10 Statistics and probability

Mastery Professional Development

10.1 Statistical measures 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.

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 *10.1 Statistical measures and analysis* into two statements of **knowledge, skills and understanding**:

10.1 Statistical measures and analysis

- 10.1.1 Understand and accurately calculate measures of central tendency and spread
- 10.1.2 Infer properties of a population from a sample

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

- 10.1.1 Understand and accurately calculate measures of central tendency and spread
  - 10.1.1.1 Understand that a data set can be represented in a frequency or a grouped frequency table
  - 10.1.1.2 Calculate measures of spread and central tendency for a data set represented in a frequency table
  - 10.1.1.3 Understand that the mean and range calculated from a data set represented in a grouped frequency table will be an estimate
  - 10.1.1.4 Estimate the mean and range for a data set represented in a grouped frequency table
  - 10.1.1.5 Locate the median and the modal class for a data set represented in a grouped frequency table
  - 10.1.1.6 Understand and calculate the interquartile range as a measure of spread
- 10.1.2 Infer properties of a population from a sample
  - 10.1.2.1 Understand how sampling can be used to gather information
  - 10.1.2.2 Understand the limitations of data collection methods
  - 10.1.2.3 Use the information from a sample to infer properties of the whole population
  - 10.1.2.4 Describe the properties of a population using appropriate summary measures

Overview

Statistical literacy will be of great importance to students throughout their lives. They will need to interpret statistical data when making life decisions, and statistics are an important lens through which to view the world. The focus of this core concept extends beyond accurate calculation of statistical measures, to interpreting statistics and developing a deeper understanding of different measures. This includes what the mean, median and mode tell us about the data and when each should be used. Emphasis is placed on how sampling can be used to gain insight into the whole population and the importance of interpreting statistical measures, based on an understanding of how they summarise the data, and may be distorted by the data they represent.

From Key Stage 3, students should understand the limitations of different measures of central tendency based on the nature of the data set. The mean includes all the values in the data set for its calculation and is the most frequently used measure of central tendency. It is important to recognise that it does not give a good measure of central tendency when there are outliers, or the data are skewed. Similarly, it is important to appreciate that the median is not affected by extreme values, so may be a better measure of central tendency when the data are very skewed. This supports students' awareness of the underlying distribution of a data set. The mode is unique, as it is the only measure that can be used for categorical data. It is important that students recognise why there can be more than one mode or no mode at all for a particular data set, and why this may contribute to the mode being the least used measure of central tendency. A key skill for students to develop further at Key Stage 4 is the ability to make an informed choice of appropriate statistical measures for a given data set.

When considering data sets with extreme outliers at Key Stage 3, students may have explored measuring the spread of the data by ignoring the extreme values and examining the distribution of the rest of the data. This foundation is built on at Key Stage 4, with the introduction of the interquartile range, which provides a more helpful measure of dispersion when considering data that contain outliers. Recognising the importance of having both a measure of central tendency and a measure of spread, to appreciate the distribution of a data set, is key to students' understanding of statistical measures.

At Key Stage 4, students develop their understanding of statistical measures to working with data that are presented in frequency tables, where the data set increases beyond simple calculations being possible from a list. Grouped frequency tables provide further progression, and it is important for students to realise the impact of working with grouped data: the information that is lost due to the way the data have been recorded (and/or collected) and the effect that working with grouped data has on the three measures of central tendency and the range.

When working with data at Key Stage 4, students need to understand how sampling can be used to gather information; recognise that a sample is a selection of data from a larger group of data, the population; and understand what the statistical term 'population' means. The importance of the sample being representative of the population – and the implications, if this is not the case – is fundamental to students developing an understanding of the assumptions that are made when using information from a sample to infer properties of the whole population.

Throughout this topic, students need to be familiar with the contexts and samples under study. One way to support this is to use real data from their lives. While this can make the data messier, it helps to reinforce the idea that statistics can be applied to real-world examples, and allows students to develop unfamiliar skills within familiar contexts. This can be supported with data that have been carefully chosen to expose different aspects of the mathematics.

Prior learning

At Key Stage 2, students learnt how to calculate the mean. The concept of central tendency was then developed at Key Stage 3 to include the mode and median, and the range was introduced as a measure of spread. Students should have not only calculated measures of central tendency but also interpreted the results of statistical analysis and understood why the three measures often have different values; it is important to check that this knowledge is secure, and Key Stage 4 can be a valuable opportunity to deepen understanding of averages. Students will have encountered just one measure of spread: the

range. It is important to check that students understand the range to be distinct from the three averages, and that it gives us different information about a data set.

A key area of learning in this core concept document is working with data presented in frequency tables. Students will have organised data into tables as far back as Key Stages 1 and 2, and will have continued to do so across different subjects within Key Stage 3. In mathematics at Key Stage 3, they will have begun to calculate averages from some data organised into frequency tables, but not yet from frequency tables where the data are grouped.

Statistical analysis provides a context in which students need to apply their existing knowledge of fractions, percentages and decimals, which will have been gradually built up from their earliest experiences of number. From the beginning of Key Stage 2 onwards, they will have started to use decimals and fractions interchangeably, with percentages also being used later in Key Stage 2. Any gaps in students' understanding of multiplicative relationships, and particularly of how percentages can be used to describe proportions of a whole, will affect the validity of their interpretation of statistical measures at Key Stage 4.

The core concept documents '*3.1 Understanding multiplicative relationships'*, '*5.1 Statistical representations and measures'* and '*5.2 Statistical analysis'* from the Key Stage 3 PD materials explores the prior knowledge required for this core concept in more depth.

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.

<table>
<tr><th>Reference</th><th>Activity</th></tr>
<tr><td>Secondary Assessment materials page 51</td><td>From 7th March 2016 to 7th March 2017 a swimming club had the same members. Complete the table to show the information about the ages of members of the club. Explain your reasoning.
<table>
<tr><th>Measure</th><th>Value</th></tr>
<tr><td>Mean (March 2016)</td><td>14 years 7 months</td></tr>
<tr><td>Range (March 2016)</td><td>4 years 2 months</td></tr>
<tr><td>Mean (March 2017)</td><td>?</td></tr>
<tr><td>Range (March 2017)</td><td>?</td></tr>
</table>
(Based on QCA, 2002)</td></tr>
</table>

<table>
<tr>
<td>Key Stage 3 PD materials document ‘5.1 Statistical representations and measures’, Key idea 5.1.1.1, Example 6</td>
<td>Students were asked to calculate the mean number of goals scored in a set of matches, as recorded in this table.
<table>
<tr><th>Goals</th><th>Matches</th></tr>
<tr><td>1</td><td>3</td></tr>
<tr><td>2</td><td>5</td></tr>
<tr><td>4</td><td>2</td></tr>
<tr><td>5</td><td>3</td></tr>
<tr><td>7</td><td>2</td></tr>
</table>
One student described their thinking, ‘I calculated 50 divided by five because the total number of goals is 50 and there are five items.’
What has this student misunderstood?

