KING EDWARD VI
HANDSWORTH GRAMMAR
SCHOOL FOR BOYS

KING EDWARD VI
ACADEMY TRUST
BIRMINGHAM

Year 12
Statistics 1
Chapter 2 – Measures of Location and Spread

HGS Maths

Dr Frost Course

Name: ____________________________________

Class: ____________________________________

Contents

1.4 Types of Data

Notes

Types of Data

Qualitative/Categorical

Non-numerical values, e.g. colour.

Quantitative

Numerical values.

Discrete

Can only take specific values, e.g. shoe size, number of children.

Note that while discrete variables only allow specific values, the range could still be infinite, e.g. "number of attempts before success".

Continuous

Can take any decimal value (possible with a specified range).

Data can be **grouped** for conciseness, at the expense of losing the exact original values.

Weight w (kg)	Frequency
$0 \leq w < 20$	3
$20 \leq w < 70$	4

$$20 \leq w < 70$$

This is known as a **class interval**.

Lower class boundary

Midpoint **= 45**

Upper class boundary

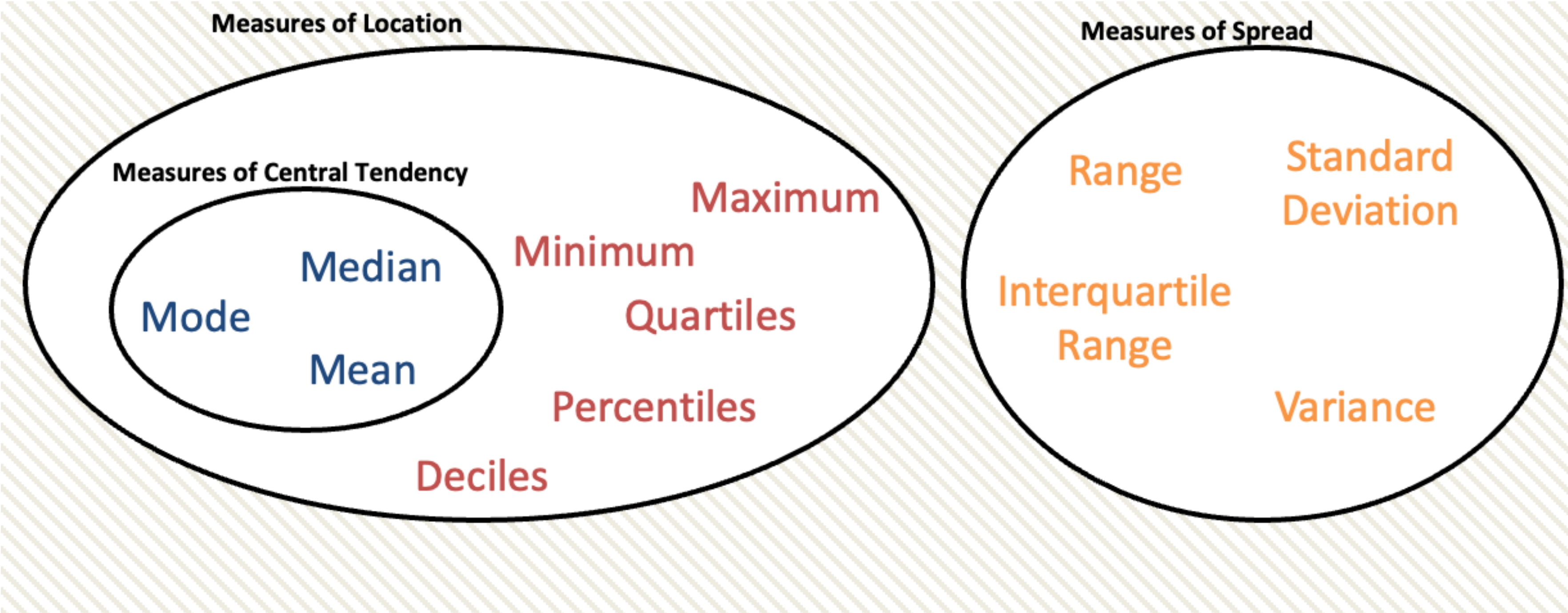
Class width = $70 - 20 = \mathbf{50}$

Worked Example	Your Turn
State the type of data: a) Type of tree b) Number of people on a train c) Time required to run 200m	State the type of data: a) Human shoe size measured as 1, 2 or 3 etc. b) Height of a tree c) Favourite colour

Worked Example	Your Turn
Which of the following are examples of discrete data: • Number of ducks in a town • Heights of mountains • Time taken to repair a car • Weights of counters in a bag • Number of counters in a sack • Number of heads when you toss 40 coins	Which of the following are examples of discrete data: • Masses of cats on a farm • Ages of assistants in a company • Number of counters in a bag • Number of sheep in a town • Time spent queueing at a doctor's office • Weights of beads in a pocket

Dr Frost 344b

Measures of…



Measures of location are single values which describe a **position** in a data set.

Of these, **measures of central tendency** are to do with the **centre of the data**, i.e. a notion of 'average'.

Measures of spread are to do with **how data is spread out**.

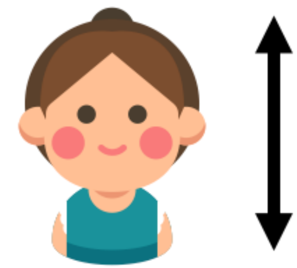
2.1 Measures of Central Tendency

Notes

Understanding Statistical Variables

x is the height, in cm, of female athletes in a running competition.

$$x = [162, 178, 150, 160, 160, 170]$$



We use **lowercase** letters for statistical variables.

x is technically not a set, because sets cannot contain duplicates, but statistical variables can.

We can refer to specific values in the collection using x_i.
For example, $x_6 = 170$

In statistics, a data variable is a **collection of data values** recorded for the **same quantity**, e.g. the weights of people in a room, or categorical values such as the favourite colours of students in a class.

n is used to refer to the **number of values in the collection**.
Here, $n = 6$

Operations on Statistical Variables

$$x = [162, 178, 150, 160, 160, 170]$$

List-to-list operations:

$$x + 2 = [164, 180, 152, 162, 162, 172]$$

We can <u>code</u> variables, which produces a new list with the operation, e.g. "+2" applied to each value in the collection. We will explore this in skill **545**.

List-to-number operations:

$$\sum_{i=1}^{n} x_i = 980$$

Σ means <u>sum</u>. This expression means, "the sum, as i varies from 1 to n, of x_i", i.e. the sum of all the values in the collection, $x_1 + x_2 + \cdots + x_n$

$$\Sigma x = 980$$

However, we typically write Σx as shorthand.

$$\bar{x} = 163.3$$

$\bar{x}$, said "x bar", determines the mean of the variable.

Formula for Mean, Using Σ

Recall that mean is calculated by **summing the values** and dividing by the **number of values.**

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

x	162	178	150	160	160	170

$$\Sigma x = 980$$

$$n = 6$$

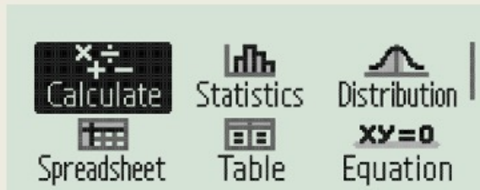
$$\bar{x} = \frac{980}{6} = 163.3$$

Mean and Σx on a Calculator

x	162	178	150	160	160	170

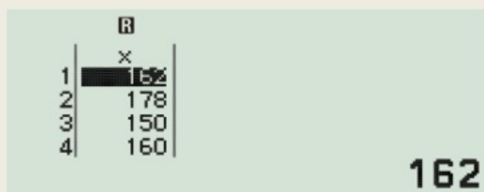
These are instructions for the **Casio fx-570/991CW**

1 Use the arrows and OK to select Statistics.

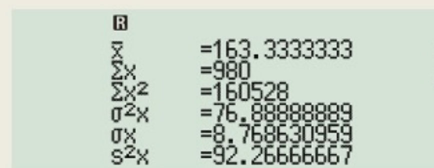


2 Choose 1-Variable.

3 Enter your values, pressing = after each value. Press = again after the last value.



4 Select **1-Var Results** and read off the values of Σx and n and $\bar{x}$.



Mean for Ungrouped Frequency Tables, Using Σ

The number of children x in different families is recorded.

Num children (x)	0	1	2	3	4	5
Frequency (f)	6	8	14	9	2	1

Determine the mean number of children per family, $\bar{x}$.

Since the frequency tells us how many times each value occurs, the data when listed out in full would be as follows:

0 0 0 0 0 0 1 1 1 1
1 1 1 1 2 2 2 2 2 2
2 2 2 2 2 2 2 2 3 3
3 3 3 3 3 3 3 4 4 5

This value of 2 appears 14 times. Therefore, the total of these 2's is

$fx = 14 \times 2 = 28$

$\boldsymbol{\Sigma f x}$ therefore represents the total of all these products, and thus the sum of x.

The number of values is the **total frequency**, i.e. $\boldsymbol{n = \Sigma f}$

Therefore:

For a grouped frequency table

$$\bar{x} = \frac{\Sigma f x}{\Sigma f}$$

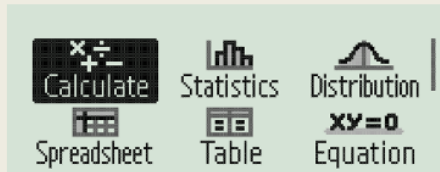
Dr Frost 532c

Ungrouped Mean on a Calculator

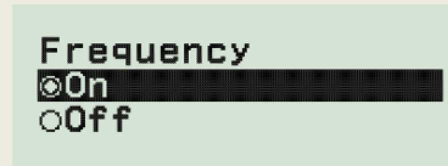
Num children (x)	0	1	2	3	4	5
Frequency (f)	6	8	14	9	2	1

These are instructions for the **Casio fx-570/991CW**

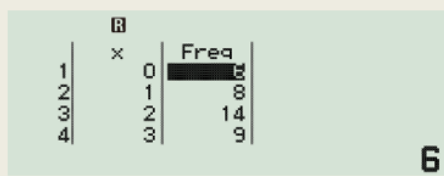
1 Use the arrows and OK to select Statistics. Choose 1-Variable.



