Edexcel International AS Maths: Statistics 1

Your notes

Statistical Measures

Contents

Your notes

Mean, Mode, Median

What are mean, median and mode?

- Mean, median and mode are measures of **location**
  - A measure of location gives information about where data is in the number system
  - Mean, median and mode are **measures of central tendency**
  - They describe where the centre of the data is
- They are all types of **averages**
  - In Statistics it is important to be specific about which average you are referring to

How are mean, median and mode calculated?

- You should already be familiar with finding the mean, median and mode from raw, ungrouped data
- The **mode** is the value that occurs **most often** in a data set
  - In a frequency table the **group** or **class** that occurs most often will be referred to as the **modal class**
  - A data set with more than one mode is **bimodal**
- The **median** is the **middle** value when the data is in **order of size**
  - If there are two values in the middle of the data set, the median is the **midpoint** of the two values
  - If finding median from a **frequency table** find the cumulative frequency first and find the group or class where the middle value will lie
  - You may have to use **linear interpolation** when finding the median from a **grouped frequency table**
- The **mean** is the **sum** of all the values **divided by the number of values** in the data set

What are summary statistics and their notation?

- Summary statistics are information that summarises a set of data values
- For n items in a data set:
  - The sum of the data is represented by

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$$

- This is usually written $\sum x$ and reads as 'sigma x'

- The mean of the data is represented by

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\Sigma x}{n}$$

- This reads as 'x bar'
- You will come across more summary statistics later in the course

How do we choose the best measure of central tendency?

- It is often better to use one of the averages over the others, depending on the data set
  - It's a good idea to be aware of the advantages and disadvantages of the use of each average
- The mean uses all of the data values, this is good for a large data set where all of the values are close together, but also means that the mean can be affected by extreme values
- The median is not affected by very high or low values so is a good average to use in data sets with extreme values
- The mode is very useful in a lot of practical situations, however often there may be more than one mode, no mode or even a mode that is nowhere near the middle of the data set

Worked Example

For the data set given below, find the mode, median and mean.

23 19 14 28 27 19

Answer:

Your notes

Your notes

23 19 14 28 27 19

Mode = most common number = 19

Mode = 19

To find the median the data must be in order first

14 19 19 23 27 28

Median is half way between the two middle numbers.

$$\text{Median} = \frac{19 + 23}{2} = 21$$

Median = 21

The mean is the sum of the numbers divided by how many there are.

$$\text{Mean} = \frac{\Sigma x}{n} = \frac{14 + 19 + 19 + 23 + 27 + 28}{6} = \frac{130}{6}$$

Mean = 21.7 (3.s.f.)

Quartiles & Range

What are quartiles and percentiles?

- **Quartiles and percentiles** are measures of location
- Quartiles divide a population or data set into **four equal sections**
  - The **lower quartile, Q_1,** splits the lowest 25% from the highest 75%
  - The **median, Q_2,** is the value that is 50% of the way through the data
  - The **upper quartile, Q_3,** splits the lowest 75% from the highest 25%
- Percentiles divide the data into 100 parts
  - The 70th percentile lies seven-tenths of the way through the data
    - 70% of the data is below it and 30% is above it

How are quartiles calculated?

- For a data set of size, n,
  - To find the lower quartile, calculate $\frac{n}{4}$
    - If $\frac{n}{4}$ is an integer then the lower quartile is the midpoint of the corresponding value and the one above it
    - If $\frac{n}{4}$ is not an integer then the lower quartile is the value corresponding to the next integer up

- To find the upper quartile, calculate $\frac{3n}{4}$
    - If $\frac{3n}{4}$ is an integer then the upper quartile is the midpoint of the corresponding value and the one above it
    - If $\frac{3n}{4}$ is not an integer then the upper quartile is the value corresponding to the next integer up
- You can also use your calculator to find the quartiles, make sure you know how to put your calculator into STAT mode, enter the data and find the values of Q_1, Q_2 and Q_3
  - You could be expected to be able to do this in the exam

What are the range and interquartile range?

