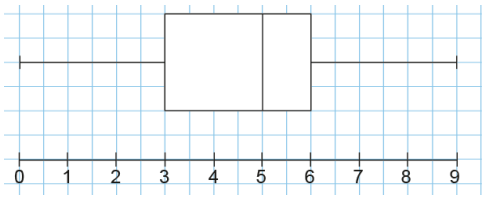

*Bar chart B:*



*Example 4:*

*The bar chart and box plot both show how GCSE Maths grades are distributed.*




*Below this example are the graphs for GCSE Engineering and GSCE Italian.*

a) *What is the same and what is different about the graphs for the three subjects? What does this tell you about the grade distribution?*

b) *What conclusions might you be able to draw about why the subjects' grade distributions are so different?*

c) *Sketch your predictions of the box plots for these two distributions.*

d) *Using the table of quartiles below, construct accurate box plots for GCSE Engineering and GCSE Italian.*

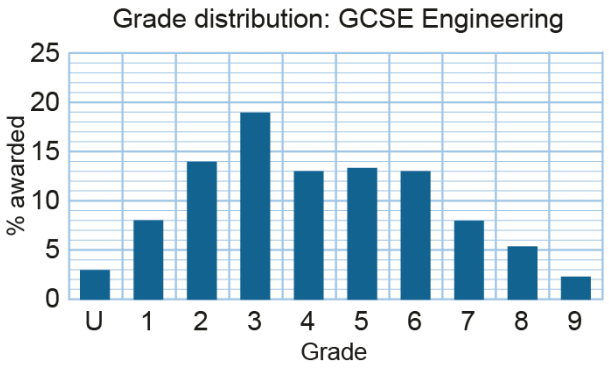
| | ***Min*** | ***LQ*** | ***Med*** | ***UQ*** | ***Max*** |
|---|---|---|---|---|---|
| *Ital.* | *U* | *5* | *8* | *9* | *9* |
| *Eng.* | *U* | *2* | *4* | *6* | *9* |

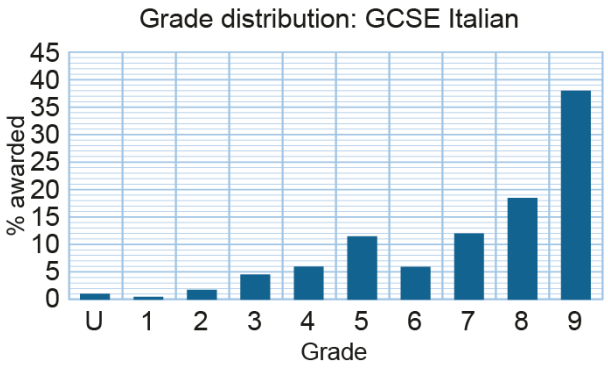
e) *How accurate were your predictions?*

In *Example 4*, students are asked to predict and then accurately construct box plots for two differently-shaped distributions. The intention is for students to make connections between the different **representations**, further consolidating their understanding of how the data are distributed. If students find this challenging, it may be helpful to shade the bars to show the data points between each quartile, as in *Example 3*.

The **language** of skew is not introduced here, to ensure the focus is still on mastering the basic concept of a box plot and how it will change when the data change. Later examples explore skew in more depth, and teachers may find it helpful to revisit this example at the same time.

In this example, students are first offered a bar chart and box plot for some data that are reasonably close to the normal distribution. The intention is to offer an anchor for students to use when comparing the other data sets, to support them in identifying the shape of the box plots. However, there is potentially a missed opportunity to explore the GCSE Mathematics data set with students first. Discuss with your colleagues what questions you might ask students to help them gain a deeper appreciation of normally-distributed data.



*Example 5:*

*A class's performance in maths and science tests is represented with box plots (presented below this example).*

*The results for the science test are also presented in a grouped frequency table, below.*

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

a) *Create a table with the the same class intervals to show the maths test scores. Use the box plot for maths below to complete the frequencies.*

b) *Is there more than one possible way to complete the grouped frequency table?*

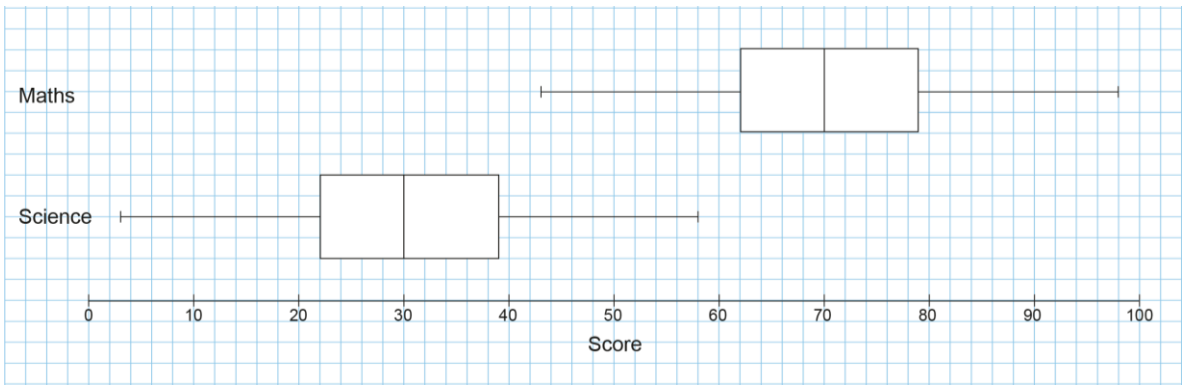
*Example 5* explores two data sets that have the same distribution, but the median for the maths test is 40 marks higher than the median for the science test. The box plots are identical in appearance, but the position of them on the number line tells us where the scores for the two tests are situated. The **variation** between the maths and science results, which maintains the same distribution of test scores for the two subjects, draws attention to the values in the box plot and what they tell us about the data.

Students are likely to read off the two medians as 30 and 70. Do they recognise that there is a difference of 40 between all five key pieces of statistical information? It is likely that students will initially imitate exactly the frequency values given in the grouped frequency table for science scores. In **deepening** their understanding, they should focus on the five key values and recognise that these remain unaffected, but that there can be some variation in the frequency amounts.

Students are asked to think about the frequency tables alongside the box plots. In exploring the connections between these **representations**, there are several misconceptions that teachers might uncover. For example, students may assume that every student scored exactly 40 more marks in maths than they scored in science. Plan ahead for questions to tackle this, such as:

- *'If the box plot values for maths are all 40 marks more than the box plot values for science, does that mean that each student scored exactly 40 more marks in maths than science?'*
- *'Is it possible for the box plot values for maths to all be 40 marks more than the box plot values for science, but the frequencies in the two grouped frequency tables to not correspond?'*

Another misconception is confusing different averages. How might you check that students are not confusing the median with the mode when identifying the highest frequency as corresponding to the class interval containing the median value?



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

*Example 6:*

*A class's performance in history and science tests is represented with box plots (presented below this example).*

*a) Match the grouped frequency tables to the correct box plot below.*

*b) Explain how you know.*

*Table A*

| *Score on test* | *Frequency* |
|---|---|
| *0-10* | *10* |
| *11-20* | *4* |
| *21-30* | *4* |
| *31-40* | *4* |
| *41-50* | *7* |
| *51-60* | *6* |

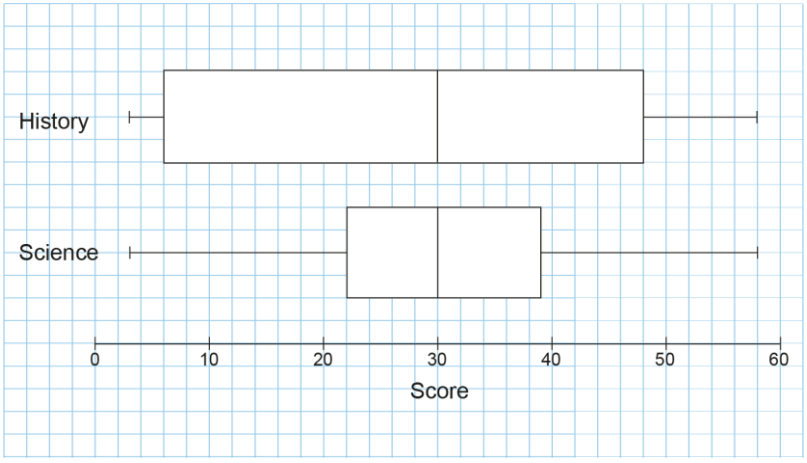
*Table B*

| *Score on test* | *Frequency* |
|---|---|
| *0-10* | *2* |
| *11-20* | *6* |
| *21-30* | *12* |
| *31-40* | *10* |
| *41-50* | *4* |
| *51-60* | *1* |

*Example 6* focuses on the distribution of two data sets that have the same median, minimum and maximum values, but the variability around the median is different. Students often assume that the size of the box on a box plot is representative of the number of data values, rather than the distribution of the data. Including the corresponding frequency table **representations** in this example addresses this misconception. Students need to recognise that the shorter box on the box plot for science scores represents a smaller interquartile range, meaning that the middle half of the data have little variability, and the test scores consistently lie around the median score. In contrast, the longer box on the history box plot indicates a larger interquartile range, implying more variable test scores. The whiskers on the box plots extend to the same minimum and maximum values for the two data sets. This tells us that although the middle half of the history test data have a lot of variability, the history test scores are not distributed any wider than the science test scores.

Establishing that the number of data values within each quartile cannot be determined from a box plot alone, but that we do know that each of the four quartiles must contain the same number of data values, is important in **deepening** students' understanding of what information can and cannot be determined from a box plot. By exploring this in conjunction with the grouped frequency table, students can begin to get a clearer picture of the shape of the distributions of the two data sets.

Discuss how students may benefit from the opportunity to devise a data set of values that satisfy both the summary values represented in the box plot (minimum, lower quartile, median, upper quartile and maximum scores) and the distribution of data outlined in the grouped frequency table. It is not necessary to do this to interpret a box plot, but it can provide valuable insight into what these representations tell us about the centre and spread of the data in question.



**Appreciate that the length of the whiskers needs to be examined together with the box to be able to compare the distribution of data values using box plot representations**

*Example 7:*

*Students sit tests in science, history and French. Box plots for all three subjects are created, and both history and French compared with science (presented below this example).*

*a) Compare the distributions of the French and science test scores.*

*b) Compare the distributions of the history and science test scores.*

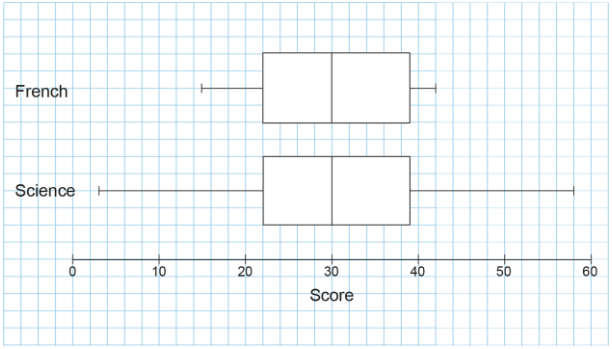
*c) Comment on the similarities and differences between your answers to parts a and b.*

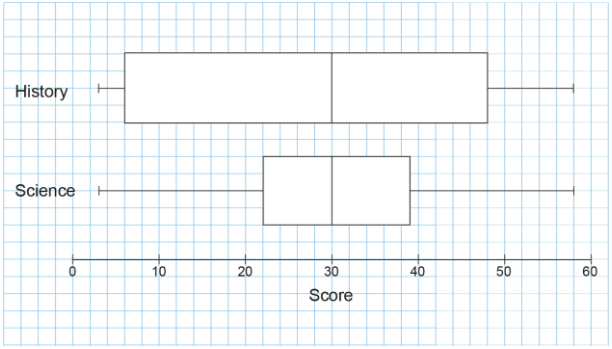
*Example 7* compares box plots where some of the summary statistics are the same and some are different. This draws attention to what the lengths of the box/whiskers within the **representation** tell us about the distribution of the data. The whiskers of a box plot represent the bottom and top 25 per cent of the data, and can be used to determine the extreme values of a data set.

