Measures of centre and spread

Year 12 General Maths
Units 3 and 4

www.maffsguru.com

Learning Objectives

By the end of the lesson, I would hope that you have an understanding and be able to apply to questions the following concepts:

- To be able to understand the mean and the median as measures of centre.
- To be able to understand the range, interquartile range and standard deviation as measures of spread.
- To be able to know whether to use the median and interquartile range, or the mean and standard deviation, for a particular distribution.
- To be able to use a CAS calculator to calculate these summary statistics.

Recap

In previous lessons we have looked at a number of key topics which will form the building blocks for the rest of the Year 12 General Maths course. We have discovered what the difference between categorical and numerical data is and how to sub-classify each of those. We know how to display and describe categorical and numerical data. We know about dot plots and stem and leaf plots and what they can be used for. Finally we have looked at log scales.

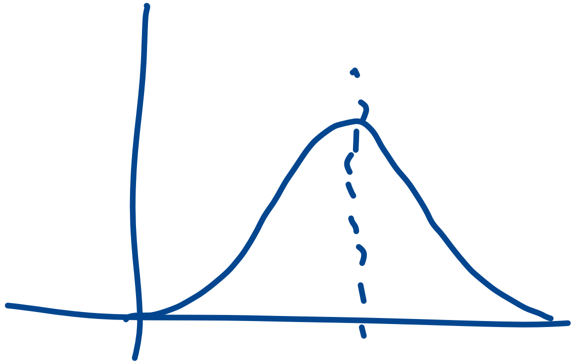
One of the more important sections of the course looks at finding the mean, median, range, and interquartile range as this can tell us a lot about the data we are trying to investigate and compare.

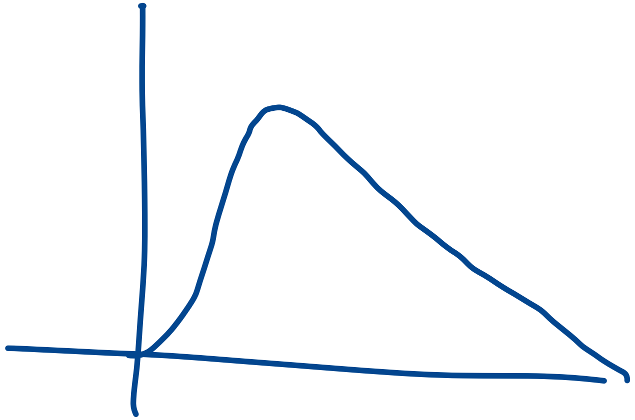
Why do we have different measures of centre and spread?

The two main measures of centre we use are mean and median. Which one we use depends on how the data is distributed. When the data is symmetrical, we can use either the mean or the median. When the data is skewed we tend to prefer to use the median as the mean can be thrown off by too many large or small values.

When we are looking at how spread out the data is, we are looking to use the range or the interquartile range.

The interquartile range tells us the spread (or variability) of the middle 50% of the data. This is less likely to be affected by outliers. When there are outliers we can see that the range can be massively affected.