Another student stated, ‘There are 19 goals and 15 matches, so I worked out the goals divided by matches, 19 divided by 15.’
Do you agree with this student’s method?

What advice would you give the students to ensure they calculate the mean from a frequency table correctly in future?</td>
</tr>
<tr>
<td>Key Stage 3 PD materials document ‘5.2 Statistical analysis’, Key idea 5.2.1.5, Example 5</td>
<td>The test scores for two classes are summarised in this table.
<table>
<tr><th></th><th>Class A</th><th>Class B</th></tr>
<tr><td>Mean</td><td>70</td><td>70</td></tr>
<tr><td>Mode</td><td>70</td><td>70</td></tr>
<tr><td>Median</td><td>70</td><td>70</td></tr>
<tr><td>Range</td><td>15</td><td>10</td></tr>
</table>
Archie thinks some students in class A have scored higher than class B because it has a larger range.

Do you agree? Explain your answer.</td>
</tr>
<tr>
<td>Key Stage 3 PD materials document ‘5.2 Statistical analysis’, Key idea 5.2.2.2, Example 1</td>
<td>Consider the following data sets on property for a given geographical area.

Average type of property
Average number of bedrooms
Average property price

For which data set would the mode be an appropriate average to use?

For which data set would the mean be an appropriate average to use?</td>
</tr>
</table>

Key vocabulary

Key terms used in Key Stage 3 materials

- (Arithmetic) Mean
- Measure of central tendency
- Median
- Mode
- Outlier
- Range

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
interquartile range	A measure of spread that excludes the lowest and highest 25% of values. It is calculated by subtracting the lower quartile from the upper quartile.
quartile	The values that divide an ordered data set into four equal parts, with a quarter of the population in each part. For example, the lower quartile (Q_1) is the median of the lower half of the data, and the upper quartile (Q_3) is the median of the upper half of the data.
population	The entirety of a group that is being described. All members of a population will share at least one characteristic that means that they belong to that group.
sample	A subset of a population. In handling data, a sample of observations may be made from which to draw inferences about a larger population.

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.

10.1.1 Understand and accurately calculate measures of central tendency and spread

At Key Stage 3, students constructed and interpreted frequency tables. When working with categorical data it is often easy for students to interpret frequency tables correctly. With quantitative data in particular, students may find interpreting the frequency table a less intuitive process. It is important that students are aware of how the data are represented and what each element of the representation means in the context of the data. Teachers should find opportunities to discuss the representation and how that relates to the underlying data. Asking students to write out the full data set can help them to appreciate what is represented and how they might calculate statistical measures of the data. For example, identifying the raw data '1, 1, 1, 1, 1, 1, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 5' from this table, found in *Example 6* of 10.1.1.4 below:

Number of stalls visited	*Frequency*
1	*6*
2	*0*
3	*8*
4	*5*
5	*1*

It is especially important to teach statistical inference conceptually as well as procedurally. Students who have not deeply understood their prior learning may use mnemonics to help recall how to calculate averages, such as 'add and divide for the mean', and many might recall that the mean involves division by 'the number of things that there are'. When presented with data in frequency tables, this could mean that students divide by the number of rows in the table, rather than the total number of values in the data set.

While this mistake can often be identified by interpreting the value of the mean in the context of the question, developing students' conceptual understanding of the mean supports them in identifying 'why' and not just 'how' we calculate this commonly-used measure of central tendency.

Students often find it difficult to calculate the median from data presented in a frequency table. As with other statistical measures, referring back to the data under study can be helpful. While the mode is more easily identifiable, from the row with the highest frequency, students can sometimes mistakenly identify the mode as the frequency itself, rather than the value it represents. It is important that students are given the opportunity to develop a deep understanding of these concepts with ungrouped data, as this will support them in progressing to estimating statistical measures for grouped data, a key focus at Key Stage 4.

Exploring the relationships between the three different measures of central tendency is important in moving students from being able to calculate the measures, to understanding what they each represent. For example, recognising that the mean and median are not the same value when the data are skewed, and why this is, provides the foundation for further exploration of the shape of distributions in future statistical studies.

An acknowledgement that the range can often be an unreliable measure of spread, as it is based on just two values and tells us little about the distribution of the rest of the data, provides a basis for the introduction of an improved measure of variability – the interquartile range. Students need to understand that the median represents a value for which 50% of the remaining data are less than, and 50% of the

data are greater than, that value. This could then be compared with the interquartile range, which describes how spread out the middle 50% of the data are. This can support them in identifying the range and interquartile range as measures of spread and not measures of central tendency, a common confusion.

10.1.1.1 Understand that a data set can be represented in a frequency or a grouped frequency table

10.1.1.2 Calculate measures of spread and central tendency for a data set represented in a frequency table

10.1.1.3 Understand that the mean and range calculated from a data set represented in a grouped frequency table will be an estimate

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

10.1.1.5 Locate the median and the modal class for a data set represented in a grouped frequency table

10.1.1.6 Understand and calculate the interquartile range as a measure of spread

10.1.2 Infer properties of a population from a sample

Students are likely to have had some informal experience of random sampling prior to Key Stage 4 – for example, picking names out of a hat, or using a random number generator. However, at Key Stage 4, students' understanding of sampling is developed to include statistical sampling, where a subset (sample) of a larger population is used to make inferences about the characteristics of that population. It is important that students recognise the importance of a sample being representative of the whole population, and the impact that an unrepresentative sample may have on the conclusions drawn.

The method of sampling used depends on various factors, and students' familiarity with simple random sampling provides a foundation to explore systematic sampling and, with further study, stratified sampling. Recognising the pros and cons of gathering information from a sample versus the whole population, is fundamental to understanding the limitations of some data collection methods and how a particular selection method may result in bias. Assuming that a sample represents the population, when the sampling method does not support this, can result in inaccurate conclusions being made. It is important that students have a deep understanding of how the sample relates to the population.

Understanding the effects that different characteristics of the data have on measures of central tendency and spread and recognising the uncertainty with which some inferences are made, can help students to draw measured conclusions. Encouraging them to engage fully with the context of the data and to consider possible causes for certain characteristics of the summary measures, is key to understanding the extent to which measures of central tendency and spread can be used to make inferences about a particular population.

10.1.2.1 Understand how sampling can be used to gather information

10.1.2.2 Understand the limitations of data collection methods

10.1.2.3 Use the information from a sample to infer properties of the whole population

10.1.2.4 Describe the properties of a population using appropriate summary measures

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.

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

Common difficulties and misconceptions

When grouping data to produce a grouped frequency table, inequalities are often used with continuous data. Students will have been introduced to this notation in primary school and begun to use it denote a range in Key Stage 3. Students often struggle to recognise when to use inequality notation and when it is not needed. Exploring different types of data when constructing grouped frequency tables can help and support them towards a deeper understanding of discrete and continuous data.