- The **range** and **interquartile range** are both **measures of spread**
- A measure of spread gives information about how spread out the data set is
- The **range** is the difference between the **largest** and **smallest** values in the data set
  - All data points in the set will be included in the range, including **extreme values**
- The **interquartile range** is the difference between the **upper quartile** and the **lower quartile**
  - Only the middle 50% of the data is included in the interquartile range
  - It is not affected by extreme values
  - Sometimes an interpercentile range could be asked for, this is the difference between two given percentiles
    - For example, the 20th to 80th interpercentile range would be the difference between the 80th percentile and the 20th percentile
- The units for range and interquartile range are the same as the units for the original data

What is an interpercentile range?

- An **interpercentile range** is the difference between two percentiles for the data set
  - If the data has **units** (seconds, cm, etc.), then the interpercentile range has the same units as the values in the data set
- It is necessary to **specify** which percentiles are to be used
  - e.g. the 30^{th} to 80^{th} interpercentile range
    - this would be 80^{th} percentile minus 30^{th} percentile
  - or the 2.5^{th} to 97.5^{th} interpercentile range
    - this would be 97.5^{th} percentile minus 2.5^{th} percentile

Your notes

- Note that the 25th to 75th interpercentile range is the same as the **interquartile range**

- The interpercentile range is a **measure of dispersion** (i.e. a measure of **spread**)
  - Like the interquartile range it is not affected by the most **extreme values** in the data set
  - You can choose the percentiles to focus on the part of the data set you are most interested in

Your notes

Worked Example

Find the range and interquartile range for the data set given below

43 29 70 31 84 56 17

Answer:

Always start by putting the values in order

17 29 31 43 56 70 84

Range is the difference between the biggest and smallest.

Range = 84 − 17 = 67

Range = 67

Interquartile range = upper quartile − lower quartile

$n = 7$

lower quartile = $\frac{7}{4}^{th}$ value = 1.75^{th} value

next integer up = 2

$Q_1 = 2^{nd}$ value = 29

Upper quartile = $3\frac{7}{4}^{th}$ value = 5.25^{th} value

next integer up = 6

$Q_3 = 6^{th}$ value = 70

This is easier to see by looking at the data directly.

17	29	31	43	56	70	84
	↑		↑		↑	
	Q_1		Q_2		Q_3	

Interquartile range = $Q_3 - Q_1 = 70 - 29 = 41$

IQR = 41

Examiner Tips and Tricks

Be aware of the difference between averages and ranges, especially when answering contextual questions asking you to describe or compare data. Remember, averages give an indication of where the data are whilst range gives an indication of how varied the data are.

Your notes

Ungrouped data

- A frequency table for **ungrouped data** shows the frequency of **individual data values**

Why use a frequency table for ungrouped data?

- When collecting large amounts of raw data, it is quicker to use a tally system and then collate the results into a frequency table
- Ungrouped frequency tables are normally used for **numerical, discrete data**
- Organising data into a frequency table makes it easier to work with
  - Calculating averages, ranges and summary statistics can be done much quicker from a frequency table than from raw data
- It gives a clear **pattern** of the data
  - It is easy to quickly see where most of the data are and to see extreme values
- A frequency table for ungrouped data keeps all of the original data values
  - It is still possible to calculate the actual averages, ranges and summary statistics

How are the mean, median and mode calculated from an ungrouped frequency table?

- You should already be familiar with finding the mean, median and mode from raw, ungrouped data
- The **mode** is the value that occurs **most often** in a data set
  - In an ungrouped frequency table the **data value** with the **highest frequency** will be the mode
- The **median** is the **middle** value when the data is in **order of size**
  - To find the median from an ungrouped frequency table, add the frequencies together until you reach the value that is half of the total
  - For a data set of values, calculate $\frac{n+1}{2}$ and the median will be whichever data value corresponds with this frequency
- The **mean** is the **sum** of all the values **divided by the number of values** in the data set
  - To find the mean, multiply each data value by its corresponding frequency, find the sum of these values and then divide by the total frequency
  - The notation for the sum of the values is Σxf

- The formula for the mean from a frequency table is $\frac{\Sigma xf}{\Sigma f}$

Your notes

Worked Example

The frequency table below gives information about the shoe size of a group of year 12 students.

Shoe Size	37	37.5	38	38.5	39	39.5	40	40.5	41
Frequency	3	3	12	5	16	9	7	4	1

(i) Write down the modal shoe size.

(ii) Explain how you know that the median shoe size is 39.

(iii) Calculate the mean shoe size.