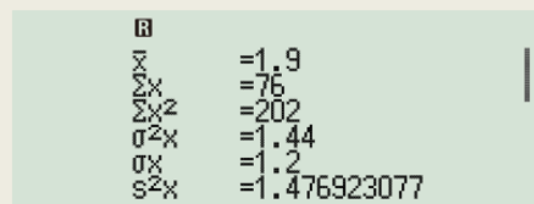
2 Press **Tools** than choose **Frequency**. Choose **On**.



3 Press the ← button. Enter your values, pressing = after each value. Use the arrow keys to navigate back to the top. Press = again after the last value.



4 Select **1-Var Results** and read off the values of $\bar{x}$. Note that $\boldsymbol{\Sigma x}$ **represents** $\boldsymbol{\Sigma f x}$ (since the x in Σx means the original data with duplicate values considered)



Worked Example	Your Turn

Worked Example

Times, x, have been rounded to the nearest minute.

a) Write down the modal class.

b) Write down the class containing the median.

c) Find an estimate for the mean time.

Time, x	**Frequency**
$0-2$	5
$3-5$	2
$6-10$	3

Your Turn

Times, x, have been rounded to the nearest minute.

a) Write down the modal class.

b) Write down the class containing the median.

c) Find an estimate for the mean time.

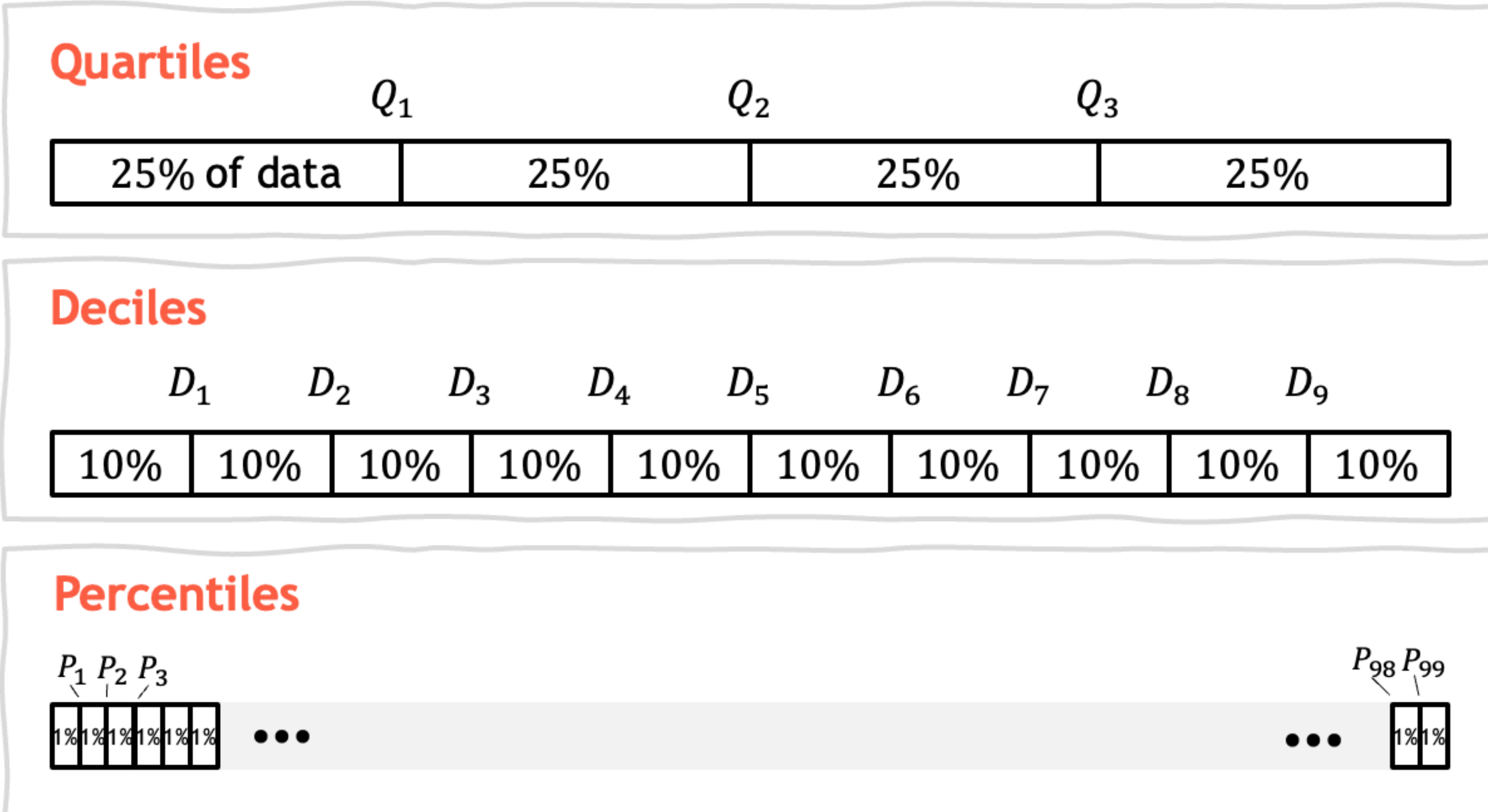
Time, x	**Frequency**
$0-3$	7
$4-8$	11
$9-10$	2

2.2 Other Measures of Location

Notes

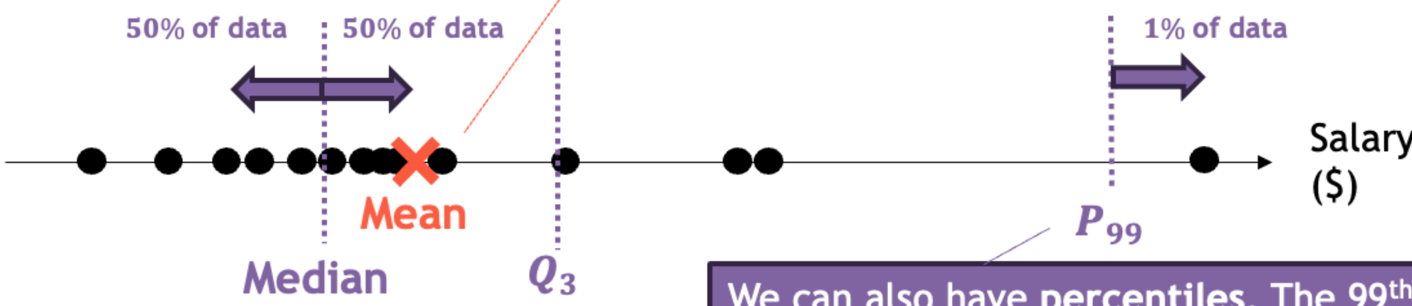
Names for Quartiles

Quantiles represent general positions across the data when ordered.



Mean vs Measures of Position

We understand mean as an ‘average’ that considers all values, but is skewed by extreme values.



We can also have **percentiles**. The 99th percentile (written P_{99}) is the value 99% along the data. When ‘the 1%’ is used in the media, it’s referring to people in the ‘top percentile’ for salary, i.e. above P_{99}.

Median is a **measure of position** because it allows us to get the value a certain position, in this case 50%, along the data.

Another measure of position is **quartiles**, which gives us the value 25%, 50%, and 75% along the data.

$Q_1 = 25\%$ along data (**lower quartile**)
$Q_3 = 75\%$ along data (**upper quartile**)

What Item to Use for Listed Data?

Items	n	Position of median	Median
1,4,7,9,10	5	3rd	7
4,9,10,15	4	2nd/3rd	9.5
2,4,5,7,8,9,11	7	4th	7
1,2,3,5,6,9,9,10,11,12	10	5th/6th	7.5

Can you think of a rule to find the position of the median given the number of values, n?

To find the <u>**position of the median**</u> for listed data, calculate $\frac{n}{2}$:

- If a decimal, round up.
- If whole, use the midpoint between this item and the one after.

What Item to Use for Grouped Data?

IQ of L6Ms2 (q)	Frequency (f)
$80 \leq q < 90$	7
$90 \leq q < 100$	5
$100 \leq q < 120$	3
$120 \leq q < 200$	2

If the data is grouped, what item do we use for the median?

$$\frac{17}{2} = \mathbf{8.5^{th}\ item}$$

To estimate the median of grouped data, calculate $\frac{n}{2}$, then use linear interpolation.

Important point: Unlike with listed values, for grouped data, do not round $\frac{n}{2}$ in any way. For example, if we were reading off a value from a cumulative frequency graph and there were 100 values, for the median we'd read across the $\frac{100}{2} = 50^{th}$ item mark, not halfway between the 50^{th} and 51^{st}.

Estimating Frequencies

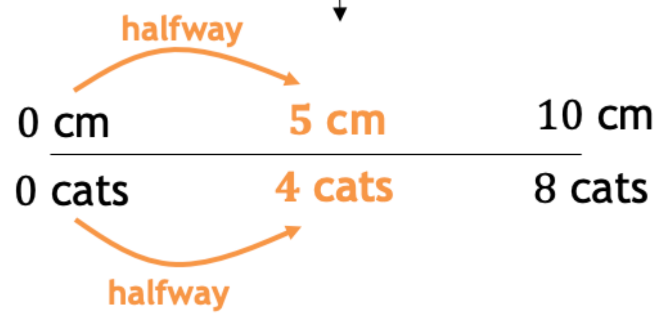
The table shows the heights of various cats.