The **variation** in this question comes from the similarities and differences between the source data. In part a, the length of the box is the same for the two box plots, so we can tell that the distribution the middle 50 per cent of scores is the same for both French and science. Although the interquartile range is the same for French and science, the overall spread is very different. This means that the bottom and the top 25 per cent of the French scores are a lot less scattered than the bottom and the top 25 per cent of the science scores. Part b compares data sets where, conversely, the minimum, median and maximum values are the same, but the interquartile ranges are quite different. The benefit of the interquartile range as a measure of spread is highlighted here, as the differences in the distribution of the history and science scores would go undetected, if just the minimum, median and maximum values were compared.

Students may assume that because the median and lower and upper quartiles in part a are the same for the French and science tests, the test scores within the interquartile range must be identical. Recognising that we cannot tell from a box plot what the actual scores are, but we know that the variation around the median is the same for the two subject areas, is fundamental to students **deepening** their understanding of what can and cannot be identified from the length of the box in a box plot.

In *Example 7* the median score of 30 marks for French, science and history has been kept constant to help students to focus on using the box plots to compare the spread of the scores. Discuss the importance of paying particular attention to what stays the same and what changes, to help develop a deep understanding of features of a box plot rather than a more generalised surface knowledge of a particular representation.



**Understand how to identify skew in a data distribution**

*Example 8:*

*The test scores for a class of 35 students are recorded in a grouped frequency table.*

| *Score on test* | *Frequency* |
|---|---|
| *0-10* | *8* |
| *11-20* | *7* |
| *21-30* | *6* |
| *31-40* | *5* |
| *41-50* | *4* |
| *51-60* | *3* |
| *61-70* | *2* |

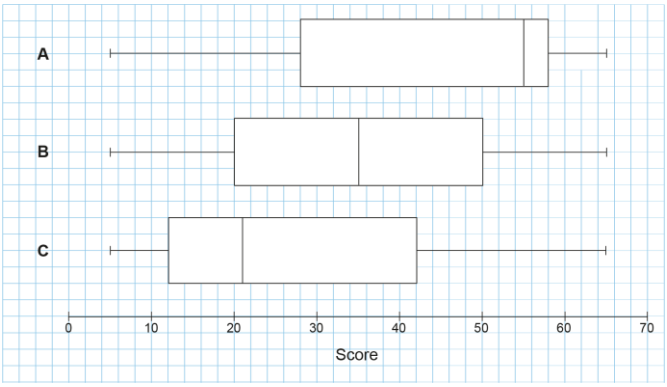
*Which of the box plots A, B or C presented below represents the data in the table? Explain your answer.*

*Example 8* focuses on how a skewed distribution can be identified from a box plot. Students may match the box plot with the grouped frequency table, simply by recognising that there are higher frequencies of lower test scores, and so you would expect the box plot to be situated towards the lower end of test scores. Discuss the box plots in some detail to support with **deepening** students' understanding of the features of skewed data sets.

The **variation** between box plots A, B and C (below) is such that the extreme values are the same (5 and 65) for all three, so that students cannot identify the correct one by considering the minimum and maximum values only. This supports them to notice the significance of the position of the median. When the median is in the middle of the box and the whiskers are fairly equal in length, as in box plot B; the distribution of the data is symmetrical; and there is no skew present. If the data are skewed, the median no longer divides the box into two equal pieces.

Students can compare the box plot and frequency table **representations** to build their sense of 'skew'. The data in the grouped frequency table consist mostly of low test scores, with a few high scores. The data are skewed to the right (positively skewed) and so when represented on a box plot, the median is closer to the lower end of the box, (so the longer part of the box is to the right of the median) and the whisker on the lower end of the box is shorter, resulting in a lopsided box plot. Checking the values for the lower and upper quartiles and the median against the grouped frequency table will help students embed this.

Although the mean is not represented in a box plot, it is possible to deduce how the data might affect the mean. A few extreme data values have an impact on the mean, pulling it to the left when the data are negatively skewed and to the right when they are positively skewed. As a result, the mean is less than the median for data that are skewed to the left (and vice versa). When there is no skew, the mean and median will have similar values and be equal for data that have a perfectly symmetrical distribution. Discuss the value in exploring this with students, and how you will mitigate against potential confusion with the median.



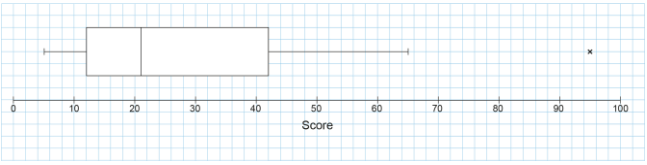
**Know how to determine whether a data value is an outlier or not**

*Example 9:*

*Zoe draws a box plot for her class's geography test scores (presented below this example).*

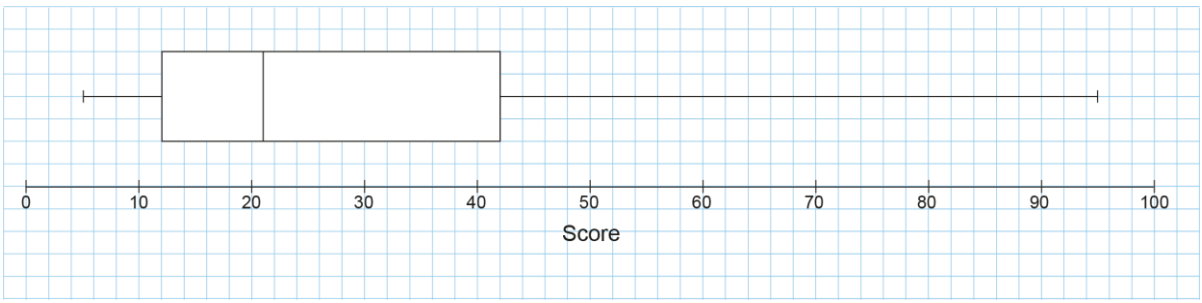
*Comment on Zoe's box plot. What advice would you give to her to improve the diagram?*

*Example 9* prompts a discussion about outliers and provides an opportunity for **deepening** students' understanding of when an extreme value has significance, and the impact it may have on statistical measures. It is likely that students will recognise that the data set represented by the box plot must contain at least one data value that is significantly bigger than the other data values, and identify this from the extended right whisker length. Establish how to check whether or not a data value is in fact an outlier (i.e., more than 1.5 times the interquartile range, below the lower quartile, or above the upper quartile) and highlight the way that outliers are usually presented on a box plot.




The data used to construct the box plot in *Example 9* contain a score of 95 and this score is 33 marks higher than the next highest score. In the context of scores on a test, it is feasible that the score of 95 is a genuine score. However, there may be situations where an outlier needs to be removed, deemed as being an erroneous data value.

It is important for students to recognise that, even though outliers do not affect or change the summary information used to construct a box plot (other than the range), outliers may have a significant effect on the mean value. This can result in it no longer being representative of the data set. Discuss with your colleagues how you might address the effect outliers can have on statistical analysis. What other contexts or real-life examples might support student understanding?



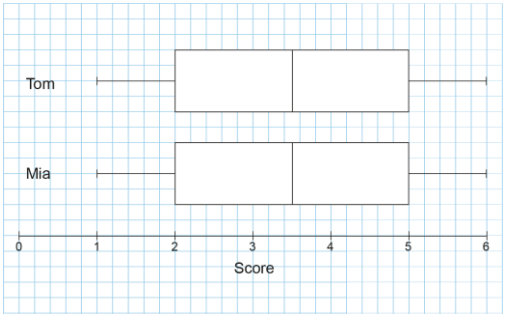
**Appreciate the information that can and cannot be determined from a box plot**

*Example 10:*

*Tom and Mia are each throwing a dice and recording the scores in a frequency table.*

| *Score* | *Frequency (Tom)* | *Frequency (Mia)* |
|---|---|---|
| *1* | *3* | *2* |
| *2* | *3* | *5* |
| *3* | *3* | *2* |
| *4* | *3* | *4* |
| *5* | *3* | *4* |
| *6* | *3* | *1* |

*They decide to draw a box plot to represent their results.*




*Explain why their box plots are identical when their results are different.*

*Example 10* highlights that, for a box plot **representation**, having the same summary information for two data sets doesn't imply that the two distributions are identical. The summary information used when constructing a box plot does not tell us anything about the distribution of the values between the minimum and lower quartile, or between the lower quartile and the median, etc. While we know that the frequency between the minimum and lower quartile must match the frequency between the lower quartile and median, we don't know which values these frequencies are allocated to.

Presenting the raw data, as well as the box plot, provides an opportunity for **deepening** students' understanding of the effects that data values have on the five pieces of information needed to construct a box plot. They should recognise that there can be more than one data list that produces the same summary information.

Students may find it helpful to use the frequency table to write out the two lists of data:

Tom: 1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6

Mia: 1, 1, 2, 2, 2, 2, 2, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 6

Ask teachers to produce a different data set that produces the same summary information. Share approaches and discuss how asking students to do the same might help them to develop a deeper understanding of how the median, quartiles and minimum and maximum values affect the box plot representation. Discuss the benefits – and challenges – of asking students to find all possible data sets.

*Example 11:*

*Thirty students took an English exam. The pass mark was a score of 40. The results are shown on a box plot (presented below this example).*

*Determine which of the following four statements A-D are 'True', 'False' or 'Cannot tell from the box plot'.*

A *Fifteen students scored at least 30 marks on the exam.*

B *Only one student scored 56 marks on the exam.*

*Example 11* exposes some common misconceptions when interpreting box plots, and provides an opportunity for **deepening** students' thinking around what information about the data set can be established from a box plot representation.

It is important for students to be able to explain their choice of category for each of the four statements A to D and to be precise in their use of **language** when referring to the summary statistics represented by the box plot. Below are some suggested prompts to help elicit the key understanding for each statement.

