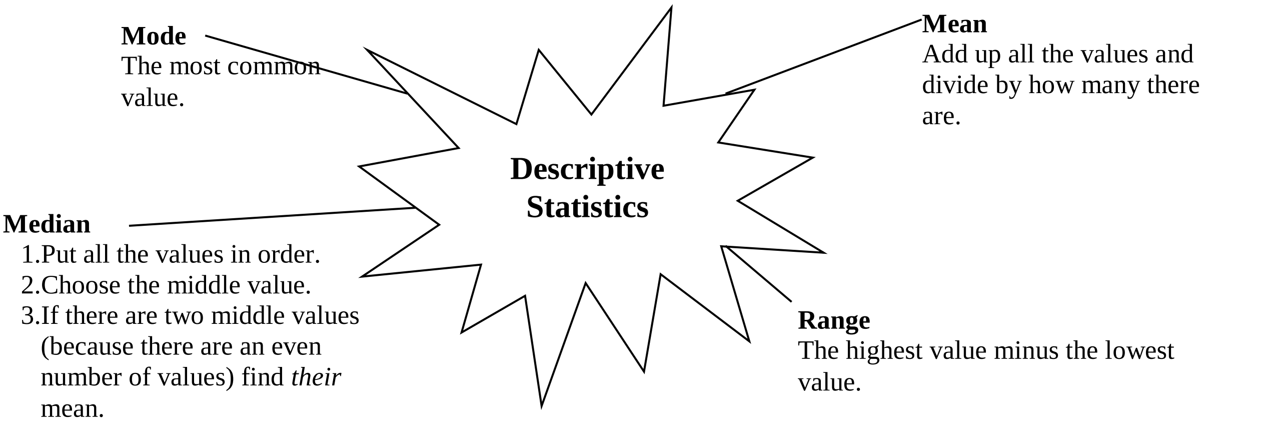
3.3 Statistical Calculations



Descriptive Statistics

Mode
The most common value.

Mean
Add up all the values and divide by how many there are.

Median
1. Put all the values in order.
2. Choose the middle value.
3. If there are two middle values (because there are an even number of values) find *their* mean.

Range
The highest value minus the lowest value.

- Here, "mean" always refers to "arithmetic mean".
  "Geometric mean" and "harmonic mean" are much less commonly used.
- To find the median of a large number of values, one way is first to make a stem and leaf diagram. Then the middle value is easily located.
- Pupils should be aware that spreadsheets can easily do these calculations using functions like **`AVERAGE`**, **`MEDIAN`**, **`QUARTILE`**, etc.

3.3.1 Spider diagram of "Descriptive Statistics" (see above). Which one is the odd one out?

Mode is closely linked to probability. If you chose a value at random, the one you'd be most likely to get would be the mode.

It's OK for the mean and median to be "impossible" values (e.g., 2.4 children), but the mode is always an actual value.

Answer:
Range is not a measure of "centre" of the values; it tells you how spread out they are.

If the data are numerical, then there's always a mean and a median, but there isn't necessarily a mode because two or more values may tie for most common, or all the values may be different from each other.
But you can have a mode in some situations where mean and median are impossible; e.g.,
a modal favourite colour could be "red", but because the data aren't numerical you can't calculate a mean or a median.

3.3.2 Imaginary scores for a game (whatever game you like).
The scores are 3, 1, 5, 0, 1.
What is the "average"?
Is the median 5?

Should we ignore the zero? Will it make any difference to the mean/mode/median/range?

Answers:
mean = 2; median = 1, mode = 1, range = 5
No; you have to put them in order (5 is larger than any of the other values, so it can't be the median).

We can't ignore it if it really happened.
In general it will make a difference. (Imagine that in a maths test one person got 10/10 and everyone else got 0/10. I might like to ignore the 0's, but I can't really!)

3.3.3 Which number is the mean of the other five?

5 4 5 1 6 3

Answer: 4
Whichever one it is, it doesn't change the mean when it's added in, so the answer will just be the mean of all six values: 24 ÷ 6 = 4.

3.3.4 Mean, Median, Mode, Range (see sheet).

This is more interesting than being given a list of numbers and asked to calculate these quantities.

Pupils can invent similar puzzles.

3.3.5 Possible or Impossible? (see sheet)

This task encourages careful logical thinking and lots of experimentation with finding mean, mode, median and range of invented data.

What if? (see same sheet)

Explores the concept of "linear coding"; the result being that if $y = ax + b$ *then* $\bar{y} = a\bar{x} + b$. *This happens because* $\sum y = a\sum x + nb$, *so*

$$\frac{\sum y}{n} = a\frac{\sum x}{n} + b = a\bar{x} + b\ .$$

3.3.6 Advantages and Disadvantages (see sheet).
Why do you think we have three different ways of working out an average?
(Is it just to make life complicated?!)