Height h (cm)	Frequency
$0 \leq h \leq 10$	8
$10 \leq h \leq 20$	60

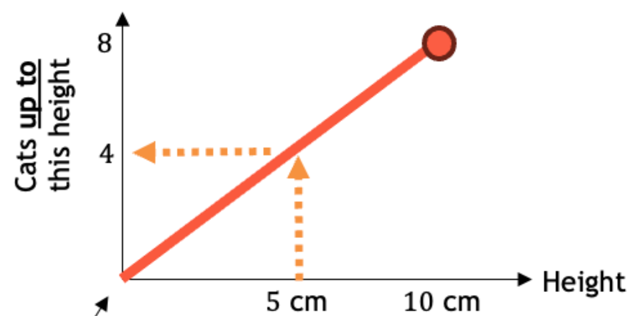
Estimate the number of cats with a height:

- a Below 5 cm
- b Above 15 cm
- c Below 12.5 cm

We are assuming the 8 cats are **equally distributed** between 0 and 10 cm, so that there are 4 between 0 – 5 cm and 4 between 5 – 10 cm.

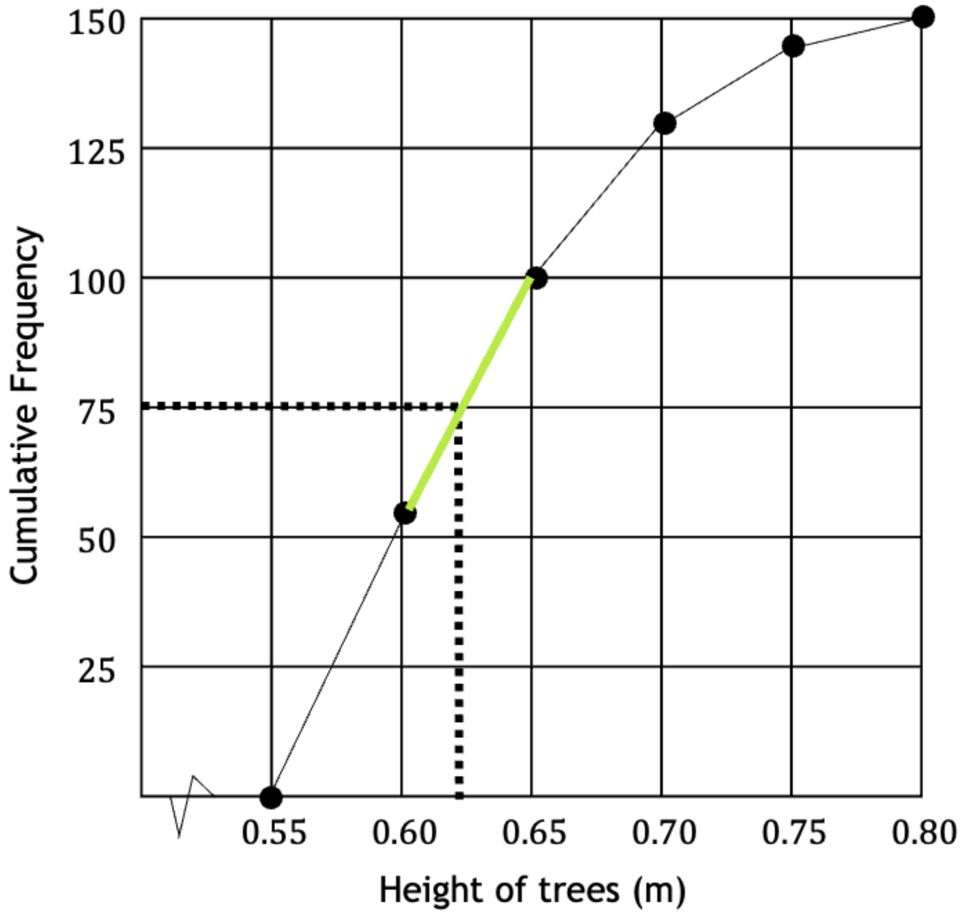


- a **4**
- b **30**
- c $8 + 15 = \mathbf{23}$



This is known as **linear interpolation**, because if we were plotting this as a cumulative frequency graph (i.e. the running total of cats up to a specific height), **the graph forms a straight line**.

Linear Interpolation



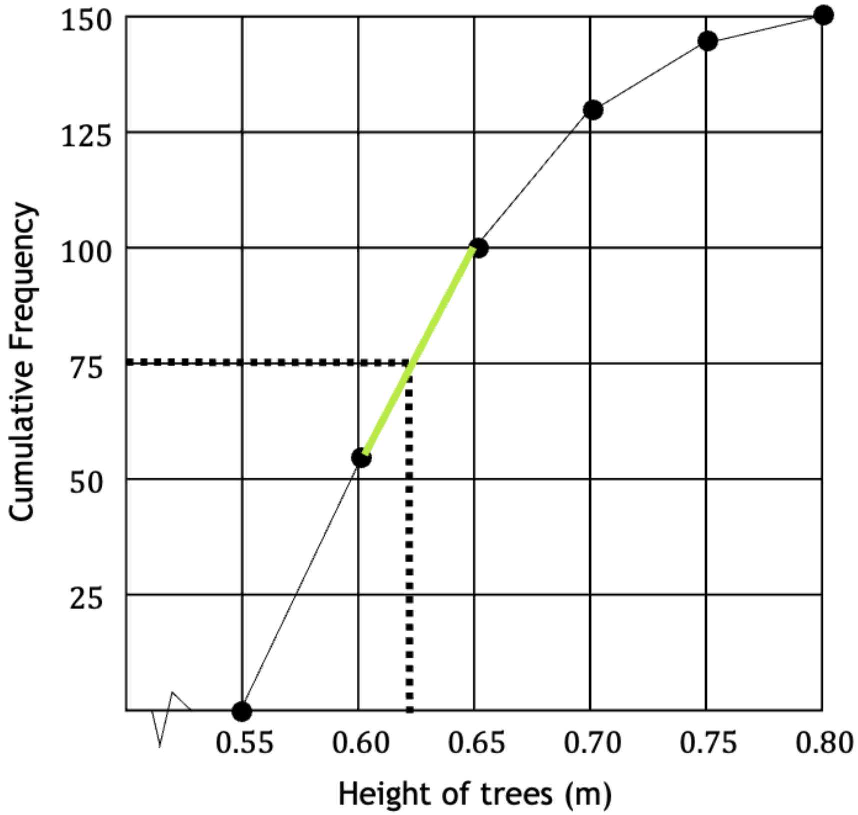
Height of tree (m)	Freq	C.F
$0.55 \leq h < 0.6$	55	55
$0.6 \leq h < 0.65$	45	100
$0.65 \leq h < 0.7$	30	130
$0.7 \leq h < 0.75$	15	145
$0.75 \leq h < 0.8$	5	150

Using a cumulative frequency graph, we know we can estimate the median by drawing a suitable line.

How could we read off this value exactly using a suitable calculation?

We could find the fraction of the way along the line segment using the frequencies, then go this same fraction along the class interval.

Linear Interpolation



Height of tree (m)	Freq	C.F
$0.55 \leq h < 0.6$	55	55
$0.6 \leq h < 0.65$	45	100
$0.65 \leq h < 0.7$	30	130
$0.7 \leq h < 0.75$	15	145
$0.75 \leq h < 0.8$	5	150

Using linear interpolation, estimate the median.

Step 1: Identify the interval in which the median item, here the $\frac{150}{2} = 75^{\text{th}}$ value, lies.

Step 2: Write the relevant data needed to make the calculation. We recommend below.

Frequency up until this interval →	55	75 ← Item number we're interested in.	100 ← Frequency by end of this interval
Height at start of interval. →	0.6 m	Q_2	Height by end of interval. → 0.65 m

Linear Interpolation

Using linear interpolation, estimate the median.

Height of tree (m)	Freq	C.F
$0.55 \leq h < 0.6$	55	55
$0.6 \leq h < 0.65$	45	100
$0.65 \leq h < 0.7$	30	130
$0.7 \leq h < 0.75$	15	145
$0.75 \leq h < 0.8$	5	150

Frequencies go at the top.

20

45

Frequency up until this interval → 55 $\xrightarrow{\frac{25}{45}}$ 75 ← Item number we're interested in. 100 ← Frequency by end of this interval

Height at start of interval. → 0.6 m $\xrightarrow{\frac{25}{45}}$ Q_2 Height by end of interval. → 0.65 m

And heights at the bottom. You may wish to put units to avoid confusing with your frequencies.

What fraction of the way across the class interval are we?