When working with grouped data, the midpoint provides an estimate of the values within each group. The process for finding the midpoint of a class interval is straightforward, but it is important that students understand why the midpoint is used when calculating the mean from a grouped frequency table. If the data are evenly distributed across a class interval, the frequency multiplied by the midpoint will have the same value as the sum of the values of the data. When the individual data values are not given, the estimate given by multiplying the frequency by the midpoint is the best approximation that can be made, which we then use to estimate the mean.

A common mistake is to divide by the number of groups the data have been arranged into, rather than the total frequency. This error is often identified when comparing the value obtained for the mean with the values within the data set. However, students need to understand how finding the mean from a grouped frequency table relates to calculating the mean from lists of raw data, in order to promote conceptual understanding. Teachers can support this understanding by drawing students' attention back to the data and context.

Students often learn the process for calculating the mean from data presented in a grouped frequency table, without fully understanding why the steps result in an estimate for the mean. There may be times

when the raw data are available but have been grouped to make statistical analysis easier. However, grouped data are often collected directly, especially when sensitive information is being collected (for example, asking someone to give the age range within which their age lies, rather than their exact age).

Recognising that the range is also an estimate may be more intuitive for students. However, for both the range and the mean, it is important that they understand the conditions for which the mean and range are true measures of central tendency and spread for the raw data.

Students need to	Guidance, discussion points and prompts
Interpret a grouped frequency table *Example 1:* *The table below shows the number of hours worked by children aged 7-14 in various countries in a representative week, according to 'Our World in Data'[1].* ***Number of hours worked*** / ***Frequency*** $1 \leq h < 5$ / *4* $5 \leq h < 10$ / *5* $10 \leq h < 15$ / *11* $15 \leq h < 20$ / *6* $20 \leq h < 25$ / *1* $25 \leq h < 30$ / *0* $30 \leq h < 35$ / *1* *For each of the statements below, state whether it is true, false, or if there is not enough information to say.* *a) The sample size is 7.* *b) Only countries where children worked for one hour or more were included in the sample.* *c) The highest number of hours worked was 34.* *d) In more than one country, children worked for 11 hours.* *e) Most children worked between 10 and 15 hours.* *f) There were no countries where children worked 25 hours.* *g) Children who worked more hours attended school less.*	*Example 1* explores a grouped frequency table with genuine data. Students should work with and draw inferences from real-world data, including data sets that they have little experience of. This example can be used both for checking understanding of frequency tables, and also for **deepening** understanding of what inferences can and cannot be made from them. The **language** in this example is carefully chosen to elicit misconceptions. Some questions check understanding of mathematical vocabulary. For example, in part a, students need to know that they should sum the frequencies to find the sample size. Others check that students understand the meaning of common language in this context. The statement in part e, for example, references 'most children'. It is true that $10 \leq h < 15$ is the modal group, but we cannot comment on how many children this is (and most countries are not in this category). Without knowing the child population of each country, we cannot know where 'most' children are within this distribution.  Do your students have enough opportunity to explore real-world data and understand what conclusions they can draw from it? What data sets might provoke the most interest or promote the most mathematical discussions? Are there any data sets that you would avoid due to their complexity or the potential for controversy? Work with your team to reflect on the data that students are currently asked to work with during their time at secondary school. Below are some suggestions for departmental activities. • Ask each member of the team to share a set of data that they think might be interesting to use in the classroom and explain why this is the case. • Share a data set with colleagues and ask them to work in pairs or small groups to generate questions. Compare and contrast the sets of questions that each group creates, then refine them for classroom use. • Select a particular learning point or misconception that you wish to address. Divide the department into two groups. Ask one group to generate an artificial data set to demonstrate the point and the other to source a real-life data set for the same purpose. Compare and contrast the data sets: which has more potential use for working with students? Why?

Example 2:

Wisal works for a charity that makes wigs. She records the lengths of the hair donations that the charity receives in a week.

a) *Which of the tables shown below could she use to collect her data?*

b) *One donation comes in that is 27 inches long. Wisal thinks she could just add a '20+ inches' category. Do you agree? Why or why not?*

The **variation** in *Example 2* is designed to expose misconceptions about how continuous data can be grouped. This is a key understanding that needs to be secure before students can begin to use grouped data to estimate values. The six groupings offered should generate discussion, as there is not a single right answer – although there are some definite wrong answers! Draw attention to the subtle difference between options A and B and ask students why rounding the length means that B is a viable grouping, while A is not.

Inequality notation forms part of the mathematical **language** that students need to be fluent with when working with continuous data that are grouped. Careful comparison of options C, D and E, including identifying any gaps or overlaps, is essential. Reference could also be made to how overlapping groups in option F are not useful in this context, but they do form the basis for a cumulative frequency table.

A

Length (inches)
0-5
6-10
11-15
16-20

B

Length (to the nearest inch)
0-5
6-10
11-15
16-20

C

Length (h) in inches
$0 \leq h < 5$
$6 \leq h < 10$
$11 \leq h < 15$
$16 \leq h < 20$

D

Length (h) in inches
$0 \leq h < 5$
$5 \leq h < 10$
$10 \leq h < 15$
$15 \leq h < 20$

E

Length (h) in inches
$0 \leq h \leq 5$
$5 \leq h \leq 10$
$10 \leq h \leq 15$
$15 \leq h \leq 20$

F

Length (inches)
Up to 5
Up to 10
Up to 15
Up to 20

Example 3:

The results on a maths test were recorded for Class 10W3.

9, 11, 14, 15, 16, 16, 18, 19, 20, 24, 26, 26, 27, 28, 30, 31, 31, 34, 35, 36, 37, 38, 38, 40, 42, 43, 43, 45, 46, 49

The test had a possible total of 50 marks.

Miss Cauchy recorded the test scores in the following table.

Mark	Frequency
0–10	1
11–20	8
21–30	6
31–40	9
41–50	6

Mr Schwarz recorded the test scores in a different table:

Score (s)	Frequency
$0 < s \leq 10$	1
$10 < s \leq 20$	8
$20 < s \leq 30$	6
$30 < s \leq 40$	9
$40 < s \leq 50$	6

Comment on the groupings used in the two tables.

Example 3 supports with **deepening** students' understanding of grouping discrete data and using inequality notation. When grouping without using inequalities, students often replicate the upper bound of the previous class as the lower bound of the following class (for example, giving 0-10, 10-20, 20-30, etc., for this data) and so it is unclear where these repeated values should be recorded. Grouping without using inequalities is appropriate when discrete data, like these test results, are being represented in a grouped frequency table.

The **variation** in this example involves presenting the same set of data in two subtly different ways. Students may assume that Mr Schwarz's method is better due to the use of inequality notation. Miss Cauchy's grouping method is actually suitable for the test marks listed, but students could be challenged to think about test marks that may result in this method of grouping becoming inadequate. A prompt might be, *'If a student was given a half mark for one of the questions in the test, how would this affect the two types of grouping?'*

Mr Schwarz's first group $0 < s \leq 10$ does not allow for a score of zero to be recorded and so needs revising to $0 \leq s \leq 10$. Discuss whether the choice of inequality use is trivial when grouping data. It is important that students are precise with their **language** when describing inequality notation. They should be encouraged to move beyond describing a class interval as 'between 10 and 20' and make explicit the distinction between 'greater than' and 'less than or equal to'.