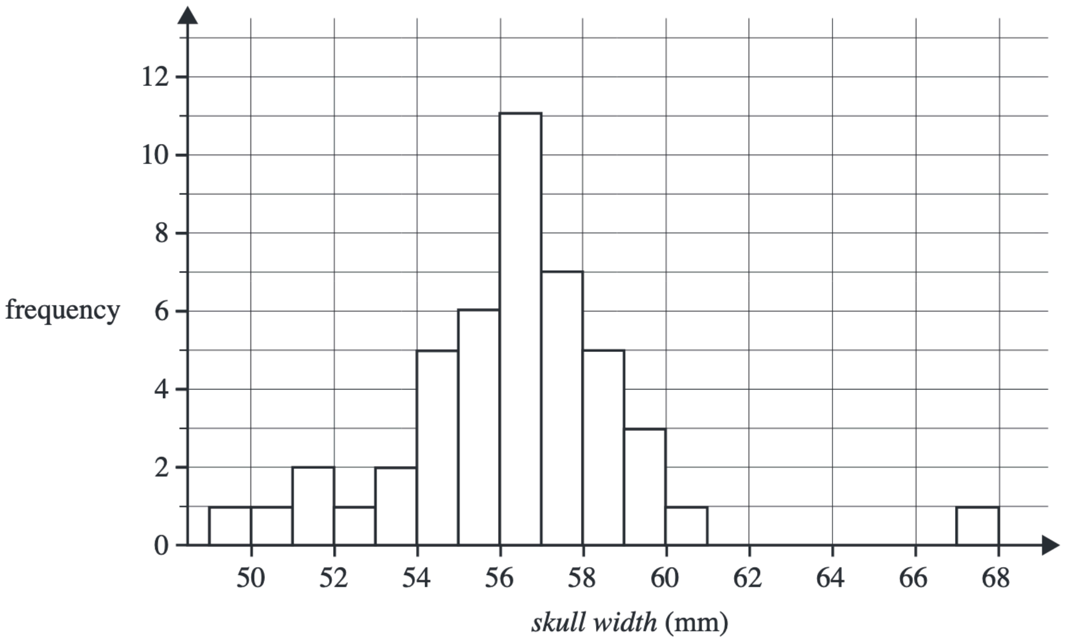
Recap: The problem with histograms

When we draw a histogram, we are losing something which might be important to us; data.

Dot plots and stem plots retain the raw data and hence allow us to accurately calculate statistical values.

Histograms can only provide is with approximate values.

The chart shown was used in the 2022 Further Maths exam. It was question 1. And it caused **a lot** of upset!



Data: adapted from DB Lindenmayer et al., 'Morphological variation among populations of the mountain brushtail possum, *Trichosurus caninus* Ogilby (Phalangeridae: Marsupialia)', *Australian Journal of Zoology*, 43(5), 1995, p. 453

Finding the median: By hand

Remember: The median is the **middle number in an ordered list.**

Let's look at an example of how to find it.

Order each of the following data sets, locate the median, and then write down its value.

- 2 9 1 8 3 5 3 8 1
- 10 1 3 4 8 6 10 1 2 9

med = 3

1 1 2 3 3 5 8 8 9

1 1 2 3 4 6 8 9 10 10

med = 5

$$\text{middle} = \frac{\text{add together}}{2}$$

Note:
The median can be found by hand or using the CAS. The number of data items will dictate which will be the quickest (and most accurate!).

Finding the median: Using the CAS

Remember: The median is the **middle number in an ordered list.**

Find the median of the following data set using the CAS.

- 2 9 1 8 3 5 3 8 1

med = 3

Note:
The median can be found by hand or using the CAS. The number of data items will dictate which will be the quickest (and most accurate!).

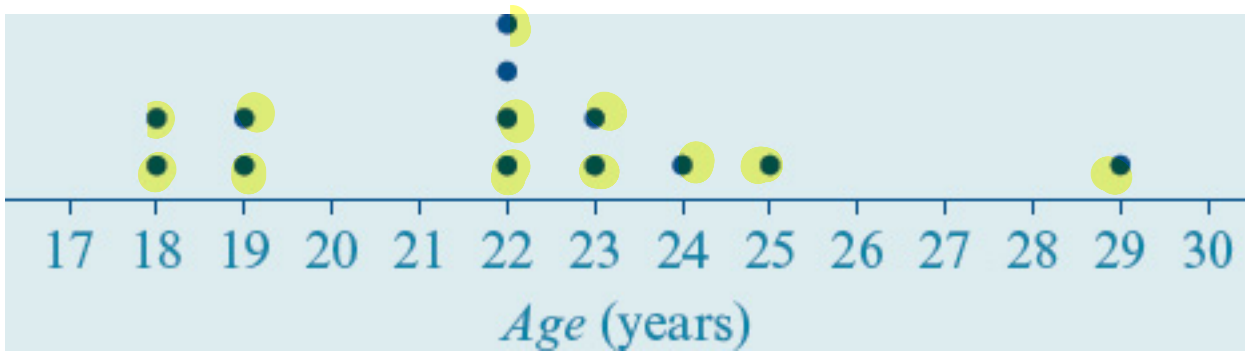
Finding the median: From a dot plot

Finding the median from a dot plot is much easier.

