Chapter 2: Measures of Location & Spread

1:: Mean, Median, Mode

"Calculate the mean of this grouped frequency table."

2:: Quartiles, Percentiles, Deciles

"Use linear interpolation to estimate the interquartile range."

3:: Variance & Standard Deviation

"Calculate the standard deviation of the maths marks."

4:: Coding

"The marks x where coded using $y = 2x + 10$. Given that the standard deviation of y is 5, determine the standard deviation of x."

Variables in algebra vs stats

$$x$$

Similarities

- Just like in algebra, variables in stats **represent the value of some quantity**, e.g. shoe size, height, colour.
- As we saw in the previous chapter, variables can be discrete or continuous.
- **Can be part of further calculations**, e.g. if x represents height, then $2x$ represents twice people's height. In stats this is known as '**coding**', which we'll cover later.

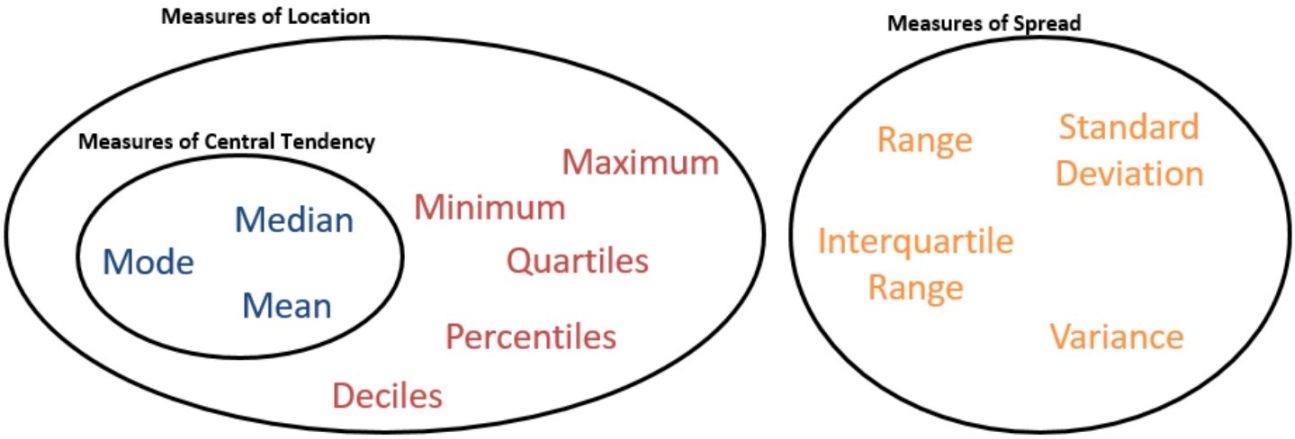
Differences

- Unlike algebra, a variable in stats represents the value of **multiple objects** (i.e. it's a bit like a set). e.g. the heights of **all** people in a room.
- Because of this, we can do **operations** on it as if it was a **collection of values**:
  - If x represents people's heights,

    $$\Sigma x$$

    gives the sum of everyone's heights. In algebra this would be meaningless: if $x = 4$, then Σx makes no sense!
  - $\bar{x}$ is the mean of x. Notice x is a collection of values whereas $\bar{x}$ is a single value.
- To each value of the variable, **we could attach an associated probability**. This is known as a **random variable** (Chapter 6).

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**.

Mean of ungrouped data

Diameter of coin x (cm)	2.2	2.5	2.6	2.65	2.9

You all know how to find the mean of a list of values.
But lets consider the notation, and see how theoretically we could calculate each of the individual components on a calculator.

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

The 'overbar' in stats specifically means 'the sample mean of', but don't worry about the 'sample' bit for now.