Discuss with your team how the data in this example can be used to explore unequal class widths, and the importance of looking at frequencies when determining class intervals. Do students have an opportunity to think about this in your current curriculum plan?

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

Example 4:

The heights of 35 apple trees are measured in metres to the nearest cm.

1.01, 1.17, 1.27, 1.38, 1.46, 1.51, 1.53, 1.64, 1.68, 1.69, 1.78, 1.83, 1.84, 1.88, 1.91. 1.93. 1.97, 2.00, 2.05, 2.10, 2.11, 2.13, 2.14, 2.18, 2.24, 2.28, 2.30, 2.37, 2.42, 2.47, 2.59, 2.64, 2.75, 2.84, 2.95

Complete the upper and lower class limits for grouping the data into 10 groups of equal class widths in the table below:

Height (h)	*Frequency*
$1.0 \leq h \leq$	*2*
$< h \leq$	*2*
$< h \leq$	*3*
$< h \leq$	*4*
$< h \leq$	*7*
$< h \leq$	*6*
$< h \leq$	*4*
$< h \leq$	*3*
$< h \leq$	*2*
$< h \leq 3.0$	*2*

All data explored in *Examples 4*, *7, 8, 11 and 13* relate to the same 35 apple tree heights, with the questions each time choosing a different aspect to highlight for students. Frequency tables are commonly used for large data sets, so students are often presented with data in grouped frequency tables, rather than being asked to group the data themselves. *Example 4* requires students to group a data set into a specified number of groups. This will support them with **deepening** understanding of how the grouped data and the raw data relate. An ordered list of the raw data is provided so students can focus their attention on the grouping, making visible the information that is lost when grouping data.

When discussing the upper and lower class limits, students need to be precise with the **language** that they use. They should refer to heights being 'greater than' or 'greater than or equal to' and 'less than' or 'less than or equal to' the lower and upper class limits.

In this example, students are not required to determine the number of groups for themselves. Discuss the value of asking students to group the data without giving a predetermined number of groups. This would support them to consider the point at which too much information is lost due to inappropriate grouping. What learning might be prompted if you asked:

- *'What would be the optimum number of groups for these data?'*
- *'Is the mean value affected by the number of groups used?'*
- *'What does affect the accuracy of the mean obtained from the frequency table compared with the actual mean calculated from the raw data?'*

Example 5:

A group of students collected data about the ages of 20 people at a school event.

Holly puts the data into a table that she had already prepared. Noel collects the same data and then constructs the table. Both tables are shown below.

a) What is the same and different about their tables?

b) What is the best guess of the range from each of their tables?

The actual ages, recorded as a list, are shown below:

2, 4, 11, 18, 23, 24, 29, 35, 41, 52, 54, 59, 61, 65, 72, 77, 81, 81, 82, 89

While it is important that students work with real-world data, it can be helpful to be selective with the data being used in the early stages of their learning. Curating data sets that help to reveal specific features is a form of **variation** and, in several examples within this key idea, this is used to highlight different features of grouped data. In *Example 5*, the focus is on recognising that, when data are grouped, any calculations with those data then become estimates. There is an opportunity to discuss how the size of the groups impacts upon the accuracy of the estimate.

Students should appreciate that the choices made around **representations** for data can also impact on the way in which those data are then processed. It is particularly important for them to understand that information about the data is lost when grouping. One of the decisions that affects calculations from a grouped frequency table is the choice of upper and lower bounds. It is intentional here that the data look evenly spread when grouped in twenties,

c) *What is the actual range of the data?*

d) *Why is the answer to part c different to the ranges calculated in part b?*

but is much more unevenly distributed when grouped in tens, to draw out this understanding.

Discuss the value of exploring a possible way of regrouping the values in this data set. Would grouping the data in nine class intervals, for example, (2-12, 13-22 and so on) provide any more insight for students? What might be the drawbacks of exploring this unconventional grouping of the data?

Holly's table:

Age group	Tally	Frequency
0–20	IIII	4
21–40	IIII	4
41–60	IIII	4
61–80	IIII	4
81–100	IIII	4
101–120		0
Total		20

Noel's table:

Age group	Tally	Frequency
1–10	II	2
11–20	II	2
21–30	III	3
31–40	I	1
41–50	I	1
51–60	III	3
61–70	II	2
71–80	II	2
81–90	IIII	4
91–100		0
Total		20

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

Example 6:

A school holds a fundraising event with five different stalls and wants to measure how successful the event is. Anoushka and Faris randomly select the same 20 people but collect different data from them.

Anoushka records the number of stalls that each person visits during the event:

Number of stalls visited	*Frequency*
1	*6*
2	*0*
3	*8*
4	*5*
5	*1*

Example 6 is the first in a series that deals more directly with finding the mean from a grouped frequency table. The **variation** here draws attention to the similarities and differences in finding the mean from a table with unique values for each row (which students should have prior experience of), and finding the mean from a table where the data are grouped. Part b checks students have remembered how to find the mean from a table, while parts c and d ask them to consider which values can be selected to best represent the group. It may be that students suggest the midpoint with little need for debate, but subsequent examples explore the reasoning for using the midpoint in more detail.

Part a deals with the misconception that can be unintentionally generated when using two identically-distributed data sets, **deepening** students' reasoning about what we know, and what we can and cannot infer, from a data set. Students might assume that the people referred to in each row are the same, as the frequencies are the same. However, all we know is that (for example) six people visited one stall, and six people stayed between 1 and 10 minutes. We do not know whether or not they are the same six people, and we cannot assume so.

Part e asks students to consider whether the data that Faris and Anoushka have collected serves the purpose they were intended for. This might seem to deviate from the stated focus of this set of key

Faris records the length of time that each person stayed at the event:

Length of time (to the nearest minute)	***Frequency***
1–10	*6*
11–20	*0*
21–30	*8*
31–40	*5*
41–50	*1*

Anoushka notices that only one person visited all five stalls. She says, 'This person also stayed for 41-50 minutes.'

a) Is Anoushka correct? Why or why not?

b) Anoushka adds the numbers in the first column and divides by five. She says that this shows the mean number of stalls visited is 3. Is she correct? Why or why not?

Faris wants to find the mean length of time spent at the event. He thinks he should use the lower value from each group (1, 11, 21, etc.).

c) What might be the problem with using these values to find the mean?

d) What alternative values could he use to represent each group?

e) Have Anoushka and Faris collected useful information to determine if the event was successful? Why or why not?

ideas – finding the mean from a grouped frequency table – but it is important that students' work with data is continually rooted in the purpose of those data. Statistics can easily be reduced in the classroom to a series of calculations or procedures which, without a sense of context or an opportunity to draw conclusions, become meaningless. It could therefore be useful to embed questions such as part e throughout your scheme of learning, referring to other stages of the data collection cycle whenever students work with data. Review your statistics materials with your team: do students have enough opportunity to do this?