Remember, we can either strike dots from each end until we meet the middle, or use the formula on the right to find the **position** of the middle dot.

The dot plot below displays the age distribution (in years) of the 13 members of a local cricket team. Determine the median age of these cricketers and mark its location on the dot plot.

med = 22



Note:
Dot plots have already placed the data into a nice order for us so makes it simpler.

Note:
Position of the middle dot (or number) can be found using:

$$position = \frac{n+1}{2}$$

Where ‘n’ is the number of terms or items in the list

$$pos = \frac{13+1}{2} = \underline{\underline{7}}$$

Finding the median: From a stem plot

Finding the median from a stem plot is much like finding the median from a dot plot.

Remember, we can either strike numbers from each end until we meet the middle, or use the formula on the right to find the **position** of the middle dot.

The stem plot opposite displays the maximum temperature (in ∘C) for 12 days in January. Determine the median maximum temperature for these 12 days.

$25 \rightarrow 27$

Key: 0|8 = 8°C

1	8	9	9				
2	0	2	5	7	8	9	9
3	1	3					

med = 26°C

Note:
Dot plots have already placed the data into a nice order for us so makes it simpler.

Note:
Position of the middle dot (or number) can be found using:

$$position = \frac{n+1}{2}$$

Where 'n' is the number of terms or items in the list

$n = 12$

$pos = \frac{12+1}{2} = 6.5$

Finding the range

We have been doing this since the start of Year 7.

The range is the **difference between the largest and smallest data item.** Remember, the numbers need to be placed in numerical order.

The stem plot opposite displays the maximum temperature (in ∘C) for 12 days in January. Determine the temperature range over these 12 days.

Exam Trick:
Data is rarely given in order. People make silly mistakes by not ordering the data first.

Key: 0|8 = 8°C

1	8	9	9				
2	0	2	5	7	8	9	9
3	1	3					

Range = 33 − 18

= 15°C

The interquartile range

There is a problem when using the range: **It is subject to outliers.**

This means that a really large value will completely throw the range. This wouldn't be a good measure of spread.

Hence, we tend to like the interquartile range which is the spread of **the middle 50% of the data.** This is much less prone to outliers.

Find the interquartile range of the weights of the 18 cats whose weights are displayed in the ordered stem plot below.

Weight (kg) 1 | 2 represents 1.2 kg

1	9					
2	1	3	5	8		
3	0	0	4	9	9	
4	0	4	5	8		
5	0	3				
6	3	4				

LQ

UQ

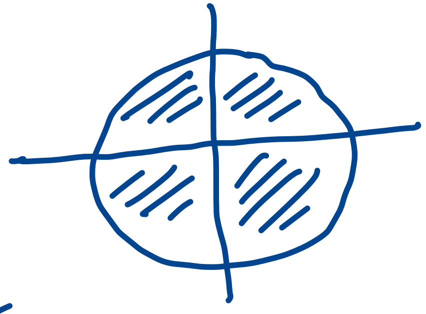
$pos = \frac{9+1}{2} = 5$

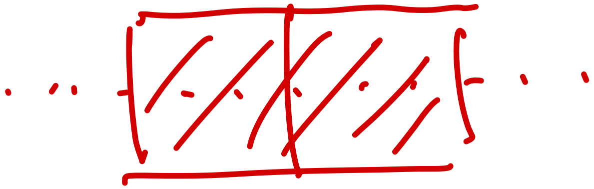
$IQR = UQ - LQ$

$= 48 - 28$

$= \underline{\underline{20}}$

$pos = \frac{18+1}{2} = \underline{\underline{9.5}}$





Notice:
A stem plot always has a key. This tells you what each number stands for.

Most people forget to place the key when drawing these and will lose marks in the exam and SACs.

Another example of finding the interquartile range

Here is another example of how to find the interquartile range.

The stem plot shows the life expectancy (in years) for 23 countries. Find the IQR for life expectancies.

Note: On its own the range or IQR doesn't really tell us anything. The power is when we compare data sets.

Notice:
A stem plot always has a key. This tells you what each number stands for.

Most people forget to place the key when drawing these and will lose marks in the exam and SACs.