$\frac{25}{45}$th **of the way**

Step 3: We therefore go the same fraction of the way between 0.6 m to 0.65 m.

$$Q_2 = 0.6 + \left(\frac{20}{45} \times 0.05\right) = \mathbf{0.622\ m}$$

Formula
$\text{lcb} + \dfrac{(Q - \text{cumulative frequency at start of class}) \times \text{class width}}{\text{class frequency}}$ where Q is the position of the required percentile

Worked Example	Your Turn
Estimate the median:	Estimate the median:

Worked Example

Estimate the median:

Score, x	**Frequency**
$0 \leq x < 1$	3
$1 \leq x < 2$	2
$2 \leq x < 4$	1
$4 \leq x < 9.5$	1
$9.5 \leq x < 10$	4

Your Turn

Estimate the median:

Score, x	**Frequency**
$0 \leq x < 1$	11
$1 \leq x < 2$	4
$2 \leq x < 4$	2
$4 \leq x < 9.5$	2
$9.5 \leq x < 10$	8

Dr Frost 544c

Worked Example	Your Turn
Estimate the lower quartile:	Estimate the upper quartile:

Worked Example

Estimate the lower quartile:

Score, x	**Frequency**
$0 \leq x < 1$	3
$1 \leq x < 2$	2
$2 \leq x < 4$	1
$4 \leq x < 9.5$	1
$9.5 \leq x < 10$	4

Your Turn

Estimate the upper quartile:

Score, x	**Frequency**
$0 \leq x < 1$	11
$1 \leq x < 2$	4
$2 \leq x < 4$	2
$4 \leq x < 9.5$	2
$9.5 \leq x < 10$	8

Worked Example	Your Turn
Estimate the 27th percentile:	Estimate the 72nd percentile:

Worked Example

Estimate the 27th percentile:

Score, x	Frequency
$0 \leq x < 1$	3
$1 \leq x < 2$	2
$2 \leq x < 4$	1
$4 \leq x < 9.5$	1
$9.5 \leq x < 10$	4

Your Turn

Estimate the 72nd percentile:

Score, x	Frequency
$0 \leq x < 1$	11
$1 \leq x < 2$	4
$2 \leq x < 4$	2
$4 \leq x < 9.5$	2
$9.5 \leq x < 10$	8

What's Different about the Intervals here?

Weight of cat to nearest kg	Frequency
10 – 12	7
13 – 15	2
16 – 18	9
19 – 20	4

You can spot this by either being aware of the word 'rounded'/'nearest' in the question, or where the endpoints of the intervals don't match, i.e. 'have gaps'.

Because the weights are **rounded** to the nearest kg, a weight of 9.8 kg for example would appear in the 10 – 12 kg interval. What interval does this **actually** represent?

$$10 - 12$$

$$9.5 - 12.5$$

Lower class boundary

Class width = 3

Upper class boundary

Worked Example	Your Turn
Times, x, have been rounded to the nearest minute. Estimate the median:	Times, x, have been rounded to the nearest minute. Estimate the median:

Time, x	Frequency
$0-2$	7
$3-5$	2
$6-10$	3

Time, x	Frequency
$0-3$	7
$4-8$	11
$9-10$	2

Worked Example	Your Turn
Times, x, have been rounded to the nearest minute. Estimate the lower quartile:	Times, x, have been rounded to the nearest minute. Estimate the upper quartile:

Time, x	Frequency
$0-2$	5
$3-5$	2
$6-10$	3

Time, x	Frequency
$0-3$	7
$4-8$	11
$9-10$	2

Worked Example	Your Turn
Times, x, have been rounded to the nearest minute. Estimate the 63rd percentile:	Times, x, have been rounded to the nearest minute. Estimate the 36th percentile:

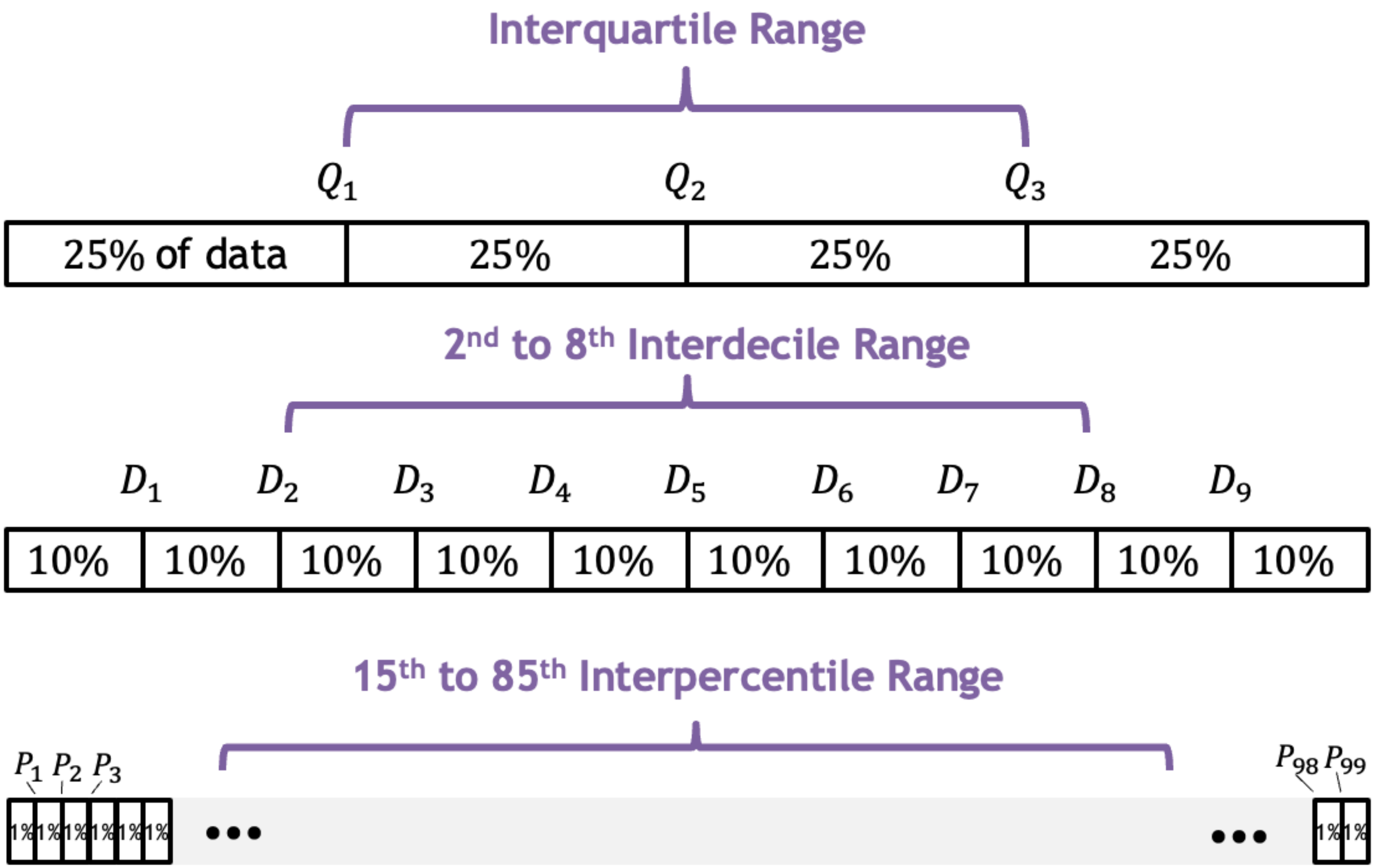
Time, x	Frequency
$0-2$	5
$3-5$	2
$6-10$	3

Time, x	Frequency
$0-3$	7
$4-8$	11
$9-10$	2

2.3 Measures of Spread

Notes

Inter-percentile and Inter-decile Ranges



Just as we can calculate the interquartile range to mean the range of the middle 50% of data, we can also use deciles and percentiles to find the range of a more general middle percentage of the data.

Worked Example	Your Turn

Worked Example

Estimate the interquartile range:

Score, x	Frequency
$0 \leq x < 1$	3
$1 \leq x < 2$	2
$2 \leq x < 4$	1
$4 \leq x < 9.5$	1
$9.5 \leq x < 10$	4

Your Turn

Estimate the interquartile range:

Score, x	Frequency
$0 \leq x < 1$	11
$1 \leq x < 2$	4
$2 \leq x < 4$	2
$4 \leq x < 9.5$	2
$9.5 \leq x < 10$	8

Worked Example	Your Turn
Estimate the 20th – 80th interpercentile range:	Estimate the 10th – 90th interpercentile range:

Worked Example

Estimate the 20th – 80th interpercentile range:

Score, x	Frequency
$0 \leq x < 1$	3
$1 \leq x < 2$	2
$2 \leq x < 4$	1
$4 \leq x < 9.5$	1
$9.5 \leq x < 10$	4

Your Turn

Estimate the 10th – 90th interpercentile range:

Score, x	Frequency
$0 \leq x < 1$	11
$1 \leq x < 2$	4
$2 \leq x < 4$	2
$4 \leq x < 9.5$	2
$9.5 \leq x < 10$	8

Worked Example	Your Turn
Times, x, have been rounded to the nearest minute. Estimate the interquartile range:	Times, x, have been rounded to the nearest minute. Estimate the interquartile range:

Time, x	Frequency
$0-2$	7
$3-5$	2
$6-10$	3

Time, x	Frequency
$0-3$	7
$4-8$	11
$9-10$	2

Worked Example	Your Turn
Times, x, have been rounded to the nearest minute. Estimate the $5^{th} - 95^{th}$ interpercentile range:	Times, x, have been rounded to the nearest minute. Estimate the $15^{th} - 85^{th}$ interpercentile range:

Time, x	Frequency
$0 - 2$	7
$3 - 5$	2
$6 - 10$	3

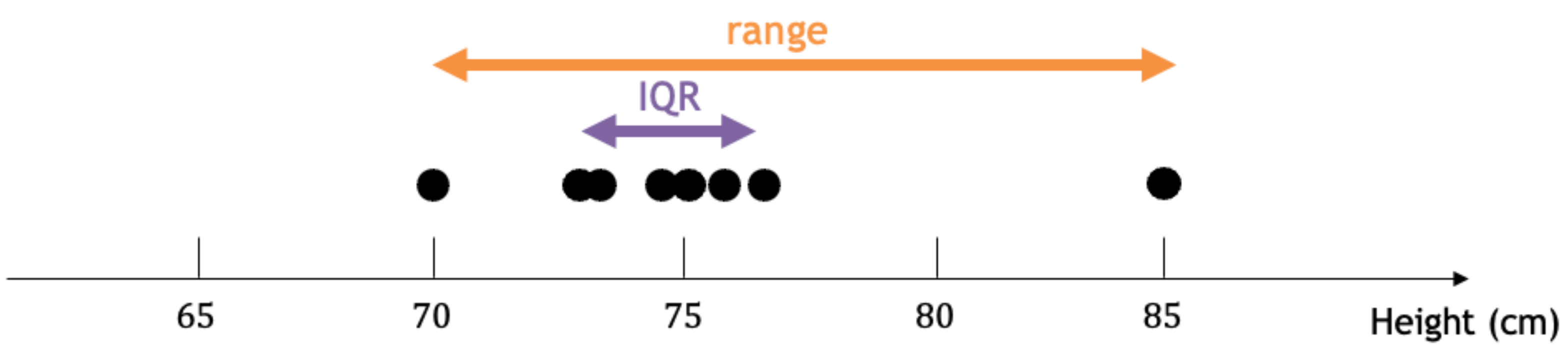
Time, x	Frequency
$0 - 3$	7
$4 - 8$	11
$9 - 10$	2

2.4 Variance and Standard Deviation

Notes

Working Towards a More Useful Measure of Spread

Consider the following heights of people in a room.