It's important to think through the advantages and disadvantages of each measure and when they are more or less appropriate.

Imagine a company. These are the annual salaries of the people who work there:
£20 000; £20 000; £20 000; £20 000; £20 000; *£200 000* (the boss).

What is the mean salary?

Answer: £50 000

The boss writes an advert for a job at his company and he says, "The average salary here is £50 000 per year."
What do you think about that?

It's misleading because everyone except him earns less than half of that! The median and the mode would be more honest here.

Think of a situation where the mode or the median would be misleading.

The boss might say to the shareholders, "Our average (median/mode) salary for the whole staff is only £20 000". This conceals his "fat-cat" salary!

It can be fun to invent similar scenarios.

3.3.7 "Birthdays are good for you. Statistics show that people who have the most birthdays live the longest!"
Why is that silly?

Answer: birthdays just measure how many years you've been alive – correlation doesn't mean that one thing causes the other to happen.

3.3.8 **NEED** "Routes to School" sheets (2 copies on the sheet).

Maths tells you the likelihood of being on time if you travel by different routes, but it can't tell you what you should *do – something may be more important to you than arriving as quickly as possible; e.g., travelling with a particular friend.*

Answers:

route	***mean***	***mode***	***median***	***range***
A	*21.95*	*22*	*22*	*16*
B	*21.6*	*24*	*21*	*11*

Pupils could mention other relevant factors, such as preferring to walk or take the bus, perhaps differently depending on the weather.

Pupils could write their answers "to James", offering him mathematical advice.

3.3.9 Discussion: "You can prove anything with statistics." Do you agree or disagree?

It was probably Disraeli (1804-1881) who said that there are "lies, damned lies and statistics".

Perhaps the answer is that statistics can be and are misused. All the more reason for understanding the subject so that the wool won't be pulled over your eyes!

Mean, Median, Mode, Range

In all these questions only use numbers which are *positive integers* (or zero).

1 Find five numbers with a mean of 10.

How many possibilities can you find?

2 Find five numbers with a mean of 10 and with a median that isn’t 10.

3 Find five numbers with a mean of 10 and no mode.

4 Find five numbers with a mean of 10, a median of 7 and a mode of 7.
(Hint: start with 7 in the middle)

5 Find five numbers with a mean of 4, a mode of 5, a median of 5 and a range of 5.

6 Find five numbers with a mean of 6, a mode of 8, a median of 6 and a range of 5.

7 Find five numbers with a mean of 9, a mode of 7, a median of 8 and a range of 7.

Challenge

Can you find four positive integers such that mean < mode < median?
Explain your answer.

What if they don’t have to be positive integers?

What other statistical inequalities can you make definite statements about?

Mean, Median, Mode, Range ANSWERS

In all these questions only use numbers which are positive integers (or zero).

1 Find five numbers with a mean of 10.
How many possibilities can you find?

Any set of 5 numbers that sum to 50 will do, because when you divide by 5 you'll get 10.
e.g.,

8	*9*	*10*	*11*	*12*
10	*10*	*10*	*10*	*10*
6	*8*	*10*	*12*	*14*
0	*0*	*0*	*0*	*50*
0	*5*	*10*	*15*	*20*

2 Find five numbers with a mean of 10 and with a median that isn't 10.

e.g.,

0	*0*	*0*	*0*	*50*
0	*1*	*2*	*22*	*25*

3 Find five numbers with a mean of 10 and no mode.

e.g.,

8	*9*	*10*	*11*	*12*
0	*1*	*2*	*22*	*25*

4 Find five numbers with a mean of 10, a median of 7 and a mode of 7.
(Hint: start with 7 in the middle)

e.g.,

6	*7*	*7*	*12*	*18*
7	*7*	*7*	*7*	*22*

5 Find five numbers with a mean of 4, a mode of 5, a median of 5 and a range of 5.

e.g.,

1	*3*	*5*	*5*	*6*

6 Find five numbers with a mean of 6, a mode of 8, a median of 6 and a range of 5.

e.g.,

3	*5*	*6*	*8*	*8*

7 Find five numbers with a mean of 9, a mode of 7, a median of 8 and a range of 7.

e.g.,

7	*7*	*8*	*9*	*1*

Challenge

Can you find four positive integers such that mean < mode < median? Explain your answer.
What if they don't have to be positive integers?