Example 7:

a) Find the mean of:

(i) 1.0 and 1.5

(ii) 1.5 and 2.0

(iii) 2.0 and 2.5

(iv) 2.5 and 3.0

Carly uses the data from Example 4 to create the below grouped frequency table. She uses it to find the mean height for an apple tree.

b) What do you notice about Carly's table and your answers to part a?

When presented with a grouped frequency table like the one in *Example 7*, we only know which groups the data is in, not each of the data values (unless the raw data are provided too). We then use the midpoint of each group as our estimate of the values in each group – not because it is convenient, but because it is the mean of the upper and lower class limits. Students must recognise this and understand that we are estimating that each data value is equal to the midpoint, helping them to see this as a sensible model of the unknown data values rather than just a process to be memorised. Some students might struggle to explain why the mean of the lower and upper limits is equal to the midpoint. Asking them to list five data values that satisfy $1.0 \leq h \leq 1.5$ and have a mean of 1.25 might

c) *Explain why Carly has multiplied the midpoint by the frequency.*

d) *What assumption is being made about the heights of the apple trees in each class interval?*

support them in **deepening** their understanding of why this is the case.

Students must be precise with the **language** they use when describing what the midpoint represents. Referring to the 'average' rather than the 'mean' may result in wrongly associating the midpoint of a class interval with finding the median. Students need to recognise the difference between finding the middle value in a set of data and the midpoint between the lower and upper limits of a class interval.

Height (h)	Frequency	Midpoint	Midpoint x frequency
$1.0 < h \leq 1.5$	5	1.25	6.25
$1.5 < h \leq 2.0$	13	1.75	22.75
$2.0 < h \leq 2.5$	12	2.25	27
$2.5 < h \leq 3.0$	5	2.75	13.75

Find the mean from a grouped frequency table

Example 8:

Dominic uses the table below to find the mean height of the apple trees.

Height (h)	Freq.	Mid-point	Midpoint x freq.
$1.0 < h \leq 1.4$	4	1.2	4.8
$1.4 < h \leq 1.8$	7	1.6	11.2
$1.8 < h \leq 2.2$	13	2.0	26
$2.2 < h \leq 2.6$	7	2.4	16.8
$2.6 < h \leq 3.0$	4	2.8	11.2

He performs the following calculation:

$4.8 + 11.2 + 26 + 16.8 + 11.2 = 70$

$70 \div 5 = 14$ metres high

a) *What has Dominic done wrong?*

b) *How would you explain to Dominic the correct way to calculate the mean from this table?*

Example 8 addresses a mistake that students often make when using a frequency table to find the mean. Students can struggle to associate the tabular description of the data with a physical list of values, as the actual data values cannot be determined from the table. As a result, they may lose sight of what the table represents and resort to following a list of steps to find the mean, without considering why they need to complete each stage of the calculation process. Explicitly demonstrating this common mistake – of dividing by the number of rows/groups in the table, rather than the total frequency – helps with **deepening** students' understanding of why the process of calculating the mean from a grouped frequency table is structured in the way that it is.

The grouped frequency table **representation** is commonly used and it is important that students are familiar with it and can engage with it. If they are struggling to work with the table, ask them to create a data set that satisfies the summarised data to help them to visualise the nature of the data presented in the grouped frequency table.

The total frequency has not been included in this grouped frequency table. Discuss with your colleagues the advantages and disadvantages of including, or adding, a row for the totals when working with frequency tables. While a 'total' row can alert students to the total number of values in the data set, they may be tempted to find the total of the midpoint column as well, which could lead to further confusion.

Recognise that calculations from grouped frequency tables will be estimates

Example 9:

Amy, Ben and Callie use the data from Example 5 to calculate the mean age of the sample of people.

Amy uses the raw data and calculates the mean as:

$$(2 + 4 + 11 + 18 + 23 + 24 + 29 + 35 + 41 + 52 + 54 + 59 + 61 + 65 + 72 + 77 + 81 + 81 + 82 + 89) \div 20 =$$

$$960 \div 20 = \mathbf{48}\ \textbf{\textit{years old.}}$$

Ben uses the data from Noel's frequency table, shown below. He calculates the mean as:

$$990 \div 20 = \mathbf{49.5}\ \textbf{\textit{years old.}}$$

Callie uses the data from Holly's frequency table, as shown below. She calculates the mean as:

$$1\,010 \div 20 = \mathbf{50.5}\ \textbf{\textit{years old}}.$$

Why do Amy, Ben and Callie get different results for the mean, when they have all used the same data? Explain your answer fully.

Example 9 makes explicit why the mean calculated from a grouped frequency table is most likely to be different to the actual mean obtained from the raw data. Presenting the raw data in an ordered list, as well in the grouped frequency table **representations** from *Example 5*, exposes the differences between the assumed average value within each of the class intervals (the midpoints) and the actual average calculable from the data.

When calculating the mean from a grouped frequency table, the midpoint values are used in the final column to calculate the total of the estimated data values. If we find the total of the midpoints and divide by the number of groups, what does this represent? In what ways can exploring the significance of the midpoints help with **deepening** understanding of the calculation of the mean? Consider how this may cause confusion and reinforce common misconceptions.

Examples 5 and *9* both use the same set of data, ostensibly showing the ages of people attending a school event – a context also shared with *Example 6*. While the context should be familiar, students may still need support in understanding why it might be useful to collect data about the attendees, and what conclusions might be drawn from it. Discuss with your colleagues how you might support students to understand the practicalities of collecting data and the purposes that data collection might serve.

Ben's calculations using Noel's frequency table:

Age group	Tally	Frequency	Midpoint	Midpoint x frequency
1–10	II	2	5.5	11
11–20	II	2	15.5	31
21–30	III	3	25.5	76.5
31–40	I	1	35.5	35.5
41–50	I	1	45.5	45.5
51–60	III	3	55.5	166.5
61–70	II	2	65.5	131
71–80	II	2	75.5	151
81–90	IIII	4	85.5	342
91–100		0	95.5	0
Total		20		990

Callie's calculations using Holly's frequency table:

Age group	Tally	Frequency	Midpoint	Midpoint x frequency
0–20	IIII	4	10	40
21–40	IIII	4	30.5	122
41–60	IIII	4	50.5	202
61–80	IIII	4	70.5	282
81–100	IIII	4	90.5	362
101–120		0		
Total		20		1008

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

Example 10:

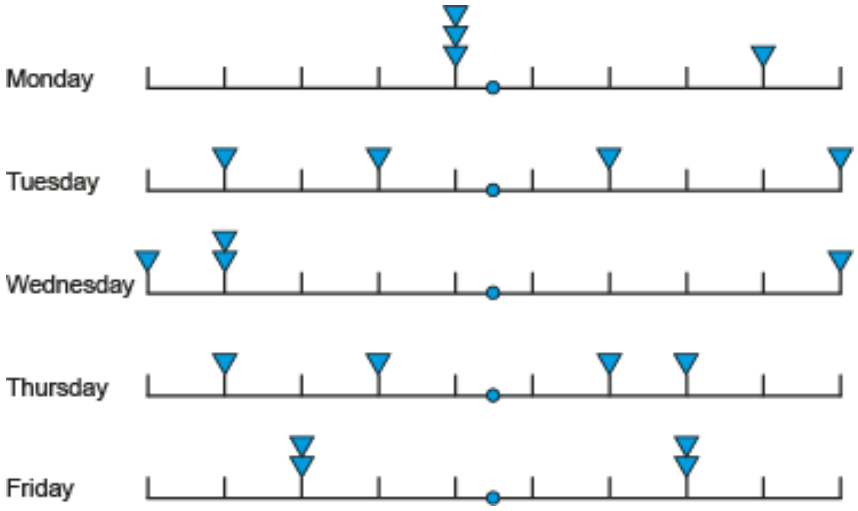
A museum collects data about the ages of its visitors. Age categories with a class width of ten years are used.

*Below are five diagrams showing the distribution of visitors in **one** age category over a week. The midpoint of the category is also shown.*

Comment on the accuracy of using the midpoint to represent the data on each day.

Example 10 asks students to think about the distribution of data within a single group, so that they can reflect on how accurate the midpoint is when used as an estimate of the values within the group. A visual **representation** shows how four imaginary data points might sit around the midpoint of the group. Values are not included on the number lines, so that students are focusing on the shape of the distribution rather than immediately trying to calculate averages. They may find that adding a hypothetical range of ten ages helps them in their interpretation of the task.

For this example to have the intended effect of **deepening** understanding, it is important that teachers relate this task to using the midpoint of an estimate of the mean from a grouped frequency table. This will support students to understand that their estimate of the mean will be more accurate if the data within the group fall in such a way that the mean of the group is close to the midpoint.



Example 11:

The mean height of 35 apple trees is determined by grouping the data in three different ways.

10 groups: mean = 2.003

5 groups: mean = 2.0

4 groups: mean = 1.993

Ernie calculates the mean from the raw data list and gets 2.001.

Comment on the accuracy of the means estimated from the three different grouped frequency tables.

Example 11 once again refers to the data set of 35 apple tree heights, Here, students are asked to compare the estimates for the mean when the data are grouped into 4, 5 and 10 class intervals. They may assume that when the data is grouped into more class intervals, the mean will be a more accurate estimate. Whilst this is often true, *Example 11* shows that this is not *always* the case, **deepening** students' understanding of what determines the accuracy of an estimate for the mean.

Discuss the value of asking students to create a list of 35 tree heights that will result in the estimated mean from a grouped frequency table being the same as the actual mean calculated from the raw data. Can the same list be used when the data are grouped using 4, 5 or 10 class intervals or does it need to be changed in some way? Is it possible to find a list of data values that works in all three cases? Why or why not?

Example 12:

Salma and Ren collect data about the number of days students have attended over a six-week term. They compare the range of two different classes.

Salma uses the raw data each time, while Ren uses a grouped frequency table.

Ren notices his tables are the same for both class A and class B.

Days attended	Frequency
0–5	1
6–10	0
11–15	1
16–20	3
21–25	5
26–30	20
Total	30

He says, 'The range for both classes was 30.'

Salma says, 'No, the range for class A was 30 but the range for class B was 27.'

a) How is this possible?

b) Suggest possible values for the raw data for each class.

c) Compare your answers to part b with another student. What is the same and what is different? Is there more than one possible maximum or minimum for either class?

Example 12 demonstrates how the range calculated from a grouped frequency table is an estimate, as the lower limit of the smallest group and the upper limit of the largest group may not be the minimum and maximum data values. As the raw data are rarely known when the **representation** of a grouped frequency table is used, an assumption must be made about the minimum and maximum data values. Using the lower limit of the smallest group and the upper limit of the largest group provides a sensible estimate.

The **variation** inherent in in this question is such that the frequency table remains the same, even though the information from Salma means that we know that it is representing two different data sets. Rather than having the values provided for them, students need to reason about what must be the same and what must be different about the raw data. This draws attention to how the minimum and maximum values relate to the limits of the groups. In class B, the minimum and maximum values must be different to the lower limit of the smallest group and the upper limit of the largest group. This is contrasted with class A, where the minimum value is equal to the lower limit used for the lowest group, and the maximum value is equal to the upper limit used for the highest group.

Students should create two different data sets that align with the corresponding grouped frequency table and then compare their sets of values across the class. This comparison offers an opportunity for **deepening** their understanding of why the range calculated from the grouped frequency table is an estimate, and also the conditions for which the estimate represents the actual range of the data values.

Example 13:

The heights of 35 apple trees measured in metres to the nearest centimetre are recorded:

1.01, 1.17, 1.27, 1.38, 1.46, 1.51, 1.53, 1.64, 1.68, 1.69, 1.78, 1.83, 1.84, 1.88, 1.91. 1.93. 1.97, 2.00, 2.05, 2.10, 2.11, 2.13, 2.14, 2.18, 2.24, 2.28, 2.30, 2.37, 2.42, 2.47, 2.59, 2.64, 2.75, 2.84, 2.95

a) Calculate the range in the heights of the trees.

The tree heights are grouped in three different frequency tables, shown below. They are used to estimate the range of the height of the apple trees. Each time

Example 13 highlights how the range of a data set represented in a grouped frequency table must be estimated as the difference between the lower limit of the smallest group and the upper limit of the largest group, because the actual data values are unknown. In this example, the raw data are given alongside the grouped frequency tables, with the **variation** lying in the different groupings. Teachers should allow students to see the actual data and make sense of them, before drawing their attention to the information about the data that is lost when it is presented in a grouped frequency table.

Exploring different ways of grouping data emphasises the importance of recognising the effects that the choice of class intervals can have on the accuracy of the estimate of the range. Students could be encouraged to generalise using repeated **language** structures, for example by

the calculation used is 3.0 − 1.0 = 2.0 metres.

Jonny says, "This means that the range from a grouped frequency table is always the same, no matter how the data are grouped."

b) Is Jonny correct? Why or why not?

*c) Group the data in such a way that the range is **greater** than 2.0.*

*d) Group the data in such a way that the range is **smaller** than 2.0.*

e) What did you have to consider in order to answer parts c and d?

unpicking the following sentence and explaining why it is true: *'The bigger the difference between the smallest value in the data set and the lower limit of the smallest group (and the largest value in the data set and the upper limit of the largest group), the bigger the difference between the estimate for the range and the actual range of the data values is likely to be.'*

It is quite common for equal class intervals to be used when grouping data in a grouped frequency table, and so the lower limit of the smallest group and the upper limit of the largest group may be affected by the number of groups used. Asking students to group the data in a different way, and predicting/observing what effects this has on the range, can help with **deepening** their understanding.