- Recognising that the median represents the value for which half the scores are greater than and half the scores are less than, is key to identifying A as being a true statement. Ask students, *'How many students scored 30 or more marks? How do you know?*

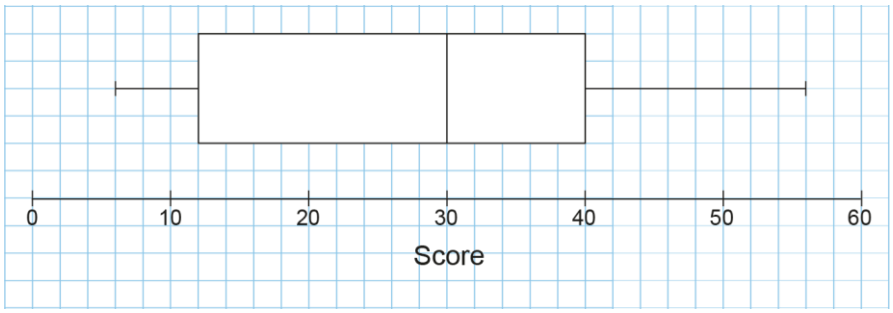
C *The mean score on the exam was 30 marks.*

D *More than half the students failed the exam.*

- Students often assume that the ends of the whiskers only represent single minimum and maximum value. We cannot tell from the box plot whether statement B is true, as it is possible that more than one student scored 56 marks. Ask students, *'What's the maximum number of students that could have scored 56 marks?'*
- Students often mistake the middle line of a box plot as representing the mean rather than the median, probably because the mean is a more commonly used measure of central tendency. However, we cannot determine if C is true from the box plot. Ask students, *'Is it possible for the mean to be equal to the median?'*
- To identify statement D as being true, students need to recognise that the upper quartile is effectively the median of the upper half of the data, and here it represents the pass mark of 40. Ask students, *'What percentage of students got at least 40 marks?'*

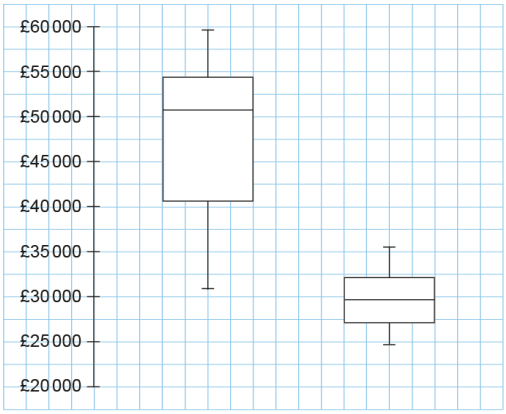
Discuss with your colleagues the value of asking students to construct a possible data list, rather than providing it for them. The act of determining data values that satisfy the summary statistics is key to developing students' understanding of how the values in a data set are reflected in a box plot representation. For example, these scores have summary statistics that match the box plot in this example, and also have a mean of 30:

6, 8, 10, 10, 12, 12, 12, 12, 16, 18, 20, 21, 29, 30, 30, 30, 33, 36, 37, 40, 40, 40, 40, 40, 41, 54, 55, 56, 56, 56



*Example 12:*

*Keir and Eluned are comparing how much money people earn in their local areas. They find some data[1] about the average net annual income, which is the amount of money earned in a year, after tax has been taken. They use this to produce the following box plots:*




a) *What can you say about the two different areas from the data presented in these box plots?*

*Eluned says, 'People in Flintshire earn less than people in Kensington and Chelsea.'*

b) *Based on this, which box plot do you think refers to which area? Why?*

c) *Is Eluned's statement accurate? Why or why not?*

d) *What can you say about the half of earners in the middle for each area?*

*The mean for Kensington and Chelsea is £48576. The mean for Flintshire is £29550.*

e) *Mark the mean with a cross on each box plot*

f) *What does the position of the mean tell us about the distribution?*

The areas selected for the two box plots in *Example 12* have an overlap and different levels of skew. Flintshire and Kensington and Chelsea have been chosen as two areas which provide high contrast. The context is intended to elicit conversations about even distributions as they relate to fairness. Careful **language** can help students to support their initial conclusions from the box plots.

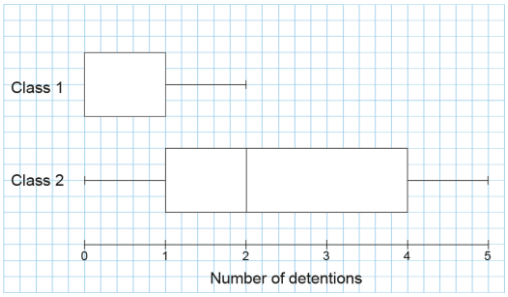
To support with **deepening** understanding of these summary statistics, the table below shows the national distribution, and data from other contrasting regions. Ask students to construct box plots for these data and to draw conclusions about the different areas.

| | England & Wales | Devon | Blaenau Gwent | Liverpool |
|---|---|---|---|---|
| Max | 67000 | 36900 | 25200 | 44900 |
| UQ | 35900 | 33625 | 25100 | 33300 |
| Median | 32000 | 32200 | 24600 | 28400 |
| LQ | 28500 | 30800 | 23900 | 25800 |
| Min | 16700 | 28200 | 23600 | 24000 |
| Mean | 32550 | 32271 | 24500 | 29816 |

Teachers may wish to explore the source data, which are freely availbale on the ONS website, and view the data for their schools' local areas. Discuss with colleagues when or if you use real-life data with students. How confident are your team with using data like these which can generate heated discussions around social issues? Are teachers comfortable finding their own data sets from reliable sources and using them in the classroom?

*Example 13:*

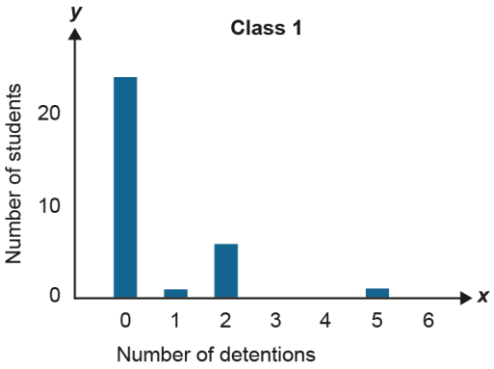
*A headteacher is looking at the number of weekly detentions that students in two different classes are given by the same teacher. The data are shown by DETENTOTRON detention tracking software as box plots:*

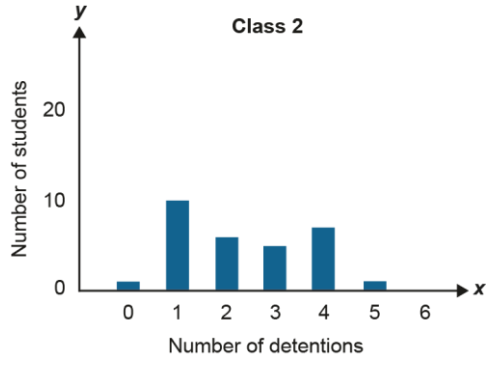



*a) What does the box plot tell us?*

*b) Which class would you rather be in? Why?*

*The headteacher represents the data as bar charts instead:*




*c) What does the bar chart tell us that the box plot doesn't?*

*d) What does the box plot tell us that the bar chart doesn't?*

*e) Which is better for deciding which class you would rather be in? Why?*

*Example 13*, the final example in this key idea, is once again an opportunity to compare the **representations** of a box plot with a bar chart, but this time with more of a focus on interpreting the data and drawing conclusions. The question prompts are deliberately open, to give teachers the opportunity to assess students' understanding more broadly.

As well as checking if students' reasoning is sound, it is worth listening carefully to the **language** that students use when answering the questions.

## 10.2.1.4 Understand the idea of accumulation and why it is useful

**Common difficulties and misconceptions**

When identifying the median and quartiles from an ordered list of data, students can count along to the value in the required position. However, when the data set becomes too extensive to be presented in a list, and is instead represented in a frequency table, they often struggle to recognise how the frequency table describes a list of data in an ordered and concise way. The idea of accumulation provides a way of interpreting the table by considering the number of data values that have gone before, providing insight into the position of certain values in the data set. Presenting the raw data alongside a frequency table can help to make this link more explicit to students.

| Students need to | Guidance, discussion points and prompts |
|---|---|
| **Use the idea of accumulation to find the median for a data set presented in a frequency table**<br><br>*Example 1:*<br><br>*Jake works in a shoe shop and is doing a stocktake of the sizes in a recent delivery. He records the results in a table:*<br><br>Shoe size / Number in stock<br>4 / 3<br>5 / 2<br>6 / 3<br>7 / 3<br>8 / 2<br>9 / 1<br>10 / 1<br><br>*Jake is struggling to work out the median shoe size, so he asks his colleague Rachel for help. She adds an extra column to the table:*<br><br>Shoe size / Number in stock / Running total<br>4 / 3 / 3<br>5 / 2 / 5<br>6 / 3 / 8<br>7 / 3 / 11<br>8 / 2 / 13<br>9 / 1 / 14<br>10 / 1 / 15 | *Example 1* explores the idea of using accumulation to support the identification of the median from data that are presented in a frequency table. The data set contains 15 values, to make it feasible to also use a list **representation** if students need support in interpreting the table. It is important that students have the opportunity to work with the frequency table first but, if they are struggling, ask them to write out the shoe sizes individually. While they should not come to rely on this method, it can help to emphasise the relationship between data presented in a frequency table and the raw data values written as an ordered list.<br><br>When there is an odd number of data values in the data set, the median is a value within the data set itself. It is important that students recognise that the median is the $\frac{n+1}{2}$th value (where $n$ is the total number of values in the data set) and can distinguish between the position of the median within the data set and its value. Building on this, further **deepening** students' understanding, might involve asking them to use the frequency table to determine the lower and upper quartiles.<br><br>Discuss with colleagues some potential misconceptions; why students might identify the wrong value as the median; and what prompts might be helpful in addressing the issue. After sharing your thoughts as a department, consider the bullet points below. Did you identify the same misconceptions and potential strategies?<br><br>• Identifying eight as the median shoe size may be a result of confusing the *position* of the median with its *value*. Support students to engage with the frequency table and the running total column. Teachers could ask students, *'If all the shoes were put in order of size, what size would the fourth pair of shoes be? What about the tenth pair?'*<br><br>• Identifying seven as the median shoe size may be a result of identifying the middle row in the table, rather than the data value that is in the middle. Teachers |

| | |
|---|---|
| *Explain how the extra column can be used to help find the median.* | could change the number in stock so that the distribution is symmetrical (so that the number of each size in stock is 1, 2, 3, 3, 3, 2, 1 respectively). This then allows them to generate and compare the running total columns for this new table with the original table. Supporting questions could include, *'Why is the median seven in this new table, but not in the original table? What would happen if there were more size fours but fewer size eights and nines?'* |
| *Example 2:*<br><br>*Simon works in a shoe shop and is doing a stocktake of the sizes of shoes he has in stock of a particular design of shoe. He records the results in a table:*<br><br>Shoe size / Number in stock<br>4 / 3<br>5 / 2<br>6 / 3<br>7 / 3<br>8 / 2<br>9 / 1<br>10 / 2<br><br>*Simon is struggling to work out the median shoe size and asks his colleague Taylor for help. Taylor adds an extra column to the table:*<br><br>Shoe size / Number in stock / Running total<br>4 / 3 / 3<br>5 / 2 / 5<br>6 / 3 / 8<br>7 / 3 / 11<br>8 / 2 / 13<br>9 / 1 / 14<br>10 / 2 / 16<br><br>*Explain how the extra column helps Simon to find the median.* | *Example 2* is an adaptation of *Example 1*, constructed so that students can still work on it without necessarily having completed the previous task. This time, an additional data value has been included, so that the data set contains an even number of values. When there is an even number of data values in the data set, the median is the average of the data values in the $\frac{n}{2}$th and $(\frac{n}{2}+1)$th positions (i.e., the mean average of the two middle values). The **variation** in the data used in *Example 2* compared with *Example 1* is such that the $\frac{n}{2}$th and $(\frac{n}{2}+1)$th values are not the same (shoe sizes 6 and 7). When the two middle values are the same, the average takes on the same value as the two middle values. In this example, the median value (6.5) makes sense within the context of shoe sizes, as it is possible to have a 6.5 size shoe, as shoes can come in half sizes.<br><br>While the median is a feasible shoe size, it is not a size in the given data set. It is worth highlighting to students that the median can be a value that is not in the data set, or even not possible in the data set. Establishing that the median represents the value for which half the data values are less than this value, and half the values are greater than this value, is key to students **deepening** their understanding of what the median tells us about the data set it is representative of, and how it relates to the data values themselves.<br><br>  These first two examples explore two possible eventualities for where the median is situated within the data set: an odd total number of values where the median is in the data set, and an even total number of values where the median is **not** in the data set. Is it possible for the converse to be true, i.e., for an odd total number of values where the median is not in the set, and an even total number of values where the median is? Share your reasoning with your colleagues and discuss whether it offered any deeper insight into the structure of the median in the context of a table. Would there be value in posing a similar question to your students? |