Answer:

(i) Modal shoe size is the most common shoe size.

the highest frequency = 16

Modal shoe size = 39

(ii) To find the median we first need to know the total number of students:

$f = 3 + 3 + 12 + 5 + 16 + 9 + 7 + 4 + 1 = 60$

The data is discrete so we use $\frac{n+1}{2}$ to find which value the media will be.

$\frac{60+1}{2} = 30.5^{th}$ value

Use the cumulative frequencies to find which data value is half way between the 30^{th} and the 31^{st}.

Shoe size	37	37.5	38	38.5	39	39.5	40	40.5	41
Frequency	3	3	12	5	16	9	7	4	1
Cumulative frequency	3	6	18	23	39				

+3 +12 +5 +16

We are looking for the 30.5^{th} value so we don't need to go further

The median shoe size is half way between the 30^{th} and 31^{st} shoe size which are both 39, therefore median = 39.

Your notes

(iii)

Shoe size, x	37	37.5	38	38.5	39	39.5	40	40.5	41
Frequency, f	3	3	12	5	16	9	7	4	1
xf	111	112.5	456	192.5	624	355.5	280	162	41

$$\text{mean} = \bar{x} = \frac{\Sigma xf}{\Sigma f} = \frac{111 + 112.5 + 456 + \ldots + 162 + 41}{60} = \frac{2334.5}{60}$$

Mean = 38.9 (3.s.f.)

Examiner Tips and Tricks

- Use common sense when checking your answers, is your mean within the range of the data? Does is seem right? A mean of 140 for example could not be correct if the data is about ages of students taking an exam at university.

Grouped data

- A frequency table for **grouped data** is usually used for large amounts of **continuous data**
- They show the frequency of **data values** that are within a particular **group** or **class**

What are the advantages and disadvantages of grouping data?

- Grouping data is especially useful when data can take a large range of different values
- Trends and patterns can be easily spotted when data has been grouped
- Calculations are much quicker with data that has been grouped
- It is important to be aware, however, that grouped frequency tables also have some negatives
  - The actual data values are lost when data is grouped
  - It is only possible to calculate estimated averages, ranges and summary statistics

What notation is used for grouped frequency tables?

- When grouping data, it is important to be clear about which **group** or **class** any data value should be entered into
  - A group entry of 10 – 20 followed by 20 – 30 would be unclear because the number 20 could be entered into both groups
    - If the data are discrete, this could be written as 10 – 19 and 20 – 29
    - For continuous data, this could be changed to 10 – and 20 –
    - This would most likely be represented as $10 \le x < 20$ and $20 \le x < 30$

- Most commonly inequalities are used to group continuous data as they leave no ambiguity
- If the data are continuous, always check that there are no gaps between upper boundary of a class and the lower boundary of the next class
  - If there are gaps you will need to close these gaps by changing the boundaries before carrying out any calculations
    - For example, the groups $10 \leq x \leq 19$ followed by $20 \leq x \leq 29$ will become $9.5 \leq x < 19.5$ and $19.5 \leq x < 29.5$
    - Check the inequality signs carefully
- Be careful when deciding what category the data falls into, taking the group $10 \leq x \leq 19$ for example,
  - if the data had been rounded it would take the form $9.5 \leq x < 19.5$ as described above
  - if the data had been **truncated** however, then the boundaries would become $10 \leq x < 20$
    - this is most likely to happen with age, which is technically a continuous variable due to being able to take any value, however we would usually consider age by counting years

How do I find averages from grouped frequency tables?

- Instead of finding the **mode**, when working with a grouped frequency table we instead find the **modal class**
  - This will be the **class** (group) with the **greatest frequency**
- It is only possible to calculate an **estimate** for the **mean** and the **median** from a grouped frequency table
- Calculating the **estimated mean** is the same as for ungrouped frequency tables, however you will need to find the **midpoints** first
  - the midpoint is the mean of the upper and lower values in the class boundaries
  - multiply the midpoint by its corresponding frequency, find the sum of these values and then divide by n
- Calculating the **estimated median** is more complicated and can involve using **linear interpolation**

How do I use linear interpolation on grouped data?