Let the four numbers in order from smallest to largest be a *,* b *,* c *,* d *(so* $a \le b \le c \le d$ *).*
If there is a mode, then two, three or all four must be equal. (If two are equal, then the other two mustn't be equal to each other – unless all four are equal.)
If all four are equal, then the inequality cannot hold because all three averages would equal the same amount.
If three are equal, then $b = c$ *necessarily, so median = mode, contradicting the inequality.*
So if it's possible at all, it must be when two of the values (but not b *and* c *) are equal.*
If $c = d \ne b$ *, then mode > median.*
So the only possibility is $a = b$ *, (and mode < median), but then the mean =* $\frac{1}{4}(a + a + c + d)$ *which must be >* a *, so mean > mode, contradicting the inequality. So, impossible, even without the restriction that the numbers have to be positive integers.*

Possible or Impossible?

I have a set of numerical values which could be anything, and I want to know whether each of these statements (separately) is possible or impossible.

Can it apply to the *mean*, the *median*, the *mode* or the *range*?

If it's possible, give an example; if impossible, try to say why.

statement	***mean***	***median***	***mode***	***range***
There isn't one.				
It's equal to zero.				
It's the highest value.				
It's the lowest value.				
It's greater than any of the values.				
It's less than any of the values.				

What if …?

You have a set of values and you've worked out the mean.

What will happen to the mean if …

1 you add 10 to all the values?
2 you subtract 10 from all the values?
3 you multiply all the values by 10?
4 you divide all the values by 10?
5 you square all the values?

Make up some numbers and try it.
Try to explain your answers.

Possible or Impossible? ANSWERS

I have a set of numerical values which could be anything, and I want to know whether each of these statements (separately) is possible or impossible.
Can it apply to the *mean*, the *median*, the *mode* or the *range*?
If it’s possible, give an example; if impossible, try to say why.

statement	*mean*	*median*	*mode*	*range*
There isn’t one.	Impossible, provided all the values are numbers.	Impossible, provided all the values are numbers.	Possible; e.g., 1, 2, 3, 4, 5.	Impossible, provided all the values are numbers.
It’s equal to zero.	Possible if all the values are zero or some are negative so that their total comes to zero.	Possible if all the values are zero or some are negative so that the median happens to be zero.	Possible; e.g., 0, 0, 1, 2, 3.	Possible if all the values are equal; e.g., 5, 5, 5, 5, 5.
It’s the highest value.	Possible only if all the values are the same.	Possible only if all the values are the same.	Possible; e.g., 1, 2, 3, 4, 4.	Possible if the lowest value is zero.
It’s the lowest value.	Possible only if all the values are the same.	Possible only if all the values are the same.	Possible; e.g., 1, 1, 2, 3, 4.	Possible if the highest value is twice the lowest value; e.g., 3, 4, 5, 6.
It’s greater than any of the values.	Impossible, because the mean represents equal shares of the total amount.	Impossible, because the “middle” value can’t be more than any of them.	Impossible, because it must be an actual value.	Possible if the lowest value is negative; e.g., –2, 3, 4, 5.
It’s less than any of the values.	Impossible, because the mean represents equal shares of the total amount.	Impossible, because the “middle” value can’t be less than any of them.	Impossible, because it must be an actual value.	Possible if the values are not very spread out compared with their size; e.g., 7, 8, 9, 10.

What if …?

You have a set of values and you’ve worked out the mean.
What will happen to the mean if …

1 you add 10 to all the values? *mean goes up by 10*
2 you subtract 10 from all the values? *mean goes down by 10*
3 you multiply all the values by 10? *mean is multiplied by 10*
4 you divide all the values by 10? *mean is divided by 10*
5 you square all the values? *no simple answer;*

$$\frac{\sum x^2}{n} \neq \left(\frac{\sum x}{n}\right)^2$$

, so the answer is not *the square of the mean.*

Make up some numbers and try it. Try to explain your answers.

Advantages and Disadvantages

What are some of the advantages and disadvantages of using the mean, the mode and the median as averages representing the whole set of data?

	advantages	***disadvantages***
mean	• takes all the data into account; • well-known and understood ("equal shares");	• swayed by extreme values; • works only with numbers (e.g., not with colours); • not always a possible value (e.g., "2.4 children"); • sometimes it's too much work to do the calculation;
median	• not affected by extreme values; • can be approximated from cumulative frequency curves; • often a possible value (e.g., it's often an integer if the data have to be integers);	• ignores the actual values of most of the data; • works only with numbers that can be ordered (e.g., not with colours); • sometimes it's inconvenient to have to put all the data in order first;
mode	• not affected by one extreme value; • works with qualitative data (e.g., colours) as well as with numbers; • always a possible value (e.g., an integer if the data have to be integers); • useful in probability work;	• ignores the actual values of much of the data; • there isn't always a mode;

Make up an example to illustrate each of the points above.