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

*Example 3:*

*A sheep farmer weighs his sheep to check if they are ready to be sold at market. He records the weights for the ewes (female sheep) in a separate table to the rams (male sheep).*

*The ewes' weights are recorded in the table below*:

| Weight (kg) | Number of ewes |
|---|---|
| 45 | 1 |
| 46 | 2 |
| 47 | 0 |
| 48 | 0 |
| 49 | 3 |
| 50 | 2 |
| 51 | 1 |
| 52 | 3 |
| 53 | 0 |
| 54 | 0 |
| 55 | 3 |

a) *Find the interquartile range for the ewes' weights.*

*The rams' weights are also recorded in a table:*

| Weight (kg) | Number of rams |
|---|---|
| 45 | 1 |
| 46 | 0 |
| 47 | 3 |
| 48 | 2 |
| 49 | 1 |
| 50 | 2 |
| 51 | 1 |
| 52 | 2 |
| 53 | 0 |
| 54 | 1 |
| 55 | 1 |
| 56 | 2 |

In *Example 3*, the **variation** between the ewes' and rams' weights is such that there is an odd number of ewes and an even number of rams. This provides an opportunity for students to think about how the number of values in a data set affects how the lower and upper quartiles are determined. The lower quartile is the value in the $\frac{n+1}{4}$th position and the upper quartile is the value in the $3(\frac{n+1}{4})$th position. When there is an odd number of values in a data set, the lower quartile and upper quartile are values within the data list (the fourth and twelfth data values in this example) and can be identified fairly easily. Students may like to think of the lower quartile as being the median of the lower half of the data (i.e., the median of 45, 46, 46, 49, 49, 49, 50) and the upper quartile as the median of the upper half of the data (51, 52, 52, 52, 55, 55, 55).

When working with data sets that contain an even number of data values, determining the interquartile range is often less straightforward. The fourth and fifth data values are not the same for the rams' weights and so a calculation is needed to find the lower quartile (47.25 kg). Similarly, the twelfth and thirteenth data values are different, so the upper quartile (53.5 kg) is not a value within the data set. Being very precise with the difference between categorical, cardinal and ordinal **language** is essential to prevent confusion here. The numerical values in the first column essentially categorise the data, so it is only the values in the second column that denote quantity (and thus are cardinal numbers). Ordinal numbers denote position, and so are a way of identifying different quantities within that column.

Grouping this data using the **representation** of a grouped frequency table might provide an accessible way for students to explore finding the interquartile range for a grouped frequency table. They can then identify the similarities and differences between and working with a frequency table where it is possible to determine the raw data, and a frequency table where it is not. Consider designing further examples that you might use with students to expose these similarities and differences.

For many students, sheep farming is perhaps not something that they can readily relate to. It is, however, important that students work with data from both familiar and unfamiliar contexts. Familiar contexts offer opportunities for students to bring their own knowledge and experience to support interpretation and sense-making – with the added benefit of teachers being able to use data to challenge biases and preconceptions. In contrast, unfamiliar contexts can initially be off-putting, but it is important that students also get used to working with data from contexts that they have little prior knowledge of – with the added benefit of students being able to learn more about their world through the lens of data. There is much to consider when sourcing or

| | |
|---|---|
| *b) Find the interquartile range for the rams' weights.*<br>*c) Comment on the differences when finding the interquartile range for each set of data.* | generating data sets for students to work with – from the different contextual factors in different school settings, to the implications of data for the students' wellbeing. It could be problematic, for example, to draw attention to weight if the context were students rather than sheep. Discuss this with your team – can you think of some familiar and unfamiliar contexts to use for your students? |
| **Begin to appreciate how cumulative frequencies can be represented on a graph**<br>*Example 4:*<br>*Are the following statements true or false?*<br>*You might find it helpful to use the data from Examples 1 to 3 to justify your answers.*<br>*a) Cumulative frequency curves do not always start at 0.*<br>*b) When two curves are plotted on the same graph, they will start and finish in the same place.*<br>*c) The median value can always be found in the middle of the $x$-axis.* | While this key idea stops short of asking students to plot cumulative frequency curves, it is important that they are prepared to connect their learning to this new **representation**. *Example 4* offers some common misconceptions and asks students to decide if they are true or false. It may be helpful to offer them the data sets and some graph paper to help them create counterexamples.<br>  The next two key ideas in this core concept – '10.2.1.5 Construct cumulative frequency graphs and use to estimate information about the data' and '10.2.1.6 Interpret and use features of data from a cumulative frequency graph' – are not exemplified. Using *Example 4* as a starting point, work with your team to design a sequence of examples that explore one of these two key ideas. |

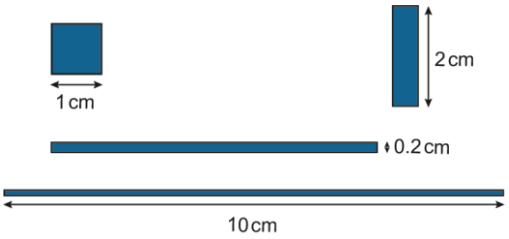
## 10.2.1.7 Construct histograms for a given data set

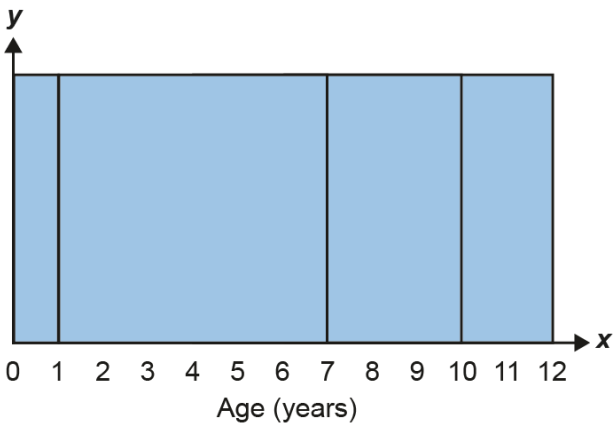
**Common difficulties and misconceptions**

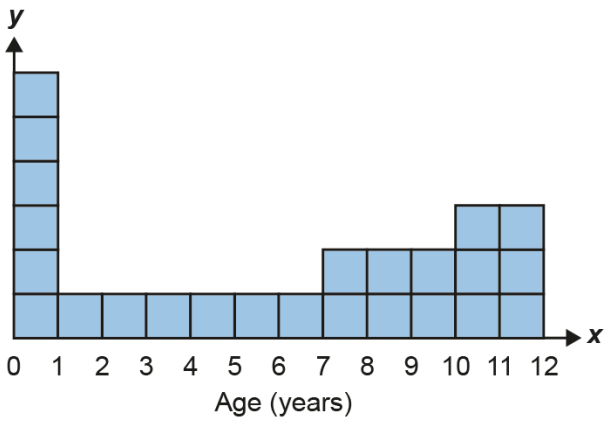
Histograms are constructed using a continuous scale with the bars touching each other. However, students often label bars with the class interval, or leave gaps between the bars, in a similar way to when constructing a bar chart.

When the classes in a grouped frequency distribution are not continuous, they need to be made continuous before the frequency densities can be calculated and represented on a histogram. This is done by finding the difference between the upper limit of a class and the lower limit of the next class, and then adding half of the difference to all the upper limits and subtracting it from all the lower limits in the distribution. Students often struggle with this necessary adjustment and can find it difficult to understand why a bar representing the class interval 11-15, for example, begins at 10.5 on the $x$-axis and ends at 15.5.

Another common mistake that students make when constructing histograms is to represent frequency, rather than frequency *density* on the $y$-axis. Recognising that the frequency of the data is measured by the area of the bars, rather than by their height, is fundamental to understanding the mathematical structure of the histogram representation.

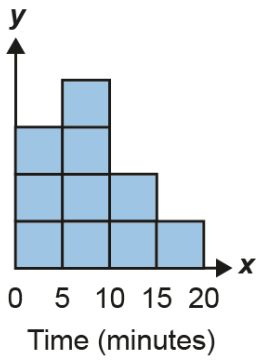
| Students need to | Guidance, discussion points and prompts |
|---|---|
| **Know that frequency density rather than frequency is plotted on the $y$-axis of a histogram**<br><br>*Example 1:*<br><br>*All the rectangles below represent one unit.*<br><br>*What are the missing dimensions for each of the rectangles?*<br><br>  <br> | Central to interpreting histograms is the concept of frequency density, which in turn requires both fluency and depth of understanding of the area of a rectangle. In *Example 1*, students are offered a range of rectangles with an area of one unit and asked to find the missing dimensions. The **representation** of rectangles, away from the context of a graph, is used so that students are building the new knowledge of histograms on an area of mathematics that they should feel confident in. Draw their attention to how, as the width increases, the height decreases, but the area remains constant. |
| *Example 2:*<br><br>*A vet records the ages of the 24 dogs that she treats in her clinic over a day. She notices that she can group them into four equal-size groups if she uses unequal age categories.*<br><br>*Which of the histograms below does this better? Give reasons for your answer.* | In *Example 2*, students are offered a scenario with a rationale for why area, rather than height, is used to denote frequency. Students need to recognise that the histogram A is not an effective **representation** of the data because the four bars look very different, and so it could easily be misinterpreted that far more dogs were in the 1-7-year age category. The $y$-axis is deliberately left unlabelled, and the bars on the second histogram divided into square units, to help students appreciate why area is a more helpful measure. Much as students will have first experienced area by counting squares, they may find histograms more accessible if they first encounter the bar areas divided into squares of one unit.<br><br>The **variation** in this example is such that the frequency within each category is constant, to support students to make the connection between the area of the bars and the relationship between frequency and class width. |



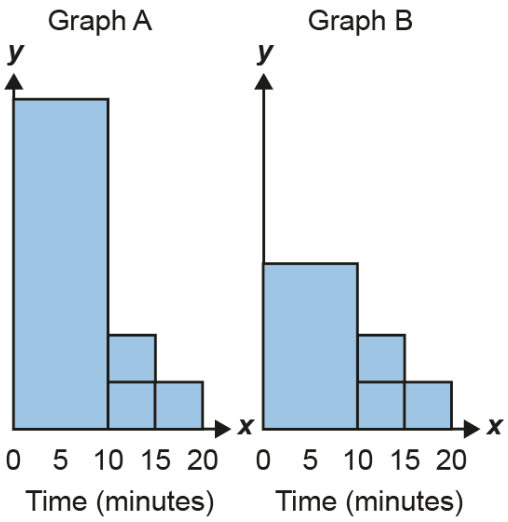
*Example 3:*

*A doctor records the lengths of ten appointments with her patients in a histogram of equal class widths, as shown below:*




*She decides to group together all appointments are ten minutes long, or less.*

*Which of her graphs below does this better? Give reasons for your answer.*




*Example 3* offers a visual **representation** to consolidate students' understanding of how frequency density differs from frequency. The misconception exposed by histogram A – using frequency on the *y*-axis even though the bar is now twice as wide – is similar to that shown in histogram A of *Example 2*. Showing the source data adds another dimension to students' understanding – because the frequencies of 3 and 4 are so close together, students can readily visualise one unit being divided equally between the two bar-widths so that the quantity is preserved. Keeping the other two bars allows students to easily compare the new bar size to the original.