- For grouped data, the **median, quartiles and percentiles** can be found using a process called **linear interpolation**
- **Interpolation** uses the assumption that the data values are evenly spread throughout each class

Your notes

- Once you know which data value you are looking for, interpolation uses proportion to find how far through each class the data value should be
- It is important to remember to add on the lower-class boundary after finding the correct data value

- Follow these steps to use linear interpolation to find the median, quartiles or percentiles:

Step 1. Find the place within the data where the value you are looking to find is positioned, e.g. the median will be the $\frac{n}{2}$ th value

Step 2. Add up the frequencies until you know which class this value will fall into (this is finding the cumulative frequency)

Step 3. Write the upper- and lower-class boundaries of this group on the upper side of a simple number line

Step 4. On the other side of the same number line add the lower and upper values of the **cumulative frequency** for this group to the ends of the line

Step 5. Use proportion to calculate the data value that is the same fraction of the way through the group on the upper side of the number line as where the value you are looking for would sit on the lower side of the number line

- To see these steps in action, go to part (iv) of the worked example
- This method can be used to find any percentile, for example the 81st percentile will be the value that is $\frac{81}{100}$ ths of the way through the data

Your notes

Worked Example

The table below shows the heights in cm of a group of year 12 students.

Height, h	**Frequency**
$150 \le h < 155$	3
$155 \le h < 160$	5
$160 \le h < 165$	9
$165 \le h < 170$	7
$170 \le h < 175$	1

(i) Write down the class that a student of exactly 160 cm should be added to.

(ii) Write down the modal class.

(iii) Calculate the estimated mean height.

(iv) Use linear interpolation to find the interquartile range of the heights

Answer:

(i) The two classes with 160 in are:

$155 \leqslant h < 160$ Not equal to

$160 \leqslant h < 165$

Or equal to

160 must go into the $160 \leqslant h < 165$ class

(ii) The modal class is the group with the greatest frequency.

Modal class = $160 \leqslant h < 165$

(iii) Add the midpoint to the table and then multiply it by the frequency.

Height, h, cm	Midpoint, x	Frequency, f	xf
$150 \leqslant h < 155$	152.5	3	152.5 x 3 = 457.5
$155 \leqslant h < 160$	157.5	5	157.5 x 5 = 787.5
$160 \leqslant h < 165$	162.5	9	162.5 x 9 = 1462.5
$165 \leqslant h < 170$	167.5	7	167.5 x 7 = 1172.5
$170 \leqslant h < 175$	172.5	1	172.5 x 1 = 172.5

$\Sigma fx = 457.5 + 787.5 + 1462.5 + 1172.5 + 172.5 = 4052.5$

$\Sigma f = 3 + 5 + 9 + 7 + 1 = 25$

Estimated mean $= \dfrac{\Sigma fx}{\Sigma f} = \dfrac{4052.5}{25} = 162.1$

Estimated mean height = 162 cm (3.s.f.)

(iv) Step 1: Find the place within the data where Q_1 and Q_3 are positioned.

Lower Quartile $= Q_1 = \frac{25}{4}^{th}$ value $= 6.25^{th}$ value

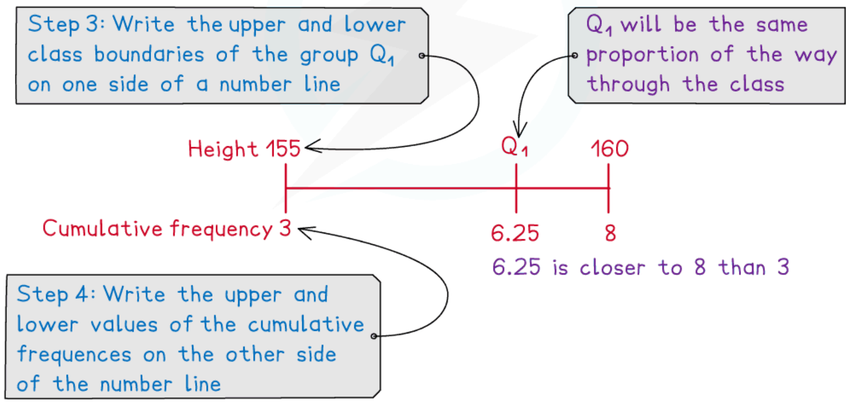
Upper Quartile $= Q_3 = \frac{3 \times 25}{4}^{th}$ value $= 18.75^{th}$ value