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

Height (h)	Frequency
$1.0 \leq h \leq 1.4$	4
$1.4 < h \leq 1.8$	7
$1.8 < h \leq 2.2$	13
$2.2 < h \leq 2.6$	7
$2.6 < h \leq 3.0$	4

Height (h)	Frequency
$1.0 \leq h \leq 1.5$	5
$1.5 < h \leq 2.0$	13
$2.0 < h \leq 2.5$	12
$2.5 < h \leq 3.0$	5

10.1.2.4 Describe the properties of a population using appropriate summary measures

Common difficulties and misconceptions

When using summary measures to describe the properties of a population, students need to recognise the importance of having both a measure of central tendency and a measure of spread to appreciate the distribution of a set of data. Students often think that the three measures of central tendency are equally valid, and struggle to appreciate the subtle differences between them and why they occur. They should recognise the way in which the mean 'levels out' the data (by keeping the total the same but distributing everything evenly), but that it does not give a good measure of central tendency when there are outliers, or the data are skewed; and why the median is a useful measure if the data are evenly spaced, but that it may give a distorted view of the data if not. Presenting summary measures alongside the raw data can help students to appreciate of the limitations of each measure when the raw data are not available.

Ideally, students will have been introduced to the range in such a way that has led them to develop a conceptual understanding of it as a measure of spread, and will have understood how a measure of spread is distinct from a measure of central tendency. However, students can sometimes conflate the two and incorrectly assume that the range is a 'fourth average'. Exploring contexts with extremes can support students to appreciate the difference between central tendency and spread, and how both values can be useful in summarising a data set. For example, you could look at data sets with a high range but low average – such as the ages of people in a Nursery class (including the teacher) – or a low range but high average – such as the ages of people in an elderly care home (not including staff).

Students often have difficulty interpreting statistical measures, especially linking the summary information to the population that it represents. It is key to use contexts that students can engage with, and to provide opportunities for them to discuss in depth the ways in which the properties of a population can significantly affect the measures of central tendency and spread. This supports them in making the transition from commenting on summary measures at a surface level, to developing a deep understanding of possible characteristics of the data that could be affecting the similarities and differences in the statistical measures.

Students need to	Guidance, discussion points and prompts
Appreciate how different measures of central tendency may be more representative depending on the data they are summarising *Example 1:* *Fi has just received some test results.* *Mathematics 56%* *Biology 54%* *Geography 97%* *Economics 56%* *Media Studies 62%* *She calculates the averages of her results.* Mean = 65% Median = 56% Mode = 56% a) *Which average best represents Fi's test results? Explain your answer.* b) *Which average might Fi choose to share with others? Explain your answer.* c) *What might be some advantages and disadvantages of using the mean? How about the median?*	*Example 1* supports with **deepening** students' thinking about measures of average, their differences and the ways in which they represent the data. In answer to part a, Fi's scores are similar in four out of the five subjects, ranging between 54% to 62%, but in Geography she has scored much more highly. This has resulted in the mean value of 65% being the least representative, as it is distorted by the 97% Geography score. The variation between the results is such that all except the result for Geography are below the mean value of 65%, drawing attention to the way in which the mean misrepresents the data. While the median and the mode are both the same value, the mode can be more suited to qualitative data. The median, which has not been distorted by the Geography result, is the most appropriate measure to represent Fi's test results. For part b, Fi may choose to share the mean value for her test results, as this is the highest value. It is not a good **representation** of the data, but it suggests that her test results are higher than they are. When discussing Fi's results, notice students' use of **language** and check whether they refer to the Geography result as being an 'outlier'. Students may have previously discounted outlier values, calculating measures of average for the remaining data values. This provides some insight into how extreme values can distort summary measures, but it is important to emphasise understanding how the three measures of central tendency represent the data when it includes values that differ significantly from the majority of the data values.  It is important to stress to students that taking each measure of central tendency separately, without the context of the other two, can sometimes give a misleading summary of the data. Discuss with colleagues what possible real-world examples of this could be explored with students. How might advertisers take advantage of this, for example?

Understand how data characteristics affect measures of central tendency and spread

Example 2:

Chris owns a cleaning business that employs 15 cleaners. The cleaners have complained that their average pay is below £18 579 per annum, the national average wage for a cleaner.

They quote the following statistics:

Mean: £18 024
Median: £16 682
Mode: £21 674
Range: £12 399

*The following four statements are all **true**:*

- *The difference between the highest and lowest earners is £12 399.*
- *At least two cleaners earn £21 674.*
- *There are seven cleaners who earn more than the median and seven who earn less than the median.*
- *The pay of the seven highest-earning cleaners is further from the median than those of the seven lowest-earning cleaners.*

a) *Explain how we know that each of these statements is true.*

*Chris tells the cleaners that the average pay **is** above the average wage for a cleaner. He includes his own wage in the calculations and quotes the following statistics:*

Mean: £19 671
Median: £16 682
Mode: £21 674
Range: £35 101

b) *What can we infer about Chris's pay from this information?*

*The following four statements are all **true**:*

- *More than one cleaner earns the median pay.*
- *At least three cleaners earn £21 674.*
- *Chris's pay is £22 702 higher than the next highest-earning cleaner.*
- *Chris is paid £44 376.*

Example 2 explores the effects of adding a value to a data set on the mean, median, mode and range. It is an opportunity for students to understand that, while an average can be used as a **representation** of a data set, there are factors (such as outliers) that determine how reliable this representation is. Furthermore, the extent of any effect is determined by how alike the additional data value is to the rest of the other data values. Students may not have any experience of the pay structure for a cleaning business, but it is likely that they will understand that a business owner may be paid more than the people that they employ. They should also glean from this example that it is possible to use statistics in such a way that they support your viewpoint: both the cleaners and Chris have used mathematically-valid calculations, even if Chris's interpretation of the results is then misleading.

In giving students true statements, rather than asking them to create their own, the focus is on understanding the mathematical structures that sit behind each statement. Pay attention to the **language** that students use in their explanations, they may find it easier to evidence their thinking with calculations or diagrams rather than with words. For some statements (such as Chris's actual pay) students' reasoning may take the form of a calculation that finds and then compares the total pay each time, whereas for others (such as explaining why we know there is more than one cleaner who earns the median) a worded explanation may be more appropriate. The reasoning behind the latter statement is key to grasping how the median is determined in a data set and the limitations of the mode as a measure of central tendency, as it does not take all the scores in the data set into consideration.

It is important that students recognise that the size of the pay gap of Chris and the cleaners determines the effect that can be seen on the mean and range. A useful exercise for further **deepening** students' understanding would be to explore the conditions for which the mean and range would remain the same: if Chris' pay is equal to the mean of the employed cleaners, or if it is the same as the highest-paid cleaner. This is fundamental to a deeper understanding of how the structure of a data set affects the summary statistics.

Providing students with summary statistics and asking them about the population they represent helps them to understand how and why the measures of central tendency and spread summarise the data in the way they do, and are affected by the way that the data values are distributed. Discuss different ways to support students who are struggling to engage with the summary statistics. If, for example, the issue is with the size or complexity of the numbers, how could they be adapted to help students to think about the structure of the data, without moving away from the context?