Stem: 5|2 = 52 years

5	2								
5	5	6							
6	4								
6	6	6	7	9					
7	1	2	2	3	3	4	4	4	4
7	5	5	6	6	7	7			

LQ

UQ

$$\frac{23+1}{2} = 12$$

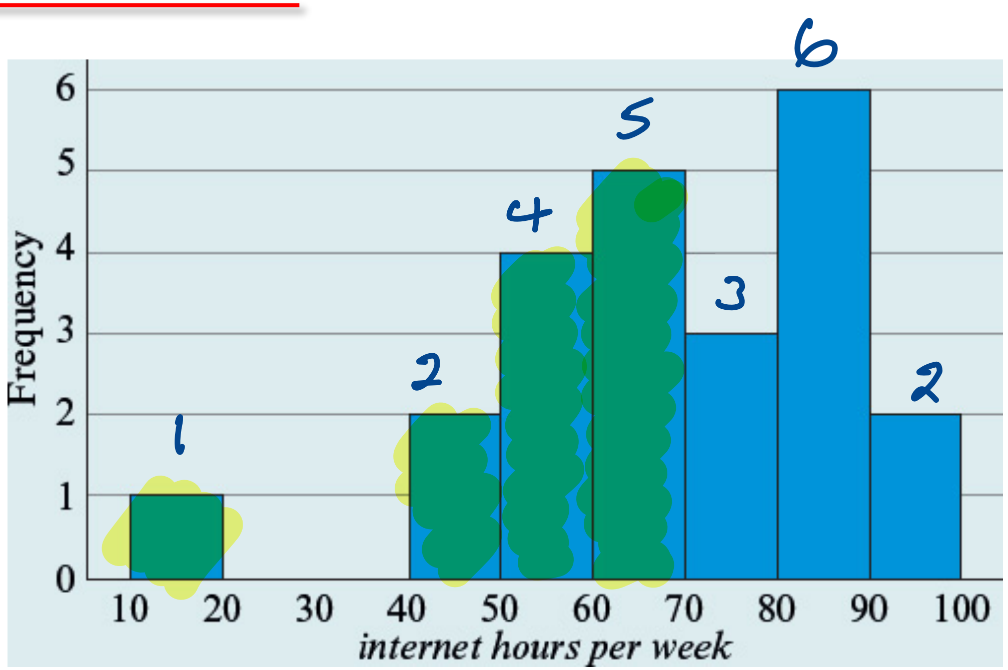
$$pos = \frac{11+1}{2} = 6$$

$$IQR = 75 - 66$$
$$= \underline{\underline{9}}$$

Finding the median and quartiles from a histogram

The histogram shows the average number of hours per week a group of 23 people spent on the internet. Find possible values for the median and quartiles of this distribution.

Note: It means the Upper and Lower Quartiles



$$pos = \frac{23+1}{2} = 12^{th}$$

med : 60 - 70 hrs/wk.

LQ = 50 - 60 hrs/wk

UQ = 80 - 90 hrs/wk

11 11

$$pos = \frac{11+1}{2} = 6$$

Finding the mean of data

The mean of a set of data is what people call the average.

Again, we have been working out how to find the average of data since Year 7.

We can express the mean in two ways; with words and mathematical symbols.

$$mean = \frac{sum\ of\ all\ data\ items}{number\ of\ data\ items}$$

Or

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

Sum

data items

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

mean

num of data items

Example: Finding the mean of data

The following is a set of reaction times (in milliseconds):

38 36 35 43 46 64 48 25

Write down the values of the following, correct to one decimal place.

- n
- $\sum x$
- $\overline{x}$

$n = \underline{\underline{8}}$

$\Sigma x = 335$

$\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{335}{8}$

$= \underline{\underline{41.9}}$ ms

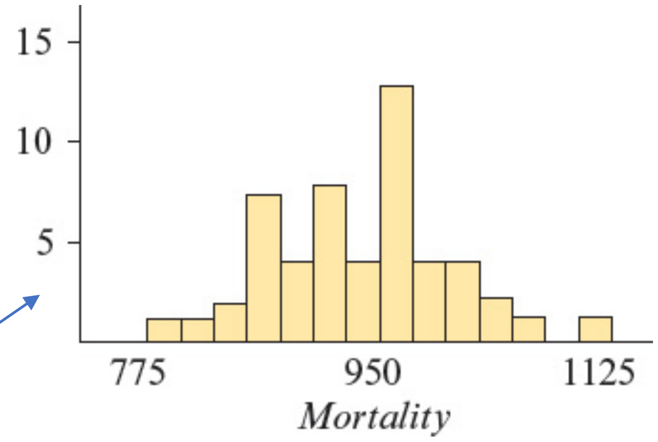
The relationship between the median and the mean