Height, h, cm	Frequency, f	Cumulative frequency, cf
$150 \leqslant h < 155$	3	3
$155 \leqslant h < 160$	5	8
$160 \leqslant h < 165$	9	17
$165 \leqslant h < 170$	7	24
$170 \leqslant h < 175$	1	25

Step 2: Find the cumulative frequency

6.25^{th} value is in this group

18.75^{th} value is in this group



Your notes

Step 5: Use proportion to calculate the value of Q_1

$$\frac{6.25-3}{8-3} = \frac{Q_1 - 155}{160-155}$$

$$\frac{3.25}{\cancel{5}} = \frac{Q_1 - 155}{\cancel{5}}$$

$Q_1 = 155 + 3.25 = 158.25$

18.75 is closer to 17 than 24

$$\frac{18.75-17}{24-17} = \frac{Q_3 - 165}{170-165}$$

$$\frac{1.75}{7} = \frac{Q_3 - 165}{5}$$

$$\frac{5}{4} = Q_3 - 165$$

$$Q_3 = 165 + \frac{5}{4} = 166.25$$

Interquartile range $= Q_3 - Q_1 = 166.25 - 158.25$
$= 8$

IQR = 8 cm

Examiner Tips and Tricks

- There can be a lot of calculations when working with grouped frequency tables, be extra careful when using your calculators as it is easy to make small errors with these questions. Use the table and add information to the table as you go as there will be marks available for showing work within the methods. Make sure you know how to use your calculator to find all of the summary statistics as this can save time and reduce errors.

Your notes

Standard Deviation & Variance

The **variance** is another measure for the spread of the data, it measures the variability from the mean of the data.

What is the variance and the standard deviation?

- The variance is a statistic that tells us how **varied** a set of data is
  - Data that is more spread out will have a greater variance
  - Data that is consistent and close together will have a smaller variance
- The **standard deviation** is the square root of the variance
- The symbol for the population standard deviation is the lowercase Greek letter sigma, **σ** and for variance is sigma squared, $\boldsymbol{\sigma}^2$
- Standard deviation and variance are used interchangeably within this course so make sure you look out for which one a question shows or asks for

How are the variance and standard deviation calculated?

- There is more than one formula that can be used for calculating the variance, and you should choose the most useful one
- For a set of n values $x_1, x_2, \ldots, x_i, \ldots, x_n$ the variance is the sum of the squares of the deviations from the mean, divided by the frequency
  - Variance = $\dfrac{\Sigma(x-\bar{x})^2}{n}$
    - This formula can be time consuming and therefore is rarely used in this statistics course
- A second, **easier to use version of the variance** is:
  - Variance = $\dfrac{\Sigma x^2}{n} - \bar{x}^2$
  - This version is easier to work with and should be used in most instances
    - The **summary statistic** $S_{xx} = \Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - \dfrac{(\Sigma x_i)^2}{n}$ can help derive different formulae as shown below:

Variance =

$$\sigma^2 = \frac{\Sigma(x-\bar{x})^2}{n} = \frac{S_{xx}}{n} = \frac{1}{n}\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right) = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2 = \frac{\Sigma x^2}{n} - \bar{x}^2$$

- An easy way to remember this is to think of it as '*the sum of x squared over n minus the sum of the mean squared*'
- Most calculators can be used to find summary statistic such as the standard deviation and variance fairly quickly, practice finding it on yours
- The **standard deviation** is the square root of the variance
  - Standard deviation = $\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}} = \sqrt{\frac{\Sigma x^2}{n} - \bar{x}^2} = \sqrt{\frac{S_{xx}}{n}}$
    - You are given the formula for S_{xx} but not for variance or standard deviation
- The units for standard deviation are the same as the units for the data and the units for variance are the same as the units for the data but squared

How are the variance and standard deviation calculated from a frequency table?

- The method for finding the variance from a frequency table is similar to that of the mean
  - If calculating from a grouped frequency table, find the midpoints, x, first
  - Multiply each x value by its corresponding frequency and use these values within the formulae
  - The formulae will become
    - Variance = $\sigma^2 = \frac{\Sigma f(x-\bar{x})^2}{\Sigma f} = \frac{\Sigma f x^2}{\Sigma f} - \left(\frac{\Sigma f x}{\Sigma f}\right)^2$
    - Standard deviation = $\sigma = \sqrt{\frac{\Sigma f(x-\bar{x})^2}{\Sigma f}} = \sqrt{\frac{\Sigma f x^2}{\Sigma f} - \left(\frac{\Sigma f x}{\Sigma f}\right)^2}$

Worked Example

Phoom recorded the length of time, t, it took him, in minutes, to answer a selection of further calculus exam questions. The data is summarised in the table below.