Diameter of coin x (cm)	2.2	2.5	2.6	2.65	2.9

On a Classwiz:

- Select 1-Variable.
- Enter each value above, pressing = after each entry.
- Press AC to start a statistical calculation.
- Press the OPTN button. "1-Variable Calc" will calculate all common statistics (including all on the left). Alternatively you can construct a statistical expression yourself – in the OPTN menu press Down. "Variable" for example contains $\bar{x}$. This will insert it into your calculation; press = when done.

Frequency Tables (ungrouped data)

Number of Children (x)	Frequency (f)
0	4
1	3
2	9
3	2

A frequency table allows us to avoid writing out repeated values.
Recall that each value x must be multiplied by the frequency (f), to ensure each value is repeated appropriately when adding up all the values.

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

Exam Tip: In the exam you get a method mark for the division and an accuracy mark for the final answer. Write:

$$\bar{x} = \frac{46.75}{40} = 1.16875$$

You're not required to show working like "$0 \times 4 + \cdots$"

Using your calculator in STATS mode

Number of Children (x)	Frequency (f)
0	4
1	3
2	9
3	2

To add a frequency column for data input, press SHIFT → SETUP, press Down, then choose Statistics. Turn frequency 'On'.
You can then input data in the usual way.

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

Grouped Data

Height h of bear (in metres)	Frequency
$0 \leq h < 0.5$	4
$0.5 \leq h < 1.2$	20
$1.2 \leq h < 1.5$	5
$1.5 \leq h < 2.5$	11

We don't know the exact values anymore. So what do we assume each value is?
The midpoint of each interval. We use the variable x to indicate the midpoint.
We can then calculate mean in exactly the same way as before.

Estimate of Mean: $\bar{x} = \frac{\sum fx}{\sum f}$

Why is our mean just an estimate? Because we don't know the exact heights within each group. Grouping data loses information.

Warning: ClassWizs will calculate the lower and upper quartiles (Q_1, Q_3) along with the median. However, this is not applicable to grouped data: When you input your midpoints in the data input, your calculator doesn't know these are midpoints – it just assumes for example that the first 4 bears **did** have a height of 0.25m. We need to take into account the class widths to estimate the median and quartiles (which we'll see later), and your calculator cannot do this.

Use your calculator's STATS mode to determine the mean (or estimate of the mean).
Ensure that you show the division in your working.

1

Num children (c)	Frequency (f)
0	2
1	6
2	1
3	1

$\bar{c} =$

2

IQ (q)	Frequency (f)
$80 < q \leq 90$	7
$90 \leq q < 100$	5
$100 \leq q < 120$	3
$120 \leq q < 200$	1

$\bar{q} =$

3

Time t	Frequency (f)
$9.5 < t \leq 10$	32
$10 \leq t < 12$	27
$12 \leq t < 15$	47
$15 \leq t < 16$	11

$\bar{t} =$

Ex 2A/B

Combined Mean

This subtopic doesn't appear in your textbook but has cropped up in exams.

The mean maths score of 20 pupils in class A is 62.
The mean maths score of 30 pupils in class B is 75.

a) What is the overall mean of all the pupils' marks.
b) The teacher realises they mismarked one student's paper; he should have received 100 instead of 95. Explain the effect on the mean and median.

Archie the Archer competes in a competition with 50 rounds. He scored an average of 35 points in the first 10 rounds and an average of 25 in the remaining rounds. What was his average score per round?

Median – which item?

You need to be able to find the median of both listed data and of grouped data.

Listed data

Items	n	Position of median	Median
$1, 4, 7, 9, 10$	5		
$4, 9, 10, 15$	4		
$2, 4, 5, 7, 8, 9, 11$	7		
$1, 2, 3, 5, 6, 9, 9, 10, 11, 12$	10		

To find the position of the median for listed data, find $\frac{n}{2}$:
- If a decimal, round up.
- If whole, use halfway between this item and the one after.

Grouped data

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

Position to use for median:

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

DO NOT round $\frac{n}{2}$ or adjust it in any way. This is just like at GCSE where, if you had a cumulative frequency graph with 60 items, you'd look across the 30th. We'll cover linear interpolation in a sec...