This is really important as I have seen it featured in SACs.

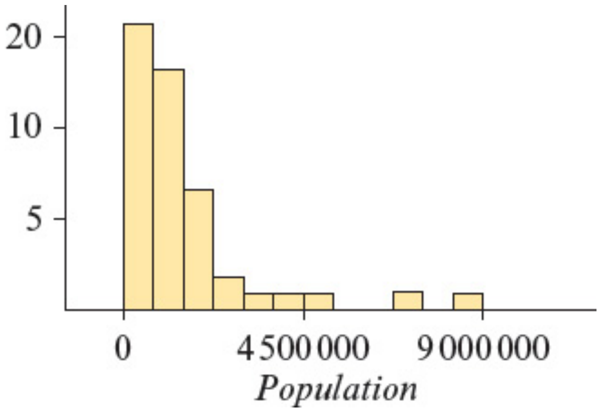
When data is **approximately symmetrical**, we can discover that the mean and median will be pretty close in terms of value.

When data is skewed, then the mean and median will be very different.

The mean values will be affected by the large and small values of the tail.



Mean ≈ 940 per 100 000 people
Median ≈ 944 per 100 000 people



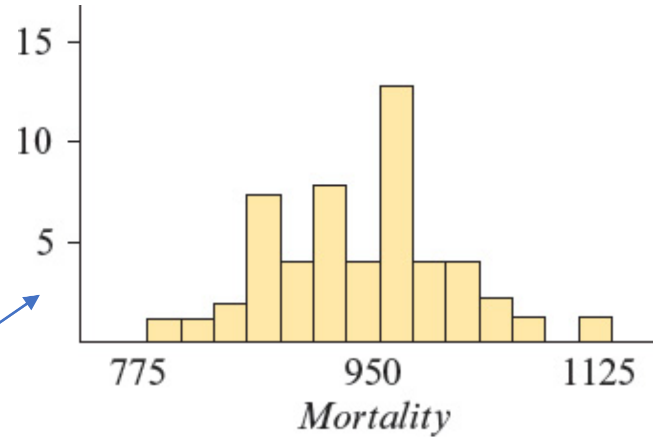
Mean ≈ 1.4 million
Median ≈ 0.9 million

The relationship between the median and the mean

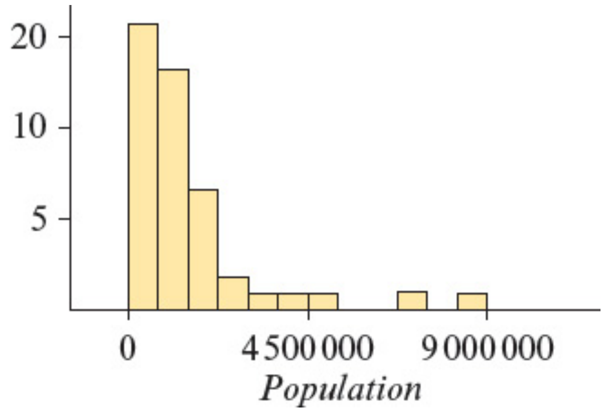
The median is known as a resistant statistic as it's not affected by extreme values.

We tend to use this over the mean.

If data is symmetric we can use either!