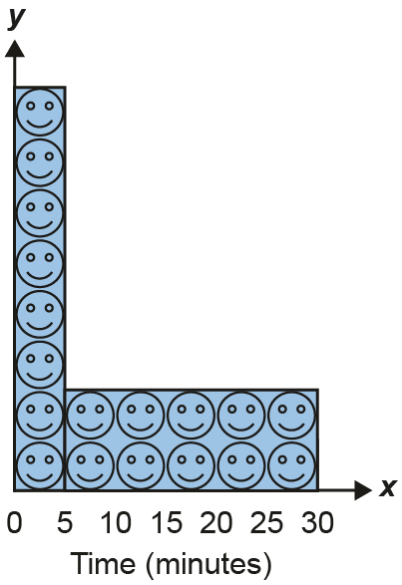
Teachers may have noticed that, in all of the examples up to *Example 7*, the *y*-axis remains unlabelled. Avoiding the **language** of frequency density at this stage is intentional, so that students can focus their attention on building a conceptual understanding of how quantity is represented by area. When frequency density is introduced, teachers are encouraged to relate this to other compound measures, so that students are supported to appreciate the idea that they are looking at the frequency per square unit.

The data values chosen in *Example 3* are deliberately small and close together to support students in comparing the areas and heights of the bars. While there is a degree of realism to the context, in that it is plausible that a doctor might record appointments of about this length, it is nonetheless not realistic data. Real-world data sets are often 'messy', with ranges and anomalies that can be hard to represent graphically without some compromise. Students need to gain experience of this but also learn key statistical concepts without distraction. Discuss with your colleagues the advantages and disadvantages of using artificially curated data versus genuine real-world data. Does your curriculum offer a balance of both?

*Example 4:*

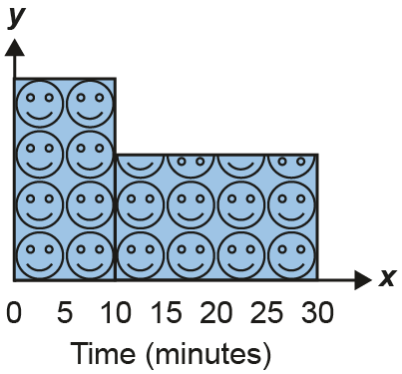
*A school records student punctuality. Any student who is up to five minutes late is given a warning. Any student who is more than five minutes late is given a detention.*

*The histogram shows how many students were late on Monday morning.*




a) *Which was the larger group? How do you know?*

*The school decides to change their guidance and only give a detention to students who were more than 10 minutes late. The deputy headteacher constructs a new histogram for the same Monday morning data.*




b) *What do you notice?*

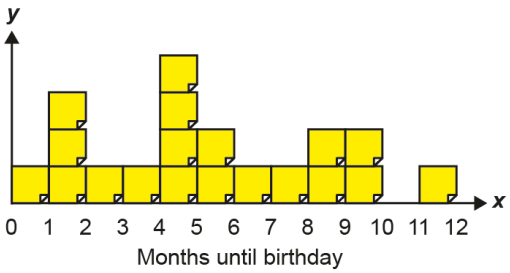
*Example 4* represents a step away from the grid-like structures of the previous three examples, but still includes visuals to support students to recognise how the number of students is represented within the histogram.

Pay attention to the **language** that students use when offering their suggestions for which is group was larger. A useful exercise might be to offer the first histogram without the images of faces or the scale on the $x$-axis, and ask students to try to identify the larger group without those visual clues. This may encourage students to consider area as a useful comparative measure, as they cannot easily compare the size of rectangles when the widths are so different.

Part b offers an opportunity for **deepening** students' understanding of what information they can and cannot deduce from a histogram. The number of students in the two new groups is identical to the original groupings, which must mean that there were actually no students who were between 5 and 10 minutes late. This is something that is particular to this set of data, as there is no other way of knowing where in the group the individual students are situated.

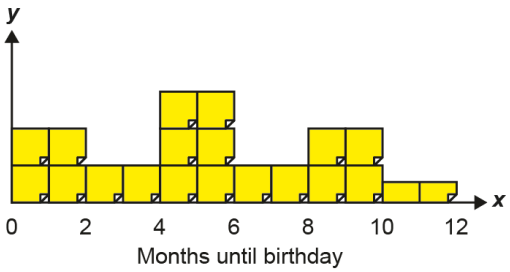
*Example 5:*

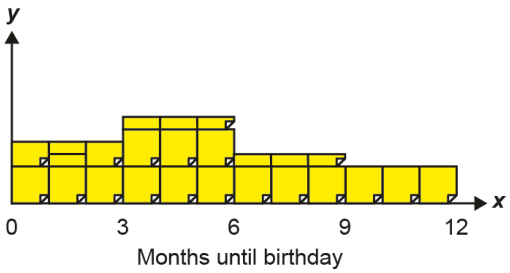
*A teacher asked her students to place one sticky note on the axes based on how many months it was until their birthday. They created the following histogram:*




*a) State three facts that you can determine from the histogram.*

*The teacher rearranged the sticky notes to create two new histograms, shown below.*

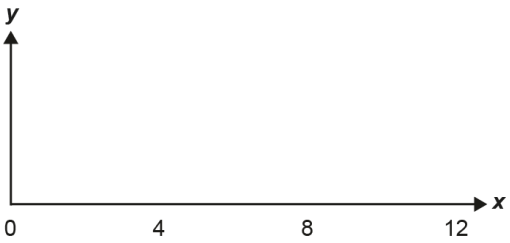


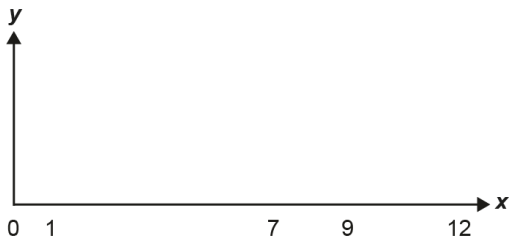




*b) What is the same and what is different about these histograms?*

*The teacher rearranges the sticky notes into two more new histograms, one with equal and one with unequal width groups.*

*c) Use the axes below to sketch what these histograms would look like.*




*Example 5* serves a similar purpose to the preceding three examples, but uses the **representation** of a sticky note as a tangible and familiar item that students can manipulate if required. Teachers may find that it is helpful to recreate the histogram shown in the rubric, and then physically rearrange the sticky notes to demonstrate how the area of each wider bar now represents the same frequency. This could be particularly impactful for helping students to appreciate how it is possible to have a decimal value for frequency density. For example, by taking the 11-12 month sticky note, halving it, and then placing the two pieces side-by-side in the new 10-12 position. This clearly shows how the same value is now represented by a shape twice as wide but half as tall.

The **variation** in this example is such that students are given opportunities to really notice the effect that class width has on the height of the bars. The number and distribution of the data (sticky notes) remain constant, but the class width changes each time. Building up from equal to unequal class widths supports students to appreciate the multiplicative structures that underpin histograms.

Students are asked to work both from existing histograms and to create their own. Working 'forwards and backwards' in this way should support with **deepening** their understanding of how frequency is represented within a histogram

*Example 6:*

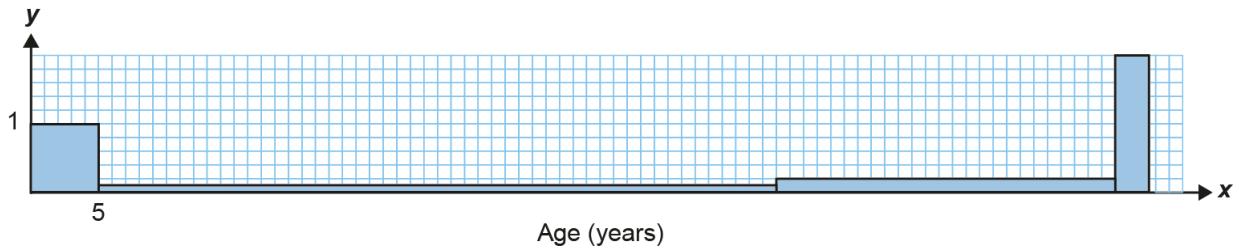
*All of the bars in the graph below represent the same number of people.*

*What are the missing class widths for each of the categories? The first class width is shown.*

*Example 6* is designed so that the bars all have the same area as the rectangles in *Example 1*. The **variation** in context between these two examples should support students to recognise that it is the same multiplicative reasoning that underpins both the area of a rectangle and frequency density in histograms. If students find this challenging, refer back to the rectangles from the first example and ask students what is the same and what is different.

As with the previous examples, the **language** of frequency density is deliberately avoided here, to focus on students' understanding of the representation. Teachers may like to ask students to suggest an appropriate label for the $y$-axis, which may generate interesting conversation about why 'frequency' is not appropriate. Students' suggestions for alternatives may reveal much about their grasp of the concept of frequency density, particularly if they have not yet been introduced to the term.

Discuss with your team what is the same and what is different about *Example 1* and *Example 6*. How might teachers best exploit the connections between the two examples, to support students to develop a deeper conceptual understanding of how histograms represent frequency?



*Example 7:*

*The post office records in a table the weights in grams of 100 letters.*

| Weight (g) | Frequency |
|---|---|
| $0 < x \leq 10$ | 20 |
| $10 < x \leq 30$ | 20 |
| $30 < x \leq 60$ | 15 |
| $60 < x \leq 75$ | 15 |
| $75 < x \leq 80$ | 15 |
| $80 < x \leq 100$ | 15 |

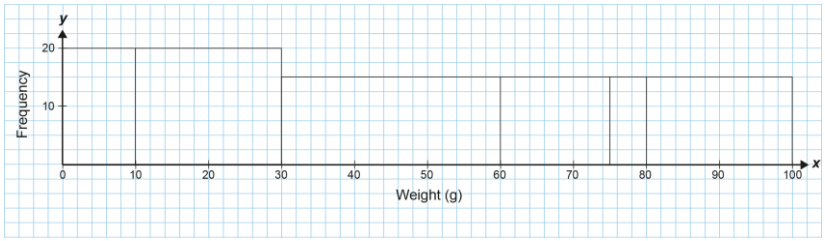
*Which of the graphs below is an accurate histogram for the data? Explain how you know.*

*Example 7* exposes the common misconception that the frequency values of a given data set are plotted on the $y$-axis. It is important that students can articulate why graphs A and B do not represent an accurate histogram, and are precise with their **language** to clearly distinguish between 'frequency' and 'frequency density'. A is a plot of frequency against weight and may be considered the odd one out, as it is the only graph that doesn't plot frequency density on the $y$-axis and so isn't actually a histogram. B appears to be very similar to A but has been generated by dividing all the frequencies by 10 (the first class width), rather than dividing the frequencies by their corresponding class widths to get the frequency density.