Time (minutes)	Frequency
$2 \le t < 4$	1
$4 \le t < 6$	4
$6 \le t < 8$	7
$8 \le t < 10$	5
$10 \le t < 12$	2

$12 \leq t < 20$	2

(i) Calculate an estimate of the variance of the time taken to complete a question, include units with your answer.

(ii) Write down an estimate of the standard deviation of the time taken to complete a question, include units with your answer.

Answer:

(i)

Find the midpoint, h, of each group:

Time, (mins)	Midpoint, t	t^2	Frequency, f	ft	ft^2
$2 \leq t < 4$	3	9	1	3	9
$4 \leq t < 6$	5	25	4	20	100
$6 \leq t < 8$	7	49	7	49	343
$8 \leq t < 10$	9	81	5	45	405
$10 \leq t < 12$	11	121	2	22	242
$12 \leq t < 20$	16	256	2	32	512

This group has a different width to the others.

Find the summary statistics:

Your calculator can be used in 'STAT' mode to find these values

$\Sigma ft = 171 \quad \Sigma ft^2 = 1611 \quad \Sigma f = 21$

(i) Substitute into the formula for variance

$$\sigma^2 = \frac{\Sigma ft^2}{\Sigma f} - \left(\frac{\Sigma ft}{\Sigma f}\right)^2 = \frac{1611}{21} - \left(\frac{171}{21}\right)^2$$

$$= \frac{510}{49} = 10.408$$

Estimated variance = 10.4 minutes^2 (3.s.f.)

(ii) The standard deviation is the square root of the variance

$$\sigma = \sqrt{\frac{510}{49}} = \frac{\sqrt{510}}{7} = 3.2261...$$

Estimated standard deviation = 3.23 minutes (3.s.f.)

Your notes

Examiner Tips and Tricks

- Look out for whether a question gives or asks for the standard deviation or variance, especially if the question is using sigma notation.
- Choose which formula to use wisely, most of the time the summary statistics will be given so only one of the formulae will be possible. On the rare occasion that you are asked to calculate directly from a table think carefully about which version of the formula is quickest and easiest to use. It will almost always be the second version given in this revision note.

Your notes

Coding data

Sometimes data needs to be coded for further use with calculations. This is particularly useful with data that deals with very small or very large numbers, or with data that needs to be classified for research purposes.

What is coding?

- Coding is a way of simplifying data to make it easier to work with
- The coding must be carried out on all values within the data set and will normally be done using a given formula
- Coding can be carried out in a number of ways:
  - Adding or subtracting a constant to each data value
  - Multiplying or dividing each data value by a constant
  - A combination of both of the above

How are statistical calculations carried out with coded data?

- If you know the **mean** or **standard deviation** of the original data it is possible to find the mean or standard deviation of the coded data and vice versa
- It is important to remember what the **mean** and **standard deviation** actually tell us about the data to understand how coding calculations work
  - The mean is a measure of **location**, changing the data set in any way will cause the mean to change in the same way
  - The **standard deviation** is a measure of spread, adding or subtracting a constant to every value within the data set will **not change** the standard deviation of the data set
    - Multiplying or dividing every value within the data set by a constant will change the standard deviation by the **modulus** of the constant
    - If the data were coded by multiplying or dividing by a **negative**, the standard deviation will change by the equivalent positive value
- Anytime calculations are carried out on data that has been coded,
  - The **original mean** can be found by solving the equation to reverse the coding
  - For example, if the data, x, was coded using the formula

$$y = ax + b$$

Then the mean of the coded data, $\overline{y}$ would be

$$\bar{y} = a\bar{x} + b$$

The original mean, $\bar{x}$, will be $\bar{x} = \dfrac{\bar{y} - b}{a}$

- The **original standard deviation**
  - Will be the same as the coded standard deviation if the data was coded by adding or subtracting a constant only
  - Can be found by reversing the coding if the data was coded by multiplying or dividing by a constant only
  - If the data was coded by a combination of both then only the multiplying or dividing will need to be reversed to find the original standard deviation
  - For example, if the data, was coded using the formula

$$y = ax + b$$

Then the standard deviation of the coded data, σ_y would be

$$\sigma_y = |a|\sigma_x$$

The original standard deviation, σ_x, will be

$$\sigma_x = \frac{\sigma_y}{|a|}$$

Worked Example

The shoulder height, h, of a group of Asian elephants living in a nature reserve are summarised in the table below.