Mean ≈ 940 per 100 000 people
Median ≈ 944 per 100 000 people



Mean ≈ 1.4 million
Median ≈ 0.9 million

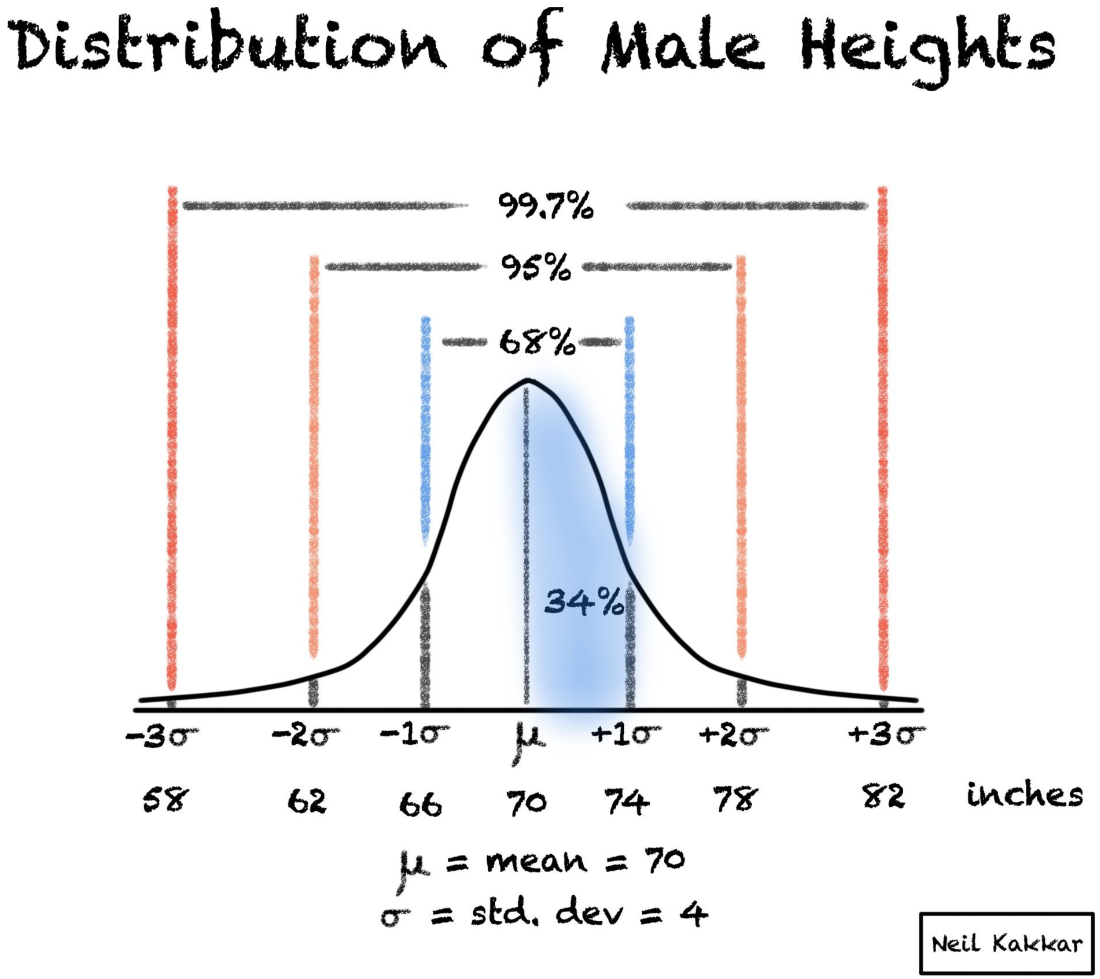
The standard deviation

When we use the interquartile range, we split the data into quarters.

There are other ways of splitting data.

Standard deviation is another one. We know, from maths more complex than we need to know, that we can split data with his normally distributed into 6 (six) main sections.

VCE results are obtained from a bell curve (which is a normal distribution). This can be split into 6 main sections. The width of each section is called a standard deviation.