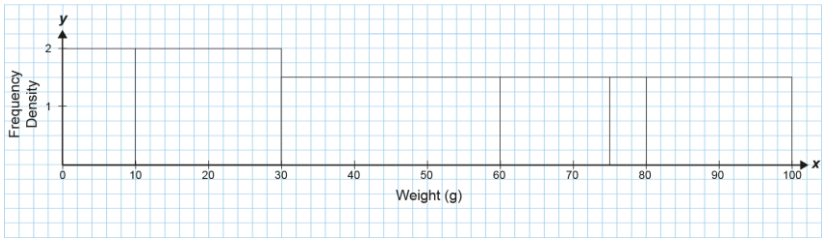
Students may select graph C as being correct, based on it having a different appearance to graphs A and B, without fully understanding why it provides an accurate histogram. Once it has been established that graph C is an accurate histogram for the data, there is opportunity for **deepening** understanding of the structure of a histogram. Suggested prompts for this include:

- *'Explain how the frequency density has been calculated from the frequency table.'*
- *'The frequencies are the same for the first two class widths ($0 < x \leq 10$ and $10 < x \leq 30$). Why are the heights of these two bars different? What is the same about these two bars?'*

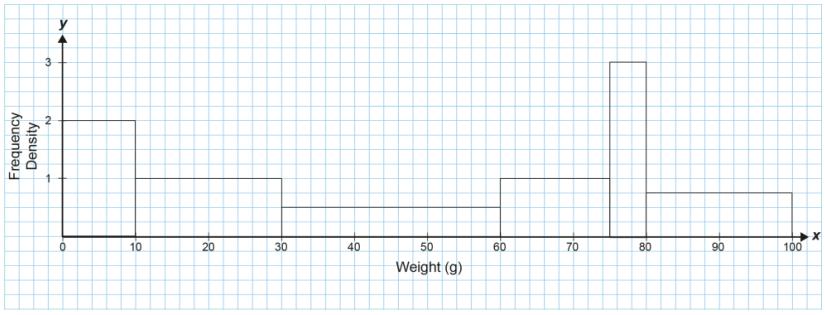
*A*




*B*




*C*



<table>
<tr>
<td>Understand the relationship between frequency, frequency density and class width<br><br>Example 8:<br><br>What would the frequency density be for a group with class width $10 < x \leq 30$ and a frequency of 10? Choose from:<br><br>A 2<br><br>B 0.5<br><br>C 20<br><br>Explain your choice.</td>
<td>Examples 8, 9 and 10 all explore the same frequency, class width and frequency density values, with a focus on deepening students' thinking about how they are related.<br><br>Students often learn a procedure to find the frequency density, without fully understanding the relationship. Thinking about the same values in detail and presenting multiple-choice options that include carefully designed variation, supports students in consolidating their understanding, and provides an opportunity for common mistakes to be exposed. In Example 8 students must identify the class width as 20, to determine the frequency density as 0.5. Ask them to explain what someone might be thinking if they selected options A or C as the frequency density.<br><br>How does your department handle student mistakes in the classroom? Are they seen as a positive opportunity to learn? Options A and C have been designed to expose some common misconceptions. Exploring these incorrect values can provide an opportunity for students to deepen their understanding, without feeling that they have made a mistake. Discuss the potential value in asking students what a hypothetical student might be thinking, and how this helps to both expose and address common misconceptions.</td>
</tr>
<tr>
<td>Example 9:<br><br>If the frequency is 10 and the frequency density is 0.5, what size is the class width? Choose from:<br><br>A 5<br><br>B 0.05<br><br>C 20<br><br>Explain your choice.</td>
<td>In Example 9 students will need to determine what the class width must be for the given frequency and frequency density. This should support with deepening their thinking about the relationship between the frequency, frequency density and class width.<br><br>How would you use this pair of examples in the classroom? Students may not recognise that Example 9 describes the same scenario as Example 8. Are there benefits of not exposing this to students at this stage, if they have yet to identify it for themselves?</td>
</tr>
<tr>
<td>Example 10:<br><br>What would the frequency be for a bar with height 0.5 and class width 20? Choose from:<br><br>A 0.025<br><br>B 40<br><br>C 10<br><br>Explain your choice.</td>
<td>In Example 10, the frequency density is not referred to explicitly. Instead, the bar's height is used to draw attention to the fact that the height of the bar in a histogram does not represent the frequency (the option of the frequency being 0.5 is not given as a choice of answer) and support students in deepening understanding of the structure of a histogram.<br><br>Using a representation may support students to be clear about this scenario: draw a bar with height 0.5 and width 20 to establish that the frequency is equal to the area of the bar.</td>
</tr>
</table>

*Example 11:*

a) *Complete the table below with the missing information.*

b) *Use the table to construct a histogram.*

*Example 11* provides students with all the information they need to be able to construct a histogram. However, some of the details are missing and need to be determined from the information that is given. Asking students to complete the missing information provides an opportunity for **deepening** their thinking about the relationships between the class width, frequency and frequency density.

The lower and upper limits of the class widths do not need to be completed in order to identify the missing frequencies/frequency densities. However they are necessary for the construction of the histogram **representation**. Students need to recognise the continuous nature of the data and that the lower limit of the second class interval needs to be the same as the upper limit of the first class interval.

This is the first example in which students are asked to construct a histogram. As a team, discuss your prior experience of this, either as students or as teachers. What mistakes do you anticipate students might make when plotting the bars along the $x$-axis?

| *Weight (g)* | *Class width* | *Frequency* | *Frequency density* |
|---|---|---|---|
| $0 < g \leq 20$ | | *80* | |
| $< g \leq$ | *10* | | *2* |
| $< g \leq$ | *30* | *90* | |
| $60 < g \leq 70$ | | | *2* |
| $< g \leq$ | *30* | *30* | |

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

*Example 12:*

*The ages of the residents on a street are recorded in a table.*

| *Age* | *Frequency* |
|---|---|
| *0-10* | *12* |
| *11-30* | *18* |
| *31-40* | *15* |
| *41-60* | *15* |
| *61-100* | *20* |

Annie produces histogram A (shown below this question). Ben says there should not be gaps between the bars.

a) Who is correct, Annie or Ben? Explain your reasoning.

*Example 12* explores some of the mistakes students may make when using the histogram **representation**, particularly referencing the continuous nature of the data. Unlike a bar chart representing discrete data, the bars on a histogram do not have gaps between them and so students need to identify that this is an error in Annie's chart. Whilst this can be easily remembered as a superficial rule, the inclusion of the second *almost* correct histogram is intended to provoke thinking and discussion around class boundaries, and how these are communicated visually.

Age is a helpful context to explore here as, while it is measured on a continuous scale, it is usually referred to discretely. For example, a 10-year-old child will usually be described as aged 10, until they reach their 11$^{th}$ birthday – in essence, we truncate rather than round their actual age. The change of context in parts c and d should help with **deepening** their understanding of how continuous data are grouped, as they need to consider instead the effects of rounding. To support students to recognise this difference, it may help to ask students to consider values on the

Ben produces histogram B (shown below Annie's histogram). Annie says that the first bar is still incorrect.

b) What is wrong with Ben's first bar?

Ben wonders whether his histogram would look the same if the data recorded was weight (to the nearest kg) rather than age.

c) Assuming the frequencies remained the same, would anything change for the histogram in this context?

Ben then wonders whether his histogram would look the same if the weights were grouped using inequalities ($0 \leq w \leq 10$, $10 < w \leq 20$ etc.)

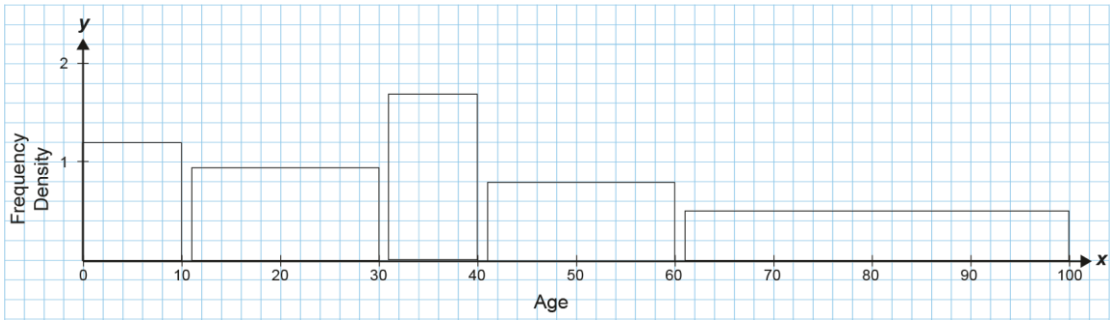
d) What would need to change now?

boundary of the class interval when they are thinking about part c. For example:

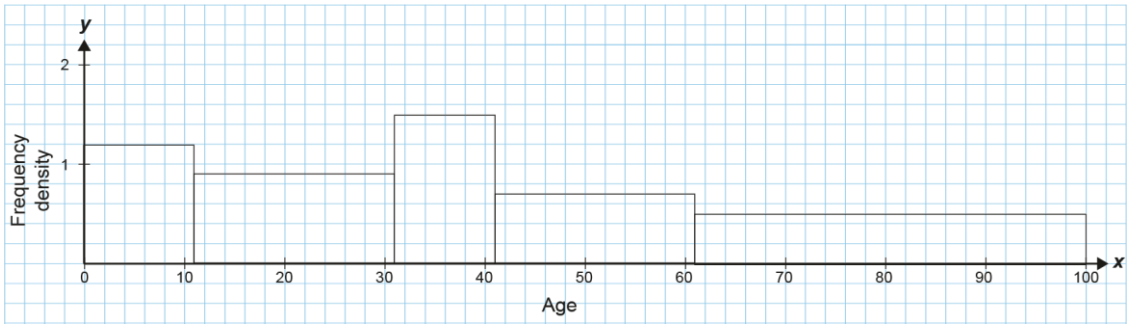
- *'In which class interval would you place someone was 10.5 years old? Would it be the same or different for someone who was 10.5 kg to the nearest kg?'*
- *'In which class interval would you place someone who was 10.4 years old? How about 10.4 kg?'*
- *'How about 10.6 years and 10.6 kg?'*
- *'How might you reflect these differences in the histogram?'*

What are the benefits of initially presenting the data to students using the class intervals 0-10, 11-20, etc., rather than using inequality notation? What does this help to expose? How does this then affect students' thinking when class intervals are introduced in part d?

*A*




*B*



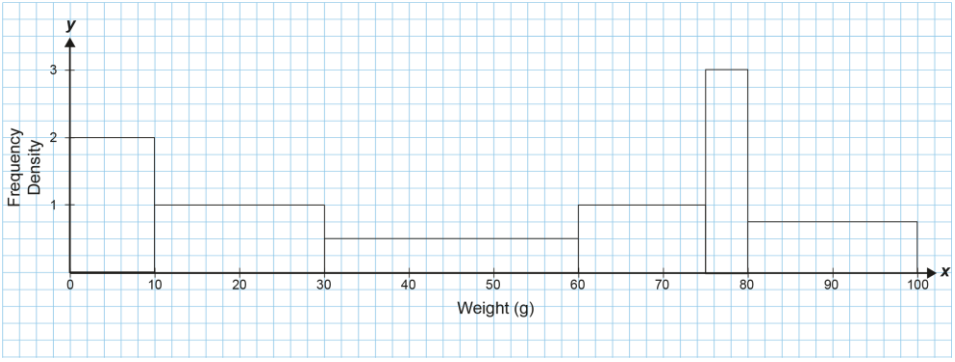
## 10.2.1.8 Interpret and use features of data from a histogram

### Common difficulties and misconceptions

Students often associate histograms with bar charts, because of the similar nature of these two representations, and a common misconception is that the height of the bars in a histogram represents frequency. While this will be true when representing continuous data grouped into equal classes of width one (as frequency density will equal frequency), histograms are more commonly used for data grouped into unequal class widths, and frequency is identifiable from the areas of the bars.