1 What is the range of the data?

It is the 'width' of the data, i.e. the total spread.
$85 - 70 = 15$ cm

2 Why is the range misrepresentative of this data?

It is sensitive to outliers. Ignoring the 85 cm person, the range of the remaining data is only 7 cm.

3 What would be a better measure of spread which overcomes this?

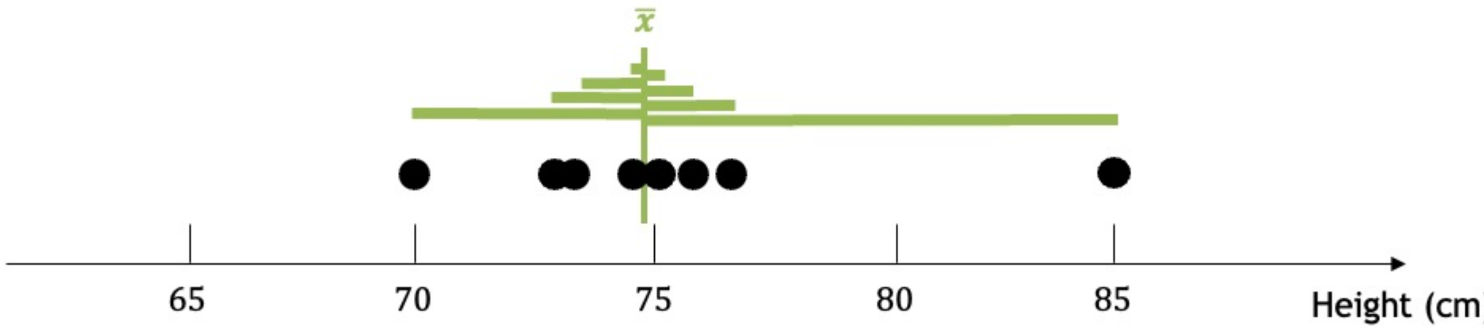
The **interquartile range** finds the range only of the 'core' data in the middle 50%.

4 Why is this still not ideal?

The IQR **still ignores half of the data**; we **want** values far away from the 'average' to have an impact on spread and **consider all values**.

Working Towards a More Useful Measure of Spread

Consider the following heights of people in a room.



We might instead consider the **average distance of values from the mean**. This seems sensible for this data set, because:

- If most of the data is close to the mean, then the average discrepancy (i.e. 'deviation') from the mean will be low.
- It considers **all data values**.
  The outlier of 85 cm will increase this measure of spread, although its impact will be limited as it is only one value as part of an average spread.

The difference between each data value x and some reference point, in this case $\bar{x}$, is known as a **deviation** $x - \bar{x}$. The deviation might be positive (if x is above the mean) or negative (if below).

Mean Absolute Deviation

Mean absolute deviation (MAD) is a measure of spread that gives the mean deviation from the data set's mean value.

$$MAD = \frac{\Sigma|x - \bar{x}|}{n}$$

Σ means 'summation' and calculates the sum across all deviations $|x - \bar{x}|$ as we consider each data point x. By adding these and dividing by the number of data values, n, we get the mean of these deviations.

| ... | is the **absolute or modulus function**, which you'll explore in skills 570–572. It makes the value positive, so if the data value x is less than the mean $\bar{x}$, the difference/deviation will always be treated as a positive difference.

The age of 4 musicians in a band are as follows:

$$18, 24, 25, 25$$

Calculate the mean absolute deviation.

$$\bar{x} = \frac{18 + 24 + 25 + 25}{4}$$
$$= 23$$

Calculate the mean of the data set.

Absolute deviations:

$$5, 1, 2, 2$$

Determine the difference of each value from the mean (treating as positive differences).

Mean absolute deviation:

$$\frac{5 + 1 + 2 + 2}{4} = \mathbf{2.5}$$

Calculate the mean of these absolute deviations.

In other words, each band member's age is, on average, 2.5 years from the mean.

Standard Deviation

The absolute/modulus function, whilst seemingly simple, causes problems* when its formula is used to derive further results.

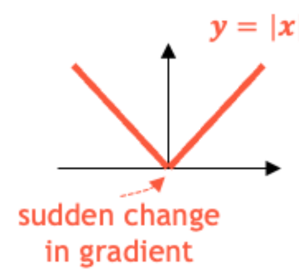
Absolute deviation

What's another way we could ensure the deviation is positive?

$$|x - \bar{x}| \longrightarrow (x - \bar{x})^2$$

Squaring makes negative values positive.

* **Advanced**: The modulus function is not **<u>differentiable</u>** because of the **discontinuity in gradient**. Differentiation is required when choosing parameters that minimises the discrepancy between the y value in the data and the predicted y value for example, e.g. when deriving the formula for the gradient and y-intercept of a straight line of best fit (the 'least squares regression line').
<u>Regression is an optimisation problem so typically requires differentiation</u>.
The same occurs in any similar optimisation that involves the deviation.
Using $(x - \bar{x})^2$ works better because the 'squared' can be easily differentiated whereas $|x - \bar{x}|$ cannot.



Standard Deviation

We average these across all values in the data set.

We find each squared deviation.

$$\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$$

Because the deviations were squared, we square root at the end to counter this.
This also ensures the unit (e.g. cm) of the standard deviation will be the same as the original data.

The **standard** deviation σ of a data set can be calculated using:

$$\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$$

Because of this sequence of steps of squaring, finding the mean, then rooting, the standard deviation is known as a **Root Mean Square (RMS)** measure.
The same principle can be used to determine how well a model, e.g. $y = ab^x$, fits some data.

σ is lowercase 'sigma' in the Greek alphabet, whereas Σ is uppercase sigma.

Standard Deviation vs Mean Absolute Deviation

The age of 4 musicians in a band are as follows:

$$18, 24, 25, 25$$

Calculate:

a the mean absolute deviation
b the standard deviation

The **standard** deviation σ of a data set can be calculated using:

$$\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$$

a **From earlier:**

$$\bar{x} = \frac{18+24+25+25}{4} = 23$$

Absolute deviations:

$$5, 1, 2, 2$$

Mean absolute deviation:

$$\frac{5+1+2+2}{4} = \mathbf{2.5\ years}$$

b $\bar{x} = 23$

Deviations $x - \bar{x}$:
$-5, 1, 2, 2$

$$\sigma = \sqrt{\frac{(-5)^2+1^2+2^2+2^2}{4}} = \mathbf{2.915\ years}$$

You may have **expected the two values to be the same**, but the standard deviation σ is always slightly greater than (or equal to) the mean absolute deviation*. The standard deviation can be interpreted as **approximately** the average distance of the values from the mean.

* **Advanced**: For normally distributed data, the ratio between the two is $\sqrt{2/\pi}$ which corresponds to the MAD being 20% less than σ (for the example above, the discrepancy is 14%).

σ on a Calculator

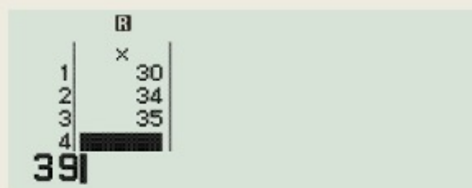
Time (secs)	30	34	35	39

These are instructions for the **Casio fx-570/991CW**

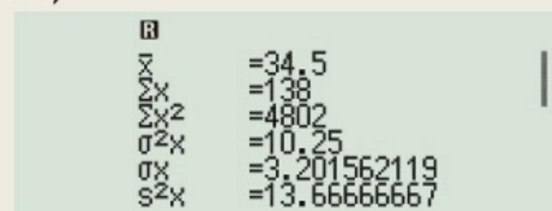
1 Use the arrows and OK to select Statistics.

2 Choose 1-Variable.

3 Enter your values, pressing = after each value, and an additional = after the last value to end your list.



4 Choose '1 Variable Results'. Read off σx (σ) and $\sigma^2 x$ (the variance σ^2)



Simplifying the Formula for σ

This formula for variance is slightly tedious to calculate, as we must subtract $\bar{x}$ from every value first.

$$\sigma^2 = \frac{\Sigma(x-\bar{x})^2}{n} = \frac{\Sigma x^2}{n} - \bar{x}^2$$

This can be simplified as follows. You can memorise it using '*msmsm*': "**mean of the squares minus the square of the mean**", i.e. we square the values first, find the mean of them, and subtract the square of the original mean.