Height, cm	Frequency, f
$200 \leq h < 220$	2
$220 \leq h < 240$	5
$240 \leq h < 260$	8
$260 \leq h < 280$	8
$280 \leq h < 300$	5
$300 \leq h < 320$	2

Your notes

(i) Code the data using the formula $x = \dfrac{h-250}{20}$

(ii) Use the coded data to find an estimate for the mean and standard deviation, you may use the summary statistics $\Sigma xf = 15, \Sigma x^2 f = 59, \Sigma f = 30.$

Answer:

Find the midpoint, h, of each group:

Use the formula to code the data from the midpoints

(i)

Height, cm	Midpoint, h	$x = \dfrac{h-250}{20}$
$200 \leqslant h < 220$	210	-2
$220 \leqslant h < 240$	230	-1
$240 \leqslant h < 260$	250	0
$260 \leqslant h < 280$	270	1
$280 \leqslant h < 300$	290	2
$300 \leqslant h < 320$	310	3

Your notes

Your notes

(ii) Given the summary statistics of the coded data, x.

$\Sigma f = 30 \qquad \Sigma xf = 15 \qquad \Sigma x^2 f = 59$

Use the x values with the formula for the mean to find the coded mean, $\bar{x}$

$$\text{Coded mean} = \bar{x} = \frac{\Sigma xf}{\Sigma f} = \frac{15}{30} = 0.5$$

Reverse the coding to find the estimated mean, $\bar{h}$:

$$\bar{x} = \frac{\bar{h} - 250}{20}$$

$$20\bar{x} = \bar{h} - 250$$

$$\bar{h} = 20\bar{x} + 250$$

$$= 20\,(0.5) + 250$$

$$= 260$$

Use the x values with the formula for the standard deviation

$$\sigma x = \sqrt{\frac{\Sigma x^2 f}{\Sigma f} - \left(\frac{\Sigma xf}{\Sigma f}\right)^2} = \sqrt{\frac{59}{30} - 0.5^2} = 1.3102...$$

Only the division part needs to be reversed as the substraction will have no effect on the standard deviation.

$$\sigma x = \frac{\sigma h}{20}$$

$$\sigma h = 20\,\sigma x = 26.204$$

The mean height of the elephants, $\bar{h} = 260$ cm
The standard deviation of the elephants' height, $\sigma h = 26.2$ cm

Examiner Tips and Tricks

- Be careful when using the formulae for the mean and standard deviation with coded summary statistics, you must make sure that you use the summary statistics consistently throughout. For example, if you use the sum of the coded data squared in the formula for the standard deviation, you must subtract the square of the coded mean.

Statistical Modelling

What is a statistical model?

- A **statistical model** allows you to use mathematics to model **real-life situations**
  - You could model the temperature throughout of a city throughout a month
  - You could model the sleeping times of a baby
- There are **advantages** to using a statistical model
  - It **simplifies** the complicated real-life situation
  - It can be made **quickly** and **easily**
  - It can be used to make **predictions** for real-life
- There are also **things to consider** when using a statistical model
  - It does **not consider all** the real-life **features** from the situation
  - It might only be applicable for **specific scenarios**
  - It might **not provide accurate predictions** for the future

What are the stages of a statistical model?

- **Stage 1**
  A real-life problem is identified
- **Stage 2**
  An initial statistical model is designed
- **Stage 3**
  The model is used to make predictions
- **Stage 4**
  Real-life data is collected
- **Stage 5**
  Comparisons are made between the expected values from the model and the observed values from the data
- **Stage 6**
  Consideration of the data selection and collection processes alongside statistical tests are used to assess the validity of the model
- **Stage 7**
  The model is adjusted and improved (if necessary) based on the results

Examiner Tips and Tricks

Questions on statistical modelling are rare but it is worth remembering the steps. Easy marks if it comes up.

Your notes