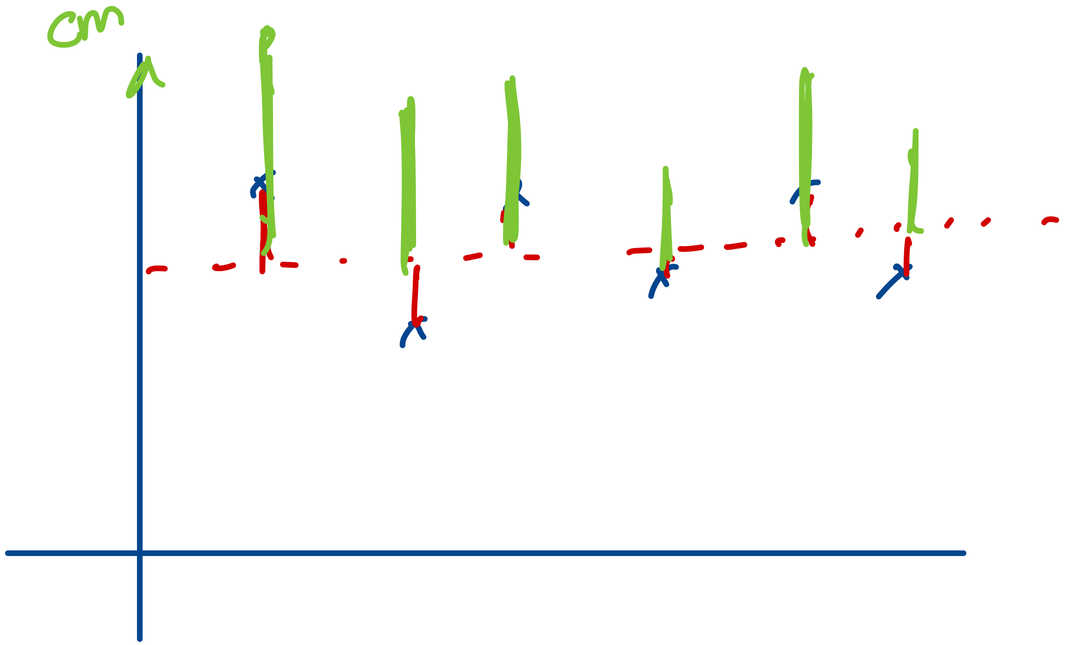


But what is the standard deviation

The standard deviation is an average of the squared deviations of each data value from the mean.

I can explain this with the use of a diagram and some logic.

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$



Calculating the standard deviation

There is a formula we can use to find the standard deviation of a set of data:

$$s = \sqrt{\frac{\sum(x - \overline{x})^2}{n - 1}}$$

Looks complicated doesn't it.

Well. It is!

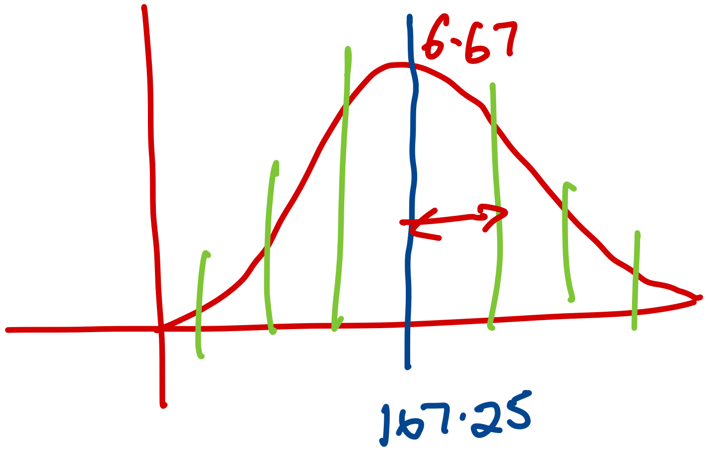
The good news is … you can use your CAS to find the standard deviation and won't have to worry about doing it using the formula.

Calculating the standard deviation: Using the CAS

The following are the heights (in cm) of a group of women.

176	160	163	157	168
172	173	169		

- Determine the mean and standard deviation of the women's heights. Give your answers correct to two decimal places.