Proof:
Note that $\bar{x}$ is constant for a fixed variable, and that in general, $\Sigma k f(x) = k \Sigma f(x)$ for a constant k, i.e. we can factor out constants out of a summation.

$$\begin{aligned}
\sigma^2 &= \frac{\Sigma(x-\bar{x})^2}{n} \\
&= \frac{\Sigma(x^2 - 2x\bar{x} + \bar{x}^2)}{n} \\
&= \frac{\Sigma x^2}{n} - \frac{\Sigma(2x\bar{x})}{n} + \frac{\Sigma \bar{x}^2}{n} \\
&= \frac{\Sigma x^2}{n} - 2\bar{x}\left(\frac{\Sigma x}{n}\right) + \frac{\bar{x}^2}{n}\Sigma 1 \\
&= \frac{\Sigma x^2}{n} - 2\bar{x}^2 + \frac{\bar{x}^2}{n} \cdot n \\
&= \frac{\Sigma x^2}{n} - 2\bar{x}^2 + \bar{x}^2 \\
&= \frac{\Sigma x^2}{n} - \bar{x}^2
\end{aligned}$$

Worked Example	Your Turn
Calculate the variance and standard deviation: a) $2, 3, 4, 5, 6$ b) $2, 4, 6, 8, 10$	Calculate the variance and standard deviation: a) $2, 3, 4, 5, 7$ b) $4, 6, 8, 10, 12$ c) $1, 2, 3, 4, 5$

Dr Frost 532c

Standard Deviation from Grouped Data

Data is...	Mean	Variance
Ungrouped	$\bar{x} = \frac{\Sigma x}{n}$	$\sigma^2 = \frac{\Sigma x^2}{n} - \bar{x}^2$
Grouped	$\bar{x} = \frac{\Sigma f x}{\Sigma f}$	$\sigma^2 = \frac{\Sigma f x^2}{\Sigma f} - \bar{x}^2$

If the data is grouped, then we previously saw that to calculate the mean, fx gives the total when x is written out f times, so accounts for the duplication of values. $\Sigma f = n$, because the total frequency is the number of data values.

The same applies for the variance formula. Each value is duplicated f times, so each squared value will also be duplicated f times, contributing fx^2 to the total.

Formulae

For a set of n values $x_1,\ x_2,\ \ldots x_i,\ \ldots x_n$

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

$$S_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$$\text{Standard deviation} = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2} \text{ or } \sqrt{\frac{S_{xx}}{n}}$$

$\bar{x} = \dfrac{\sum fx}{\sum f}$ where x is the midpoint of each class.

$$\text{Standard deviation} = \sqrt{\frac{\sum fx^2}{\sum f} - \bar{x}^2}$$

σ on a Calculator for Grouped Data

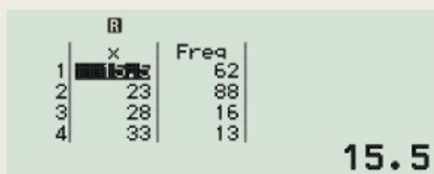
Time (minutes) t	11 – 20	21 – 25	26 – 30	31 – 35	36 – 45	46 – 60
Midpoints	15.5	23	28	33	40.5	53
Number of students f	62	88	16	13	11	10

These are instructions for the **Casio fx-570/991CW**

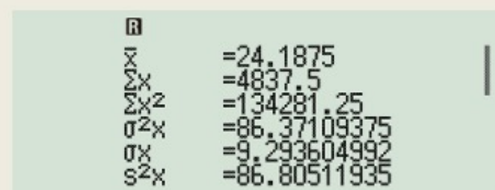
1 Use the arrows and OK to select Statistics.

2 Choose 1-Variable. Then press the "…" button, select 'Frequency', then 'On', then press the ← button twice.

3 Enter your values, pressing = after each value, using the arrow keys to get back to the top of the table for entering frequencies. Double press = to complete your table.



4 Choose '1 Variable Results'. Read off σx (σ) and $\sigma^2 x$ (the variance σ^2)



Important note: Σx^2 here means $\Sigma f x^2$, consistent with the value given in the question on the previous slide. The frequencies are implicitly 'baked in'.

Worked Example	Your Turn
Calculate the variance and standard deviation:	Calculate the variance and standard deviation:

Score	Frequency
0	3
1	2
2	1
3	1
4	4

Score	Frequency
0	6
1	4
2	2
3	2
4	8

Worked Example	Your Turn
Work out how many people had a score more than one standard deviation below the mean.	Work out how many people had a score more than one standard deviation below the mean.

Worked Example

Score	Frequency
0	3
1	2
2	1
3	1
4	4
5	9
6	5

Your Turn

Score	Frequency
0	6
1	4
2	2
3	2
4	8
5	18
6	10

Worked Example	Your Turn
Estimate the variance and standard deviation:	Estimate the variance and standard deviation:

Score, x	Frequency
$0 \leq x < 1$	8
$1 \leq x < 2$	2
$2 \leq x < 4$	1
$4 \leq x < 9.5$	1
$9.5 \leq x < 10$	4

Score, x	Frequency
$0 < x \leq 1$	6
$1 < x \leq 3$	4
$3 < x \leq 6$	2
$6 < x \leq 6.5$	2
$6.5 < x \leq 10$	8

Worked Example	Your Turn
Times, x, have been rounded to the nearest minute. Estimate the variance and standard deviation:	Times, x, have been rounded to the nearest minute. Estimate the variance and standard deviation:

Time, x	**Frequency**
$0-2$	5
$3-5$	2
$6-10$	3

Time, x	**Frequency**
$0-3$	7
$4-8$	11
$9-10$	2

Worked Example	Your Turn
The scores, x, were recorded for 20 people. The summary data is: $S_{xx} = 235$ Calculate the standard deviation.	The scores, x, were recorded for 40 people. The summary data is: $S_{xx} = 532$ Calculate the standard deviation.

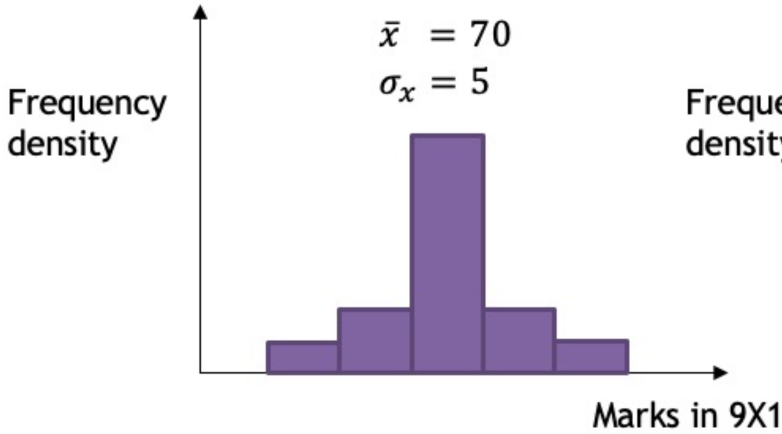
Worked Example	Your Turn
The scores, x, were recorded for 20 people. The summary data is: $\sum x = 34$, $\sum x^2 = 567$ Calculate the mean and standard deviation.	The scores, x, were recorded for 40 people. The summary data is: $\sum x = 76$, $\sum x^2 = 543$ Calculate the mean and standard deviation.

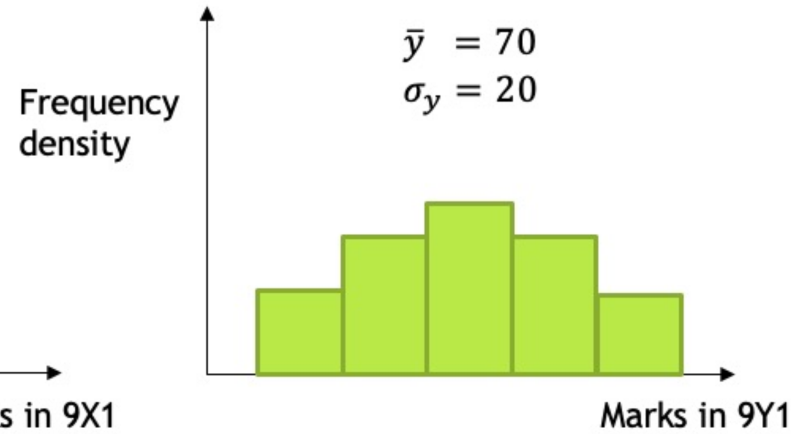
Dr Frost 534f

Worked Example	Your Turn
The scores, x, were recorded for 20 people. The summary data is: $\sum x = 34$, $\sum x^2 = 567$ The highest score was 8.5 The lowest score was 0.2 Estimate the number of scores which were greater than one standard deviation above the mean.	The scores, x, were recorded for 40 people. The summary data is: $\sum x = 76$, $\sum x^2 = 543$ The highest score was 5.8 The lowest score was 0.3 Estimate the number of scores which were greater than one standard deviation above the mean.

Worked Example	Your Turn
In classes 8A and 8B, consisting of 5 and 10 students respectively, the means of their maths scores were 60 and 70, and the standard deviations 6 and 7. Determine the standard deviation across all 15 students.	Throughout the year Alice took nine written tests and five practical tests in Music. The mean mark of her nine written tests was 69. The mean mark of her five practical tests was 83. The standard deviation of her nine written tests was 15.4. The standard deviation of her five practical tests was 7.8. Calculate the standard deviation of all 14 tests.

Comparing Sets using Standard Deviation





Compare the marks of the two classes.

The classes had the same mean mark, but 9X1 had a lower standard deviation, suggesting the marks were **more consistent**.