Quick questions: What position do we use for the median?

Lengths: 3cm, 5cm, 6cm, …

$n = 11$

Median position:

Lengths: 4m, 8m, 12.4m, …

$n = 24$

Median position:

Age	Freq
$10 \leq a < 20$	12
$20 \leq a < 30$	5

Median position:

Score	Freq
$150 \leq s < 200$	3
$200 \leq s < 400$	7

Median position:

Ages: 5, 7, 7, 8, 9, 10, …

$n = 60$

Median position:

Score	Freq
$150 \leq s < 200$	15
$200 \leq s < 400$	6

Median position:

Weights: 1.2kg, 3.3kg, …

$n = 35$

Median position:

Volume (ml)	Freq
$0 \leq v < 100$	5
$100 \leq v < 200$	6
$200 \leq v < 300$	2

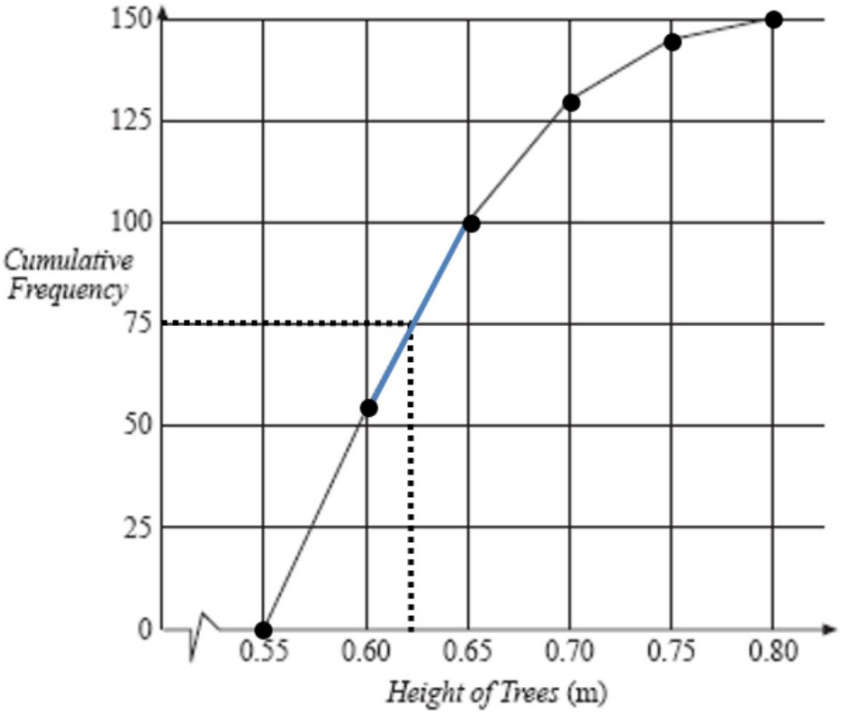
Median position:

Weights: 4.4kg, 7.6kg, 7.7kg…

$n = 18$

Median position:

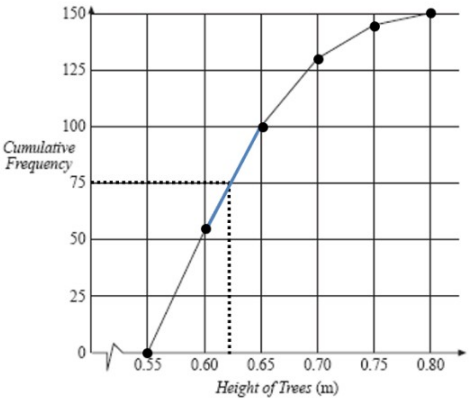
Linear Interpolation



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

At GCSE we could find the median by drawing a suitable line on a cumulative frequency graph. How could we read off this value ***exactly*** using a suitable calculation?

1. *find the fraction of the way along the line segment using the frequencies*
2. *then go this same fraction along the class interval*



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

Tip: To quickly get frequency before and after, just look for the two cumulative frequencies that surround the item number.