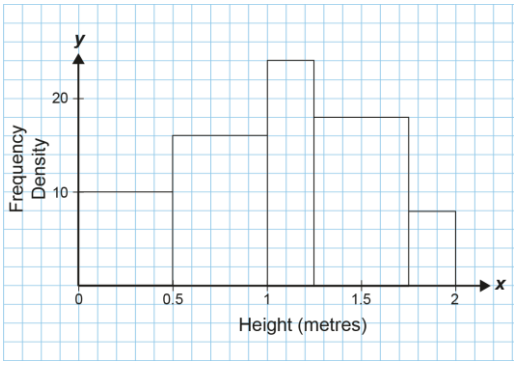
If the left side of a histogram resembles a mirror image of the right side, the distribution of the data is said to be symmetrical. When skewed data are represented on a histogram, students often confuse left (negative) and right (positive) skew as the peak of the histogram veers to the opposite side to which it is skewed, which can be counterintuitive.

| Students need to | Guidance, discussion points and prompts |
|---|---|
| **Understand what the heights of bars in a histogram represent**<br>*Example 1:*<br>*The histogram below shows the weights in grams of 100 letters.*<br>*How many letters had a weight of between 75 and 80 g? Choose from:*<br>*A 3*<br>*B 24*<br>*C 15*<br>*D 5*<br>*Explain how you know.* | *Example 1* aims to expose some common misconceptions students have when interpreting histograms, by providing alternative options that make these mistakes identifiable. Exploring possible misconceptions supports students in **deepening** their understanding of the features of histograms.<br>Teachers should consider how students have misinterpreted the **representation** to generate possible incorrect answers. The most common wrong answer is likely to be 3 (option A), where students read off the height of the bar. Establish that the $y$-axis represents the frequency density, which is the frequency per gram for the data in the class the bar represents. To find the frequency, we need to multiply the frequency density by the class width. Students may select option B if they have counted the squares contained within the bar, or option D if they are confusing the number of letters with the class width.<br>Multiple-choice questions can be a helpful tool to anticipate, expose and correct possible misconceptions. Designing multiple choice questions can also be a productive professional development activity. What prompts could be given to students who have selected options B or D, instead of the correct answer of C, 15, to help them to recognise their error? Are there any other possible wrong answers students might give? |



*Example 2:*

*The histogram shows the heights of some children.*




*a) How many children are there in total?*

*b) Explain how you know.*

*Example 2* moves beyond interpreting what one bar in a histogram represents, to **deepening** thinking about the total frequency. The heights have been described in metres rather than centimetres, so that the frequency densities are values that could feasibly be frequencies (when described in centimetres the frequency densities are 0.1, 0.16, 0.24, 0.18 and 0.08), making it possible to establish whether students are interpreting the heights of the bars as frequencies or not.

Incorrect answers can reveal where students have gaps in their understanding of the **representation**. For example, if students are still interpreting the $y$-axis values as frequencies rather than frequency densities, they might give an answer of 76. An answer of 38 might suggest students used the class width of 0.5 for all bars, rather than just the first two; assuming equal class widths suggests that students may not recognise how the construction of histograms is more appropriate when data are grouped into unequal, rather than equal class widths.

Discuss with colleagues what other incorrect answers students might possibly give, and what these answers might suggest about students' understanding of histograms.

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

*Example 3:*

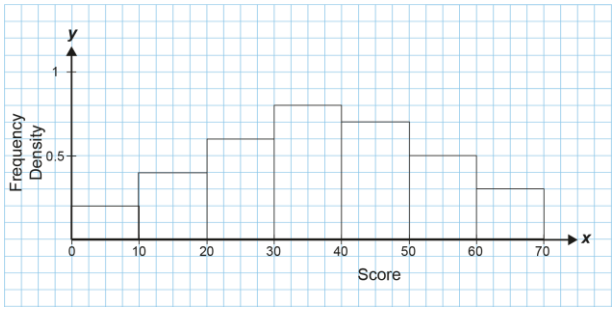
*The histogram below represents a class's scores on a test.*

*What can you say about the average score and distribution of test results for the class?*

*Example 3* focuses on the information that can be identified from a histogram, making links with measures of central tendency. The distinction between the 'mode' and 'modal class' is an important one to make with students, to ensure that they use the correct **language** when describing the most common values. The most common test scores for the class were between 30 and 40 marks (i.e., the modal class, and highest bar, is 30-40 marks).

Thinking about the distribution can support in **deepening** students' understanding of how the measures of central tendency relate to the data set. When we have a symmetrical histogram, whilst we cannot identify the exact values of the mean and median values, we know that they must be similar and will lie within the modal class (located at the centre of the distribution).

Discuss some possible prompts that will challenge students further. For example, *'Is it possible for the most common test score to be a value that is not situated within the modal class?'*



*Example 4:*

*The histogram below this example represents a class's scores on a test.*

*Which statements are true for this histogram? Choose from*

*A The mean and median have a similar value.*

*B The mean is greater than the median.*

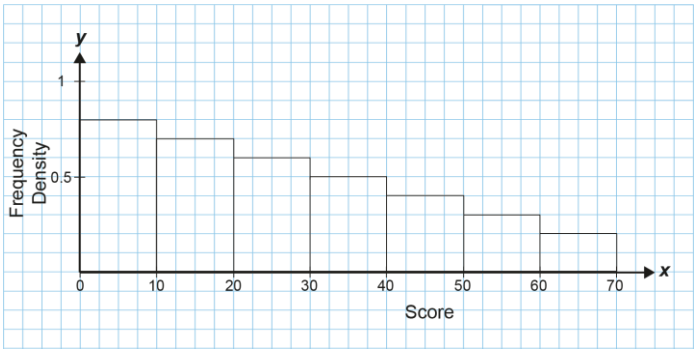
*C The mean is less than the median.*

*Explain how you know.*

*Example 4* introduces a skewed distribution and explores the effect a positive skew has on the relationship between the three measures of central tendency. Students will be able to identify the modal class as the tallest bar (0-10) from the histogram, but should recognise that it is not possible to determine the exact values of the mean and median from a histogram **representation**.

Students are likely to assume that the skewed nature of the data implies that the mean and median will not be similar. To establish which measure will be greater, they need to recognise the effect that extreme values have; understand that they have more of an effect on the mean than the median; and understand why this is. This example therefore offers an opportunity for further **deepening** their understanding of averages in relation to the data set. Can they explain why this distribution is described as a positive rather than a negative skew?

Discuss some possible follow-up activities that students could be asked to do, once they have correctly identified that the mean is greater than the median for positively-skewed data. For example, constructing a data set that satisfies the features of the histogram, and use this to determine the mean and median.



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

*Example 5:*

*The mean scores on a science test are the same for two different classes.*

*The results are plotted on two histograms, shown below.*

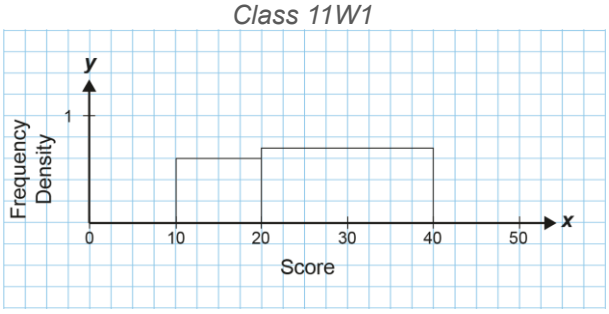
*Explain how it is possible for the histograms to look so different when the mean score for the two classes are the same.*

In *Example 5*, the **variation** between the two data sets is such that the means are the same, but the distributions are different. This highlights the limitations of what a single summary statistic can tell us about a data set. While the mean provides useful information about a data set, it summarises the data as a single numerical statistic and tells us nothing about the shape of the distribution. Representing the data on a histogram shows which scores are more common or less common, as well as displaying the dispersion of values within the data set.

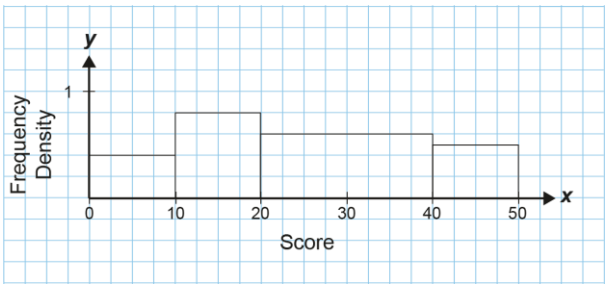
The mean scores for classes 11W1 and 11W3 are the same, but the histograms show that the spread of scores are very different. The scores for class 11W1 lie between 10 and 40 marks; for class 11W3 the range is 0-50. There is an opportunity for **deepening** students' understanding of the concept of the mean, by supporting them to identify and explain how this scenario is possible. The total number of students in each class can be determined from the histograms. We know that six students in class 11W1 scored between 10 and 20 marks and 14 students scored between 20 and 40 marks, so the total number of students in class 11W1 is 20. Similarly, we can determine that there are 30 students in 11W3. For the means to be the same, we know that the total marks scored on the test by each class, divided by the number of students in the class, must have the same value. Recognising that this is feasible, and creating an example of data sets where this is the case to confirm it, provides an important insight into the value of using statistical measures and representations simultaneously to gain a better understanding of the features of a data set.

Discuss possible data sets that could be represented by the two histograms. Consider which values can be varied without affecting the histogram representation or resulting in a change in the mean. Reflect on your thinking during this activity. What knowledge or representations did you need to draw upon? Has your understanding of the mean shifted at all?