When σ is lower, it means there's less variation, which means **more consistency**.

Previously when comparing two data sets, we might compare:

1. A measure of **central tendency** (usually the mean)
2. A measure of **spread**

For the latter, you may have previously used the range or interquartile range. **Now you can use the standard deviation**!

2.5 Coding

Notes

General Case

The general case:

If a set of data values X is related to a set of values Y so that $Y = a X + b$, then:

$$\text{mean of } Y = a \times \text{mean of } X + b$$

$$\text{standard deviation of } Y = a \times \text{standard deviation of } X$$

$$\text{variance of } Y = a^2 \times \text{variance of } X$$

Coding Variables

x	179	160	165	172

Let $y = x + 10$

y	189	170	175	182

When we write a **statistical variable within an expression**, it **replaces** each data value according to that expression. This is known as **coding**.

Coding Variables

x	179	160	165	172

Let $y = x + 10$

y	189	170	175	182

a Calculate $\bar{x}$
b Calculate $\bar{y}$
c What do you notice?

a $\bar{x} = 169$
b $\bar{y} = 179$

c In increasing each value by 10, we also increased the mean by 10, i.e.

$$\bar{y} = \bar{x} + 10$$

Does this work on all Transformations?

x	1	2	3	4

Let $y = x^2$

y	1	4	9	16

a Calculate $\bar{x}$
b Calculate $\bar{y}$
c What do you notice?

a $\bar{x} = 2.5$
b $\bar{y} = 7.5$

c $2.5^2 \neq 7.5$. Squaring the values does **not** square the mean, i.e. although $y = x^2$, $\bar{y} \neq (\bar{x})^2$

Coding Averages

If a variable is coded using $y = ax + b$, then $\bar{y} = a\bar{x} + b$

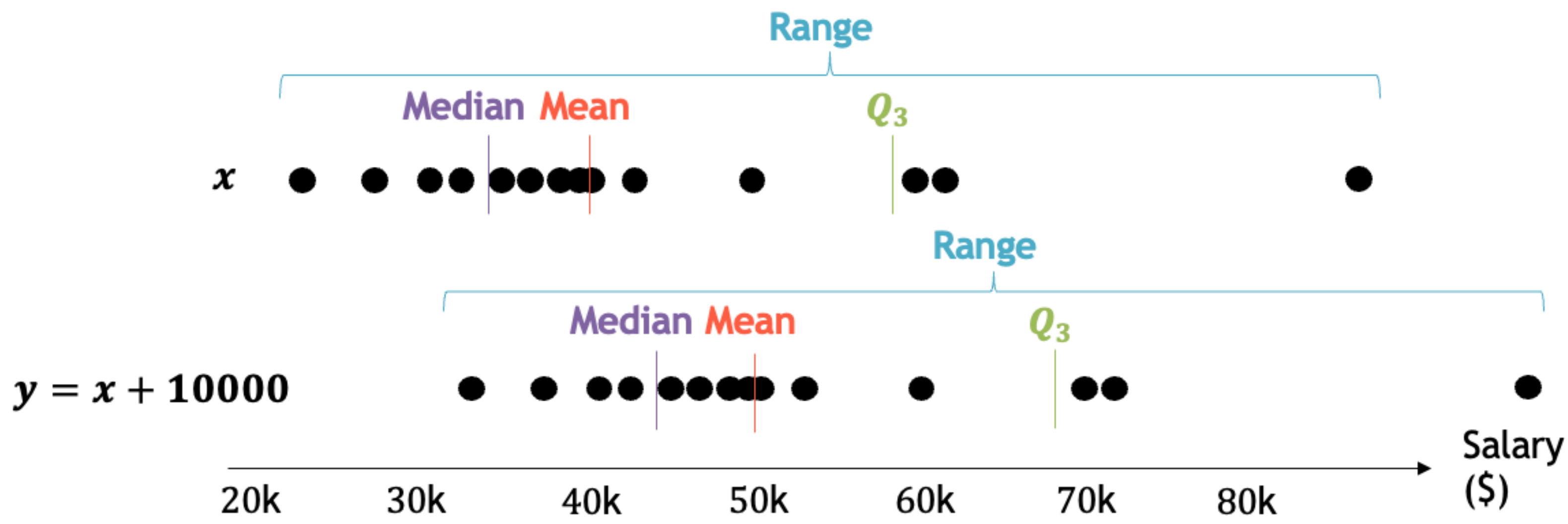
This means that any **linear** transformation to the variable (i.e. adding, subtracting, multiplying, dividing) will affect the mean in the same way.

Example:

The weights w of 10 cats in kg is recorded. $\bar{w} = 8$
If $y = 3w + 2$, when calculate $\bar{y}$.

$$\bar{y} = 3\bar{w} + 2$$
$$= (3\times8) + 2$$
$$= \mathbf{26} \text{ kg}$$

Effect of Coding on Other Statistics



It's visually easy to see that as the values shift up, so does the mean in the same way.

What about the median?
The median will be affected in the same way as the mean.

What about quartiles?
Any quantile (e.g. median, quartiles) will be affected in the same way.

What about the range?
As 10 000 is being added to both the minimum and maximum value, the difference between these, i.e. the range, will be unaffected.

If the values were doubled however ($y = 2x$), then this would double the range.

Summary of Effects of Coding

	Effect of + and −	Effect of × and ÷
Averages and measures of position (mean, mode, median, quartiles, minimum)	**As per operation**	**As per operation**
Measures of spread (range, interquartile range, standard deviation)	**No effect**	**As per operation**

Old mean $\bar{x}$	Old σ_x	Coding	New mean $\bar{y}$	New σ_y
10	2	$y = 4x$	**40**	**8**
15	4	$y = x + 12$	**27**	**4**
22	5	$y = 2(x + 1)$	**46**	**10**
10	**1**	$y = 3x - 5$	25	3

$$\bar{x} = \frac{\bar{y} + 5}{3}$$

σ is not affected by + or −, so

$$\sigma_x = \frac{\sigma_y}{3}$$

To work backwards, we can rearrange for x

$$x = \frac{y + 5}{3}$$

Worked Example	Your Turn
The daily mean pressure x in Jacksonville is recorded over a selected period. The mean is 1005.4 and the standard deviation is 66.9 The data is coded using $y = \frac{6x-23}{5}$ Find the mean and standard deviation of the coded data. Give your answers correct to 3 significant figures.	The windspeed x in Leuchars is recorded over a selected period. The mean is 14.3 and the standard deviation is 1.55 The data is coded using $y = \frac{10x-3}{8}$ Find the mean and standard deviation of the coded data. Give your answers correct to 3 significant figures.

Dr Frost 545c

Worked Example	Your Turn
The daily mean temperature x in Hurn is recorded over a selected period. The data is coded using $y = \frac{4x-24}{2}$ The mean of the coded data is 37.6 and the standard deviation of the coded data is 4.20 Find the mean and standard deviation of x Give your answers correct to 3 significant figures.	The daily mean pressure x in Beijing is recorded over a selected period. The data is coded using $y = \frac{7x-34}{9}$ The mean of the coded data is 794 and the standard deviation of the coded data is 24.0 Find the mean and standard deviation of x Give your answers correct to 3 significant figures.

Dr Frost 545f

Worked Example	Your Turn
Scores, x: $2090, 2080, 2070, 2060, 2050$ a) Use the coding $y = \frac{x-2000}{10}$ to code this data b) Calculate the mean and standard deviation of the coded data c) Use your answer to b) to calculate the mean and standard deviation of the original data	Scores, x: $1010, 1020, 1030, 1040, 1050$ a) Use the coding $y = \frac{x-1000}{10}$ to code this data b) Calculate the mean and standard deviation of the coded data c) Use your answer to b) to calculate the mean and standard deviation of the original data

Worked Example	Your Turn
Scores, x, of 20 people were recorded. The data was coded using $y = 5x - 10$ and the following summations were obtained: $\sum y = 23$, $\sum y^2 = 147.6$ Calculate the standard deviation of the actual scores.	Scores, x, of 40 people were recorded. The data was coded using $y = 10x - 5$ and the following summations were obtained: $\sum y = 32$, $\sum y^2 = 764.1$ Calculate the standard deviation of the actual scores.

Worked Example	Your Turn
A teacher standardises scores, x, of his class by adding 10 to each score and then reducing the score by 8%. The following summary statistics are calculated for the standardised scores, y: $n = 30, \bar{y} = 23.4, S_{yy} = 5.6$ Calculate the mean and standard deviation of the original scores	A teacher standardises scores, x, of his class by adding 8 to each score and then reducing the score by 10%. The following summary statistics are calculated for the standardised scores, y: $n = 25, \bar{y} = 43.2, S_{yy} = 6.5$ Calculate the mean and standard deviation of the original scores

Most Common Exam Errors

- ❑ Thinking $\Sigma f x^2$ means $(\Sigma f x)^2$. It means the sum of each value squared!
- ❑ When asked to calculate the mean followed by standard deviation, using a rounded version of the mean in calculating the standard deviation, and hence introducing rounding errors.
- ❑ Forgetting to square root the variance to get the standard deviation.

ALL these mistakes can be easily spotted if you check your value against "σx" in STATS mode.

Worked Example

A lake contains three different types of carp.

There are an estimated 450 mirror carp, 300 leather carp and 850 common carp.

Tim wishes to investigate the health of the fish in the lake.

He decides to take a sample of 160 fish.

As part of the health check, Tim weighed the fish.

His results are given in the table below.