Frequency up until this interval

Item number we're interested in.

Frequency by end of this interval

Height at start of interval.

Height at end of interval.

Tip: Put the units to avoid getting frequencies confused with values of the variable.

Find an estimate for the median

Weight of cat (kg), w	Freq
$1.5 \leq w < 3$	10
$3 \leq w < 4$	8
$4 \leq w < 6$	14

Find an estimate for the median

Time (s), t	Freq
$8 \leq t < 10$	4
$10 \leq t < 12$	3
$12 \leq t < 14$	13

True class limits

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

Weight of cat to nearest kg	Frequency
	7
	2
	9
	4

Class width =

Distance $\boldsymbol{d}$ travelled (in m)	...
$0 \leq \mathrm{d} < 150$	
$150 \leq \mathrm{d} < 200$	
$\mathbf{200 \leq d < 210}$	

Lower class boundary =
Upper class boundary =
Class width =

Time $\boldsymbol{t}$ taken (in seconds)	...
$0 - 3$	
$\mathbf{4 - 6}$	
$7 - 11$	

Lower class boundary =
Upper class boundary =
Class width =

Weight $\boldsymbol{w}$ in kg	...
$10 - 20$	
$21 - 30$	
$\mathbf{31 - 40}$	

Lower class boundary =
Upper class boundary =
Class width =

Speed $\boldsymbol{s}$ (in mph)	...
$10 \leq \mathrm{s} < 20$	
$20 \leq s < 29$	
$\mathbf{29 \leq s < 31}$	

Lower class boundary =
Upper class boundary =
Class width =

Linear Interpolation/True class limits

Summarised below are the distances, to the nearest mile, travelled to work by a random sample of 120 commuters.

Distance (to the nearest mile)	Number of commuters	
0 – 9	10	10
10 – 19	19	29
20 – 29	43	72
30 – 39	25	97
40 – 49	8	105
50 – 59	6	111
60 – 69	5	116
70 – 79	3	119
80 – 89	1	120

Find an estimate for the median

Here are some extra questions for you...

Ex 2C
Q1a, 3, 4a, 5a ONLY

1 The number of patients attending a hospital trauma clinic each day was recorded over several months, giving the data in the table below.

Number of patients	10 - 19	20 - 29	30 - 34	35 - 39	40 - 44	45 - 49	50 - 69
Frequency	2	18	24	30	27	14	5

Use linear interpolation to estimate the median of these data.

2 The ages of 300 houses in a village are recorded given the following table of results.

Age a (years)	Number of houses
$0 \leq a < 20$	36
$20 \leq a < 40$	92
$40 \leq a < 60$	74
$60 \leq a < 100$	39
$100 \leq a < 200$	14
$200 \leq a < 300$	27
$300 \leq a < 500$	18

Use linear interpolation to estimate the median.

3 A cyber-café recorded how long each user stayed during one day giving the following results.

Length of stay (minutes)	Number of houses
$0 \leq l < 30$	15
$30 \leq l < 60$	31
$60 \leq l < 90$	32
$90 \leq l < 120$	23
$120 \leq l < 240$	17
$240 \leq l < 360$	2

Use linear interpolation to estimate the median of these data.

4 The following table summarises the times, t minutes to the nearest minute, recorded for a group of students to complete an exam.

Time (minutes) t	11 – 20	21 – 25	26 – 30	31 – 35	36 – 45	46 – 60
Number of students f	62	88	16	13	11	10

[You may use $\sum ft^2 = 134281.25$]

(a) Estimate the mean ~~and standard deviation~~ of these data. **(5)**

(b) Use linear interpolation to estimate the value of the median. **(2)**

1) 37.2
2) 45.9
3) 73.125
4a) 24.1875
4b) 22.7

Quartiles – which item?

You need to be able to find the **quartiles** of both listed data and of grouped data. The rule is exactly the same as for the median.