*Class 11W1*




*Class 11W3*

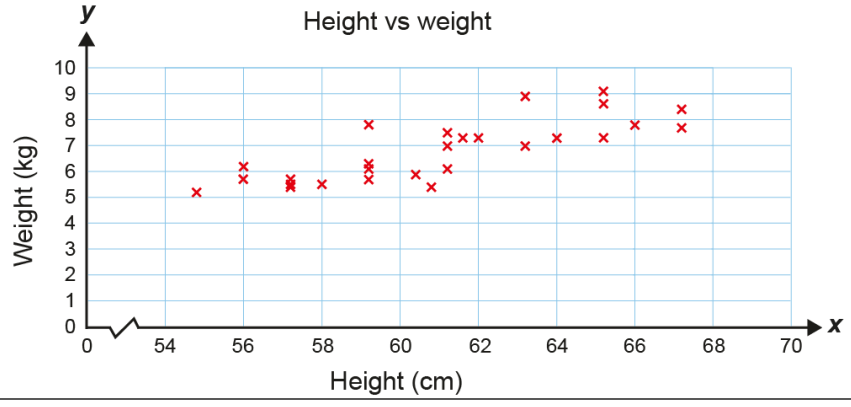


## 10.2.3.2 Understand that correlation alone does not indicate causation

### Common difficulties and misconceptions

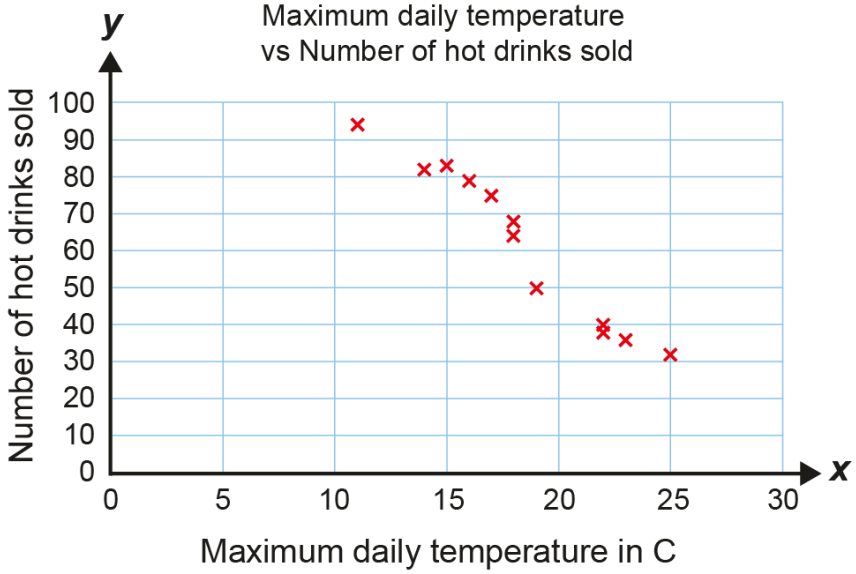
Scatter graphs are a good way of displaying bivariate data to determine whether or not there is correlation, the existence of a relationship between two variables. However, observing a relationship between two variables in a scatter graph does not mean that changes in one variable are responsible for changes in the other. It is important that students recognise the difference between correlation, the existence of a relationship between variables; and causation, where one variable causes an effect on another. Recognising that there may be other causal variables associated with both correlated variables, and being able to identify what these might be, is key to students developing a deeper understanding of the difference between correlation and causation.

| Students need to | Guidance, discussion points and prompts |
|---|---|
| **Understand the difference between correlation and causation**<br><br>*Example 1:*<br><br>*A health visitor measures babies' heights and weights at a clinic. She plots these values on a scatter graph, shown below.*<br><br>*Which of the following statements best describes the relationship between the babies' heights and weights? Choose from:*<br><br>*A No correlation*<br><br>*B Correlation but no causation*<br><br>*C Correlation and causation*<br><br>*Explain your choice.* | In *Example 1*, students explore the relationship between height and weight for a group of babies, **deepening** thinking about what it means for two variables to be correlated, and how this differs to one variable *causing* a change in the other.<br><br>The **variation** between the three options draws students' attention to the fact that correlation does not necessarily imply causation. Students may assume that because the points on the scatter graph are not in a straight line there is no correlation, even though the pattern of points suggests that there is a positive correlation between height and weight. It is important that students can describe what this means, i.e., that, in general, heavier babies are also taller, but that weight does not cause height. For causation to be present, an increase in a baby's weight would **result** in an increase in their height. Students may argue that, if a baby grows in weight, their bones have become bigger and so they will also be taller. However, there are many other factors that can affect a baby's height and weight, including genetics and feeding issues. This means that, although there is a correlation between the two variables, it does not imply causation.<br><br>Discuss possible misconceptions about correlation and causation, and how students' own varying life experiences will contribute to their perceptions in different contexts. What strategies might be helpful for developing students' thinking? |

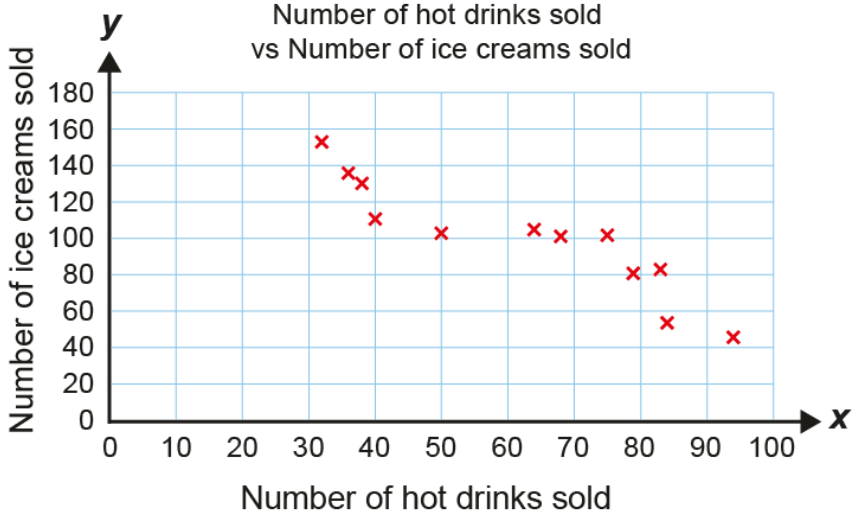


| | |
|---|---|
| *Example 2:*<br><br>Describe the possible relationship between the following pairs of variables, commenting on possible correlation and causation.<br><br>a) *The number of ice creams sold and the number of air-conditioning units sold*<br><br>b) *Shirt size and shoe size*<br><br>c) *The number of hours a person works and the amount they get paid*<br><br>d) *The speed of a train and the time it takes to travel 30km*<br><br>e) *The price of a product and the number of products sold* | In *Example 2*, students explore five different relationships in terms of possible correlation and causation. It uses worded statements, rather than a **representation** such as a scatter graph. This is to ensure that students think carefully about the possible relationship between pairs of variables, and hypothesise about what might happen to one variable as the other one increases/decreases, without the support of a graphically-represented data set to explore.<br><br>When discussing the scenarios with students, use **language** that supports them in recognising the presence of correlation and causation. This may involve supporting them to recognise factors in real-life scenarios that they are potentially not familiar with. For example, ask:<br><br>• *'Will an increase in the number of ice creams sold, **cause** there to be an increase in the number of air-conditioning units sold?'*<br><br>• *'Could the relationship between these variables be a coincidence?'*<br><br>• *'Will there always be causation and correlation between hours worked and amount paid? How about if someone is paid a fixed salary with no overtime?"* |
| **Understand that a causal variable may be present and be able to identify it**<br><br>*Example 3:*<br><br>*A takeaway café is open from Monday to Saturday and sells ice creams and hot drinks. The owner of the café records the maximum daily temperature and the number of ice creams and hot drinks sold over a two-week period. The results are represented in three scatter graphs, shown below.*<br><br>*Explain why Graph B is the odd one out.* | *Example 3* may initially be counterintuitive to students, as Graph C (being the only one with positive correlation) seems potentially a more obvious odd one out. Asking students to describe why Graph B is the odd one out, rather than to identify one graph as not belonging to the group, provides them with an opportunity for **deepening** their thinking about the relationships between the three variables represented in the scatter graphs.<br><br>Precision with **language** is important in helping to build a nuanced understanding of causation and correlation. While it may not be possible to prove causation, it can be argued that an increase in temperature may cause an increase in the number of ice creams sold and a decrease in the number of hot drinks sold. Temperature can therefore be identified as being a causal variable. In graph B, while a negative correlation can be identified, we cannot conclude that an increase in the number of hot drinks sold **causes** there to be a decrease in the number of ice creams sold. Graph B, therefore, is the only graph that suggests correlation only.<br><br>  Discuss possible prompts to support students when exploring the three different relationships represented in Graphs A, B and C. For example:<br><br>• *'How are Graphs A and B the same and how are they different?'*<br><br>• *'Which graph(s) are the most useful to the café owner?'* |

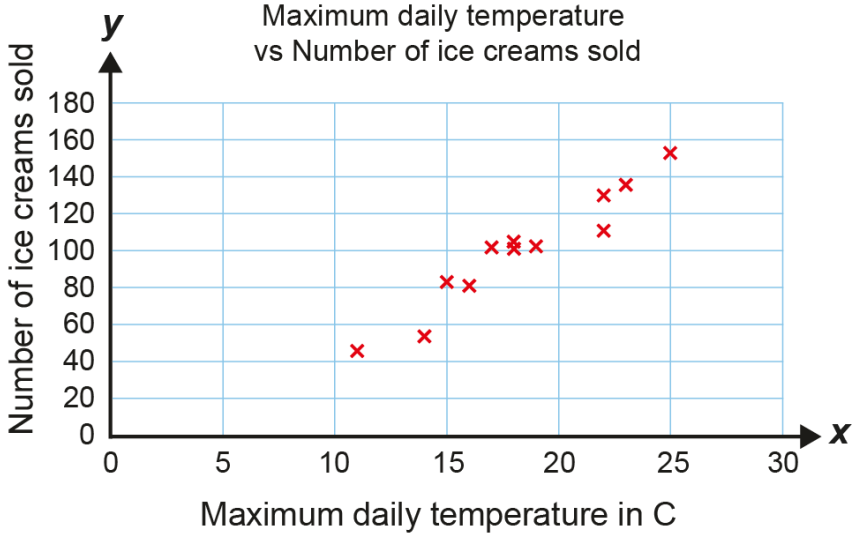
*Graph A*




*Graph B*




*Graph C*



## Collaborative planning

Although they may provoke thought if read and worked on individually, the materials are best worked on with others as part of a **collaborative professional development** activity based around planning lessons and sequences of lessons.

If being used in this way, is important to stress that they are not intended as a lesson-by-lesson scheme of work. In particular, there is no suggestion that each key idea represents a lesson. Rather, the fine-grained distinctions offered in the key ideas are intended to help you think about the learning journey, irrespective of the number of lessons taught. Not all key ideas are of equal weight. The amount of classroom time required for them to be mastered will vary. Each step is a noteworthy contribution to the statement of knowledge, skills and understanding with which it is associated.

Some of the key ideas have been extensively exemplified in the guidance documents. These exemplifications are provided so that you can use them directly in your own teaching but also so that you can critique, modify and add to them as part of any collaborative planning that you do as a department. The exemplification is intended to be a starting point to catalyse further thought rather than a finished 'product'.

A number of different scenarios are possible when using the materials. You could:

- Consider a collection of key ideas within a core concept and how the teaching of these translates into lessons. Discuss what range of examples you will want to include within each lesson to ensure that enough attention is paid to each step, but also that the connections between them and the overall concepts binding them are not lost.
- Choose a topic you are going to teach and discuss with colleagues the suggested examples and guidance. Then plan a lesson or sequence of lessons together.
- Look at a section of your scheme of work that you wish to develop and use the materials to help you to re-draft it.
- Try some of the examples together in a departmental meeting. Discuss the guidance and use the PD prompts where they are given to support your own professional development.
- Take a key idea that is not exemplified and plan your own examples and guidance using the template available at Resources for teachers using the mastery materials | NCETM.

Remember, the intention of these PD materials is to provoke thought and raise questions rather than to offer a set of instructions.

## Solutions

Solutions for all the examples from *Theme 10 Statistics and probability* can be found here:
https://www.ncetm.org.uk/media/r2idmejd/ncetm_ks4_cc_10_solutions.pdf

## Data sources

1 Office For National Statistics (2023). *Income estimates for small areas, England and Wales - Office for National Statistics*. [online] www.ons.gov.uk. Available at: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/datasets/smallareaincomeestimatesformiddlelayersuperoutputareasenglandandwales.