What other advantages and disadvantages can you think of?

Routes to School

James has two possible ways of getting to school:

A He can catch a bus into the town centre and another bus out; or
B He can walk in a different direction and then just catch one (different) bus.

He isn’t sure which way is quicker so he decides to time his journey from leaving his front door to arriving at the school gate. He times 20 journeys by route **A** and 20 by route **B**.
Here are his results (in minutes).

Route A

24	18	14	19	22	23	18	15	22	30
30	26	22	27	16	22	26	19	19	27

Route B

25	18	21	18	28	21	17	19	26	24
25	24	21	17	20	22	24	20	24	18

Work out the mean, median, mode and range for the times for route **A** and for route **B**.
Which route do you think James should use, and why?

Routes to School

James has two possible ways of getting to school:

A He can catch a bus into the town centre and another bus out; or
B He can walk in a different direction and then just catch one (different) bus.

He isn’t sure which way is quicker so he decides to time his journey from leaving his front door to arriving at the school gate. He times 20 journeys by route **A** and 20 by route **B**.
Here are his results (in minutes).

Route A

24	18	14	19	22	23	18	15	22	30
30	26	22	27	16	22	26	19	19	27

Route B

25	18	21	18	28	21	17	19	26	24
25	24	21	17	20	22	24	20	24	18

Work out the mean, median, mode and range for the times for route **A** and for route **B**.
Which route do you think James should use, and why?

Statistical Calculations

- **Discrete** data – only certain values are possible (e.g., shoe sizes: 1, 1.5, 2, etc.).
- **Continuous** data – any value is possible within a certain range (e.g., heights of pupils measured in cm).
- **Stem and leaf diagrams** – remember to include a **key** and to use LOW and HIGH for extreme values that are hard to fit on the scale. If there's a lot of data on each line you can do a **stretched** version with, say, 0-4 on one line and 5-9 on the next.
  Good for finding the median and seeing the general shape.

- **Mean** (arithmetic mean): add up all the values and divide by how many there are.

$$\bar{x} = \frac{\sum x}{n}$$

  This shares out the total x evenly, but it can be distorted by exceptionally large or small values, and the answer may not be a possible actual value (e.g., "2.4 children").
  For data where there are f occurrences of each x, we use the formula

$$\bar{x} = \frac{\sum fx}{\sum f}$$

  For grouped data, x has to be the mid-point of the class.
- **Mode** (or **modal class** for grouped data): most frequent value.
  Often easy to find, but there isn't always one.
  There may be just one mode (**unimodal**) or there may be two values – not necessarily equally frequent – that are more common than the rest (**bimodal**).
  Useful if you want the most probable value or if you're using qualitative data (e.g., modal favourite colour – you can't work out a "mean colour").
  Data may be **skewed positively** (mode to the left, long tail to the right) or **negatively** (the opposite).
- **Median**: the middle value when they're listed in order (or the mean of the two middle values where there are an even number of data).
  Often easy to find, especially from a stem and leaf diagram.
  Not affected by extreme or missing values.
- **Mid-range**: the mean of the highest and lowest values.
  Good for fairly symmetrical data but dependent on the most extreme values.

- **Range**: the highest value minus the lowest.
  Easy to calculate but dependent on the most extreme values.
- **Mean absolute deviation from the mean**: $\dfrac{\sum f\left|x-\bar{x}\right|}{\sum f}$.
- **Variance**: defined as the mean of the square deviations from the mean: $\dfrac{\sum f\left(x-\bar{x}\right)^2}{\sum f}$, but for calculations it's easier to use $\dfrac{\sum x^2}{n} - \bar{x}^2$, or for data where there are f occurrences of each x we use the formula $\dfrac{\sum fx^2}{\sum f} - \bar{x}^2$.
- **Standard deviation**: the positive square root of the variance. Data which are more than two standard deviations above or below the mean are exceptional and may be classed as **outliers**. There may be a special reason for these values.

- **Linear coding**: Imagine you have a set of data (x) with mean $\bar{x}$ and standard deviation $\mathrm{sd}(x)$. If you modify each value so that you create a new variable y so that $y = ax + b$, then that will affect the mean and standard deviation so that the new mean $\bar{y}$ and new standard deviation $\mathrm{sd}(y)$ will be $\bar{y} = a\bar{x} + b$ and $\mathrm{sd}(y) = a \times \mathrm{sd}(x)$.