$\bar{x} = 167.25$ cm

sd = $\underline{\underline{6.67\text{cm}}}$

VCAA Questions

VCAA 2022 Further Maths
Exam 1

Question 3

The third quartile (Q_3) for this distribution, in millimetres, could be

A. 55.8
B. 56.2
C. 56.9
D. 57.7
E. 58.3

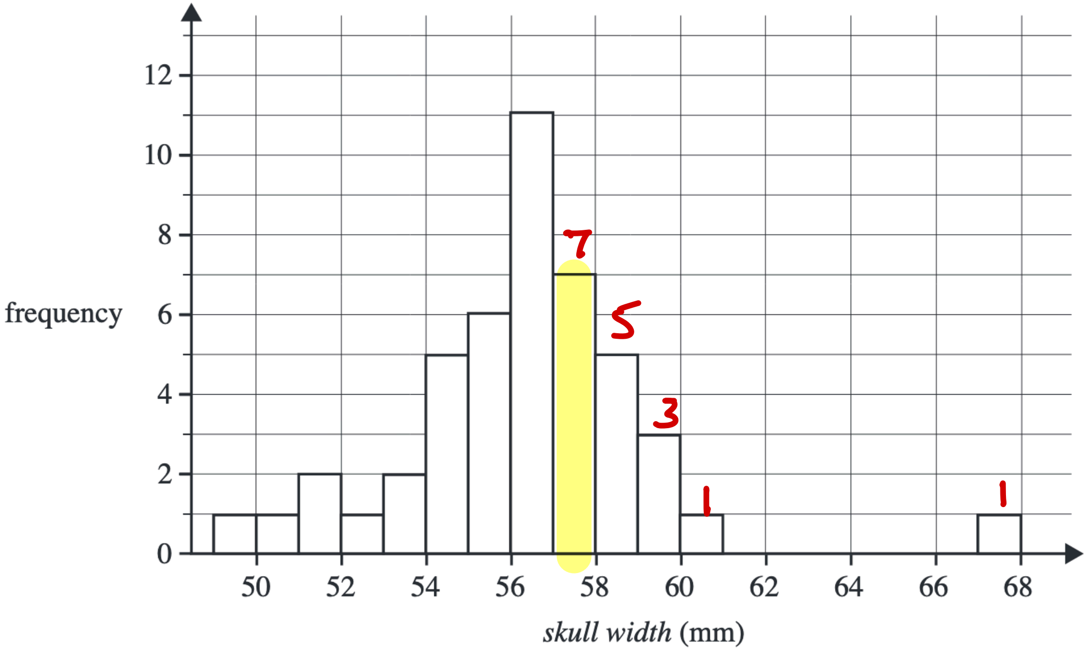
23 | 23 ← 12

The histogram below displays the distribution of *skull width*, in millimetres, for 46 female possums.

$n = 46$

$pos = \frac{46+1}{2} = 23.5$

57 → 58



Data: adapted from DB Lindenmayer et al., 'Morphological variation among populations of the mountain brushtail possum, *Trichosurus caninus* Ogilby (Phalangeridae: Marsupialia)', *Australian Journal of Zoology*, 43(5), 1995, p. 453

VCAA Questions

pos = $\frac{30+1}{2}$

= 15.5

Question 1 (6 marks)

The back-to-back stem plot below displays the distribution of daily maximum *wind speed*, in kilometres per hour, recorded at a weather station in April and November 2021.

key: 4 | 2 = 42

wind speed (km/h)

April 2021 $n = 30$		November 2021 $n = 30$
3 3 3 3 3 1 1 1	1	
9 7 5 5	1	5 7 9 9
2 0 0 0 0 0	2	0 0 2 2 4 4 4
8	2	6 8 8 8 8
3 3 3 1 1 1 1	3	0 0 0 1 3
9 9	3	5 5 9
	4	3
8	4	6 8
4	5	4
	5	9
	6	
	6	
	7	0

a. For April 2021, determine

i. the median *wind speed*, in kilometres per hour 20 1 mark

VCAA 2020 Further Maths
Exam 2

Question 1 (3 marks)

Body mass index (*BMI*), in kilograms per square metre, was recorded for a sample of 32 men and displayed in the ordered stem plot below.

key: 21 | 6 = 21.6 $n = 32$

21	6	9	9				
22	1	2	5	6			
23	0	1	4	6	6	7	8
24	4	5	6	7	7	9	
25	6	8					
26	1	7	9				
27	3	7					
28	2						
29	1	8					
30	4						
31	1						

$n = 32$

positively skewed

a. Describe the shape of the distribution. 1 mark

www.maffsguru.com