Weight (w kg)	Frequency (f)	Midpoint (m kg)
$2 \leqslant w < 3.5$	8	2.75
$3.5 \leqslant w < 4$	32	3.75
$4 \leqslant w < 4.5$	64	4.25
$4.5 \leqslant w < 5$	40	4.75
$5 \leqslant w < 6$	16	5.5

(You may use $\sum fm = 692$ and $\sum fm^2 = 3053$)

(c) Calculate an estimate for the standard deviation of the weight of the carp.

(2)

Tim realised that he had transposed the figures for 2 of the weights of the fish.

He had recorded in the table 2.3 instead of 3.2 and 4.6 instead of 6.4

(d) Without calculating a new estimate for the standard deviation, state what effect

(i) using the correct figure of 3.2 instead of 2.3

(ii) using the correct figure of 6.4 instead of 4.6

would have on your estimated standard deviation.

Give a reason for each of your answers.

(2)

Your Turn

A company sells assorted chocolate discs of different sizes.

In each batch, there are an estimated 90 dark chocolates, 270 milk chocolates and 180 white chocolates.

Lorraine wishes to carry out quality control checks.

She decides to take a sample of 90 chocolates.

As part of the quality control check, Lorraine measured the diameters of the chocolate discs.

Her results are given in the table below.

Diameter (d mm)	**Frequency (f)**	**Midpoint (m mm)**
$10 \le d < 15$	8	12.5
$15 \le d < 20$	15	17.5
$20 \le d < 22$	36	21
$22 \le d < 27$	24	24.5
$27 \le d < 30$	7	28.5

(You may use $\sum fm = 1\,906$ and $\sum fm^2 = 41\,811.5$)

(c) Calculate an estimate for the standard deviation of the diameter of the chocolates.

(2)

Lorraine realised that she had incorrectly recorded 2 of the diameters of the chocolate discs.

She had recorded 10.1 in the table instead of 21.9 and 23.5 instead of 25.3.

(d) Without calculating a new estimate for the standard deviation, state what effect

(i) using the correct figure of 21.9 instead of 10.1

(ii) using the correct figure of 25.3 instead of 23.5

would have on your estimated standard deviation.

Give a reason for each of your answers.

(2)

Worked Example

Stav is studying the large data set for September 2015

He codes the variable Daily Mean Pressure, x, using the formula $y = x - 1010$

The data for all 30 days from Hurn are summarised by

$$\sum y = 214 \qquad \sum y^2 = 5912$$

(a) State the units of the variable x

(1)

(b) Find the mean Daily Mean Pressure for these 30 days.

(2)

(c) Find the standard deviation of Daily Mean Pressure for these 30 days.

(3)

Your Turn

Stav is studying the large data set for June 1987

He codes the variable Daily Mean Temperature, x, using the formula $y = x - 7$

The data for all 30 days from Hurn are summarised by

$$\sum y = 199.5 \qquad \sum y^2 = 5824.648$$

(a) State the units of the variable x

(1)

(b) Find the mean Daily Mean Temperature for these 30 days.

(2)

(c) Find the standard deviation of Daily Mean Temperature for these 30 days.

(4)

(d) What type of variable is daily mean pressure?

(1)

Worked Example

Dian uses the large data set to investigate the Daily Total Rainfall, r mm, for Camborne.

(a) Write down how a value of $0 < r \leqslant 0.05$ is recorded in the large data set.

(1)

Dian uses the data for the 31 days of August 2015 for Camborne and calculates the following statistics

$$n = 31 \qquad \sum r = 174.9 \qquad \sum r^2 = 3523.283$$

(b) Use these statistics to calculate

(i) the mean of the Daily Total Rainfall in Camborne for August 2015,

(ii) the standard deviation of the Daily Total Rainfall in Camborne for August 2015.

(3)

Your Turn

Dian uses the large data set to investigate the Daily Total Rainfall, r mm, for Leuchars.

Dian uses the data for the 31 days of July 1987 for Leuchars and calculates the following statistics

$$n = 31 \qquad \sum r = 61.4 \qquad \sum r^2 = 463.88$$

(b) Use these statistics to calculate

(i) the mean of the Daily Total Rainfall in Leuchars for July 1987,

(ii) the standard deviation of the Daily Total Rainfall in Leuchars for July 1987.

(3)

Worked Example

Ben is studying the Daily Total Rainfall, x mm, in Leeming for 1987

He used all the data from the large data set and summarised the information in the following table.

x	0	0.1–0.5	0.6–1.0	1.1–1.9	2.0–4.0	4.1–6.9	7.0–12.0	12.1–20.9	21.0–32.0	tr
Frequency	55	18	18	21	17	9	9	6	2	29

(a) Explain how the data will need to be cleaned before Ben can start to calculate statistics such as the mean and standard deviation.

(2)

Using all 184 of these values, Ben estimates $\sum x = 390$ and $\sum x^2 = 4336$

(b) Calculate estimates for

(i) the mean Daily Total Rainfall,

(ii) the standard deviation of the Daily Total Rainfall.

(3)

Your Turn

Mike is studying the Daily Total Rainfall, x mm, in Leuchars for 1987

He used all the data from the large data set and summarised the information in the following table.

x	0	0.1–0.5	0.6–1.0	1.1–1.9	2.0–4.0	4.1–6.9	7.0–12.0	12.1–20.9	21.0–32.0	tr
Frequency	42	35	12	9	21	19	9	5	2	30

(a) Explain how the data will need to be cleaned before Mike can start to calculate statistics such as the mean and standard deviation.

(2)

Using all 184 of these values, Mike estimates $\sum x = 423$ and $\sum x^2 = 4373$

(b) Calculate estimates for

(i) the mean Daily Total Rainfall,

(ii) the standard deviation of the Daily Total Rainfall.

(3)

Worked Example

Ming is studying the large data set for Perth in 2015

He intended to use all the data available to find summary statistics for the Daily Mean Air Temperature, x °C.
Unfortunately, Ming selected an incorrect variable on the spreadsheet.
This incorrect variable gave a mean of 5.3 and a standard deviation of 12.4

The correct values for the Daily Mean Air Temperature are summarised as

$$n = 184 \qquad \sum x = 2801.2 \qquad \sum x^2 = 44\,695.4$$

(b) Calculate the mean and standard deviation for these data.

(3)

Your Turn

Colin is studying the large data set for Jacksonville in 2015

He intended to use all the data available to find summary statistics for the Daily Mean Air Temperature, x °C.
Unfortunately, Colin selected an incorrect variable on the spreadsheet.
This incorrect variable gave a mean of 1016 and a standard deviation of 4.0

The correct values for the Daily Mean Air Temperature are summarised as

$n = 184 \qquad \sum x = 4569.4 \qquad \sum x^2 = 114811.5$

(b) Calculate the mean and standard deviation for these data.

(3)

Standard deviation

Standard deviation = √(Variance)

Interquartile range = $\mathrm{IQR} = \mathrm{Q}_3 - \mathrm{Q}_1$

For a set of n values $x_1, x_2, \ldots x_i, \ldots x_n$

$$\mathrm{S}_{xx} = \Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}$$

Standard deviation = $\sqrt{\dfrac{\mathrm{S}_{xx}}{n}}$ or $\sqrt{\dfrac{\sum x^2}{n} - \bar{x}^2}$

Summary

1 The **mode** or **modal class** is the value or class that occurs most often.

2 The **median** is the middle value when the data values are put in order.

3 The **mean** can be calculated using the formula $\bar{x} = \frac{\Sigma x}{n}$.

4 For data given in a frequency table, the mean can be calculated using the formula $\bar{x} = \frac{\Sigma xf}{\Sigma f}$.

5 To find the **lower quartile** for discrete data, divide n by 4. If this is a whole number, the lower quartile is halfway between this data point and the one above. If it is not a whole number, round *up* and pick this data point.

6 To find the **upper quartile** for discrete data, find $\frac{3}{4}$ of n. If this is a whole number, the upper quartile is halfway between this data point and the one above. If it is not a whole number, round *up* and pick this data point.

7 The **range** is the difference between the largest and smallest values in the data set.

8 The **interquartile range** (IQR) is the difference between the upper quartile and the lower quartile, $Q_3 - Q_1$.

9 The **interpercentile range** is the difference between the values for two given percentiles.

10 **Variance** $= \frac{\Sigma(x-\bar{x})^2}{n} = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2 = \frac{S_{xx}}{n}$ where $S_{xx} = \Sigma(x-\bar{x})^2 = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$

11 The **standard deviation** is the square root of the variance:

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

12 You can use these versions of the formulae for variance and standard deviation for grouped data that is presented in a frequency table:

$$\sigma^2 = \frac{\Sigma f(x-\bar{x})^2}{\Sigma f} = \frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2 \qquad \sigma = \sqrt{\frac{\Sigma f(x-\bar{x})^2}{\Sigma f}} = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2}$$

where f is the frequency for each group and Σf is the total frequency.

13 If data is coded using the formula $y = \frac{x-a}{b}$

- the mean of the coded data is given by $\bar{y} = \frac{\bar{x}-a}{b}$
- the standard deviation of the coded data is given by $\sigma_y = \frac{\sigma_x}{b}$ where σ_x is the standard deviation of the original data.

4
- Variables or data associated with numerical observations are called **quantitative variables** or **quantitative data**.
- Variables or data associated with non-numerical observations are called **qualitative variables** or **qualitative data**.

5
- A variable that can take any value in a given range is a **continuous variable**.
- A variable that can take only specific values in a given range is a **discrete variable**.

6
- When data is presented in a grouped frequency table, the specific data values are not shown. The groups are more commonly known as **classes**.
- Class boundaries tell you the maximum and minimum values that belong in each class.
- The midpoint is the average of the class boundaries.
- The class width is the difference between the upper and lower class boundaries.