- *The lowest pay in the company is £9 275.*

c) *Explain how we know that each of these statements is true.*

Example 3:

The three tables A to C summarise ages in different populations in a secondary school:

A

Age (years)	
Mean	13.6
Median	14
Mode	15
Range	5

B

Age (years)	
Mean	39
Median	39
Mode	39
Range	39

C

Age (years)	
Mean	26.3
Median	14
Mode	15
Range	50

a) *Describe the population that each table summarises, explaining your answer.*

b) *Create another table that could summarise the ages for your own school year group.*

Example 3 considers possible populations within a secondary school, **deepening** students' thinking about the ways in which the measures of central tendency and spread are affected by the characteristics of these different populations.

Asking students to consider the similarities and differences between the three tables is a helpful starting point and an opportunity to explore why the median and mode are the same in tables A and C. The **variation** between the three tables draws attention to which measures change and which remain the same when a data set is added to or has values removed. When the whole school population, C (students and teachers), is compared with the student population A, the only measure of central tendency that is affected is the mean. The range is also affected, as the oldest age changes while the youngest age remains the same. The table for the teacher population, B, prompts discussion and explores what the mean, median and mode having the same value tells us about how the ages of the teachers are distributed. Students need to recognise that the absence of skew in a data set such as this suggests that the ages are evenly distributed. Students may be surprised that the range of the teachers' ages is also 39 years. Emphasise that the range is a measure of spread, whereas the mean, median and mode are measures of central tendency.

When thinking about the whole school population, stress that the mean is noticeably greater than the median, which means that there are some markedly high values in the data set. A representation can help to visualise this skew. Check that students can explain why this is and are familiar with the **language** used. Students often get confused between positively skewed (extended tail to the right) and negatively skewed (extended tail to the left). It is important that they recognise that the ages in the population summarised in C are positively skewed as there are fewer teachers and so most of the ages are between 11 and 16 years old.

It would be impractical to list data values for ages of the population of a school, but the context provides a setting which is familiar and accessible to students. Discuss what other data sets might exploit students' existing knowledge to support their thinking about how the summary statistics relate to raw data.

Example 4: *Mr and Mrs Weston have two children.* *Mr Weston is tidying up and fills a basket with 20 pairs of shoes that he finds lying around the house. Shoes belonging to all four family members are collected in the basket.* *Mrs Weston notices that:* • *The smallest shoe is a size 1.* • *The mean shoe size is 5.* • *The most common shoe size is 4.5.* • *The median shoe size is 4.5.* • *The range of shoe sizes is 8.* *Suggest possible sizes for the 20 pairs of shoes.*	*Example 4* provides summary statistics for a data set and asks students to 'work backwards' to suggest possible data values that satisfy the measures of central tendency and spread described. This encourages them to rely less on remembering and applying processes, instead **deepening** thinking about how the measures of central tendency and spread represent the data. Students may arrive at a suitable list of 20 shoe sizes through trial and improvement, and may need prompting to engage with the information provided. Ask: • *'What size are the biggest shoes in the basket? How do you know?'* • *'Are any other shoe sizes known?'* • *'If there are four family members, how many different shoes sizes are there likely to be? Could a family member own shoes in more than one size?'* • *'In the context of shoes, what sizes are impossible or unrealistic?'* This task will naturally produce multiple different answers from your class, and selecting which answers to discuss and compare acts as a form of **variation**. Draw attention to students whose strategies reflect the mathematical structures being exploited here. For example, recognising that the total of the 20 shoe sizes must be 100 to give a mean value of 5 demonstrates a good grasp of the structure of the calculation of the mean and allows students to gain a better understanding of how the mean value represents a given set of data. Another key to identifying a possible data set is recognising that both the mode and median are 4.5, which means that the median shoe size is the actual size of a shoe in the basket and not the result of taking the average of the two middle sizes.
Example 5: *At the end of term, students sit exams in five subjects. They are given their scores as percentages.* *Flo is exploring the averages and ranges for her class's results. She creates the table shown below.* *Flo summarises the test results for each subject with the following five statements:* *A Everyone in the class got similar marks.* *B Some students had not been told they were having a test, whereas others had been given time to prepare.*	In *Example 5*, students consider the measures of central tendency and spread from a generalised viewpoint of how they interrelate and interpret this in terms of the distribution of the test results for each subject. The **variation** between the five summary results A to E draws attention to the effects that outliers and skew have on measures of central tendency and spread, and how this differs to summary measures for evenly-distributed data. When discussing the summaries of the results, explore how the measures of central tendency and spread relate to one another and what this tells us about the data set. Pay careful attention to the **language**, as students need to make inferences from statements that do not explicitly reference the mathematical terminology. For example, do students connect the word 'majority' in statement C with

C *The majority of the class found the test too difficult.*

D *A couple of students did much better than the rest of the class.*

E *There was a very wide range of evenly-distributed results in this subject.*

Match each of statements A to E to the subject it is describing.

Explain how you know.

the mode? Similarly, will students recognise the potential for bimodal data in the scenario described by statement B?

Discuss the importance of asking students to match up the statements based on the summary measures, rather than giving them access to the data lists. Spend time looking at the test results in detail, to consider the distributions and explore ways to support and prompt students to engage with the summary statistics when the data values are unknown.

	Science	***Geography***	***Mathematics***	***RE***	***French***
Mean	*50.5*	*50.5*	*34.5*	*26.25*	*56*
Median	*50.5*	*25*	*33*	*16*	*57*
Mode	*No mode*	*20 and 90*	*80*	*16*	*57*
Range	*95*	*87*	*74*	*72*	*13*
IQR	*49*	*69.5*	*28*	*28.5*	*5*

The raw test scores for the five different tests are provided for reference:

Science	3, 7, 10, 14, 17, 21, 24, 28, 31, 35, 38, 42, 45, 49,52, 56, 59, 63, 66, 70, 73, 77, 80, 84, 87, 91, 94, 98
Geography	10, 12, 13, 15, 17, 18, 19, 20, 20, 20, 21, 22, 23, 25, 25, 75, 75, 81, 83, 85, 88, 90, 90, 90, 92, 93, 95, 97
Mathematics	6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 80, 80
RE	3, 5, 7, 8, 8, 9, 11, 11, 12, 13, 14, 15, 16, 16, 16, 17, 18, 25, 28, 32, 36, 43, 49, 56, 57, 64, 71, 75
French	49, 49, 50, 51, 52, 52, 53, 54, 54, 55, 56, 56, 56, 57, 57, 57, 57, 57, 58, 58, 58, 59, 59, 60, 60, 61, 61, 62

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 “Data Page: Average weekly working hours of children”. Our World in Data (2025). Data adapted from International Labour Organization, UNICEF and World Bank. Retrieved from https://ourworldindata.org/grapher/average-working-hours-of-children [online resource]