Listed data

Items	n	Position of LQ and UQ	LQ & UQ
$1, 4, 7, 9, 10$	5		
$4, 9, 10, 15$	4		
$2, 4, 5, 7, 8, 9, 11$	7		
$1, 2, 3, 5, 6, 9, 9, 10, 11, 12$	10		

To find the position of the LQ/UQ for listed data, find $\frac{1}{4}n$ or $\frac{3}{4}n$ then as before:
- If a decimal, round up.
- If whole, use halfway between this item and the one after.

Grouped data

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

To find the LQ and UQ of grouped data, find $\frac{1}{4}n$ and $\frac{3}{4}n$, then use linear interpolation.

Position to use for LQ:

Position to use for UQ:

Again, **DO NOT** round this value.

Percentiles and Deciles

- Quartiles split the data up into *quarters*
- Percentiles split the data up into *hundredths*
- Deciles split the data up into *tenths*

The LQ, median and UQ give you 25%, 50% and 75% along the data respectively.
But we can have any percentage you like!

$n = 43$
Item to use for 57th percentile? → $\mathbf{43 \times 0.57 = 24.51}$

You will always find these for grouped data in an exam, **so never round this position**.

If there were 150 items (n=150), which item/position would you use to find the …

- 33rd percentile?
- 3rd quartile?
- 7th decile?

Notation:

Lower Quartile:

Median:

Upper Quartile:

64th Percentile:

3rd Decile:

25th Percentile:

Ex 2C
Remaining questions

Measures of Spread

The range, interquartile range and interpercentile range are examples of **measures of spread.**

Measures of spread tell us how similar or variable the data it is.

Small spread? All the data is quite close.
Big spread? It's really varied!

Interquartile Range = Upper Quartile − Lower Quartile

Why might we favour the interquartile range over the range?

We can control this further by having for example the "*10^{th} to 90^{th} interpercentile range*", which would be $P_{90} - P_{10}$. This would typically be symmetrical about the median, so that we could interpret this as "*the range of the data with the most extreme 10% of values at either end excluded*".

Age of relic (years)	Frequency
0-1000	24
1001-1500	29
1501-1700	12
1701-2000	35

Find the IQR

Shark length (cm)	Frequency
$40 \leq x < 100$	17
$100 \leq x < 300$	5
$300 \leq x < 600$	8
$600 \leq x < 1000$	11

Find the 10^{th} to 90^{th} interpercentile range:

Ex 2D
Extra Qs below

Extra Questions

Q1) May 2013 Q4 (continued)

The following table summarises the times, t minutes to the nearest minute, recorded for a group of students to complete an exam.

Time (minutes) t	11 – 20	21 – 25	26 – 30	31 – 35	36 – 45	46 – 60
Number of students f	62	88	16	13	11	10

(*c*) Show that the estimated value of the lower quartile is 18.6 to 3 significant figures. **(1)**

(*d*) Estimate the interquartile range of this distribution. **(2)**

Q2) June 2005 Q2

The following table summarises the distances, to the nearest km, that 134 examiners travelled to attend a meeting in London.

Distance (km)	Number of examiners
41–45	4
46–50	19
51–60	53
61–70	37
71–90	15
91–150	6

(*c*) Use interpolation to estimate the median Q_2, the lower quartile Q_1, and the upper quartile Q_3 of these data.

Q3)

The ages of 300 houses in a village are recorded given the following table of results.

Age a (years)	Number of houses
$0 \leq a < 20$	36
$20 \leq a < 40$	92
$40 \leq a < 60$	74
$60 \leq a < 100$	39
$100 \leq a < 200$	14
$200 \leq a < 300$	27
$300 \leq a < 500$	18

Use linear interpolation to estimate the lower quartile, upper quartile and hence the interquartile range.

Q4)

A cyber-café recorded how long each user stayed during one day giving the following results.

Length of stay (minutes)	Number of houses
$0 \leq l < 30$	15
$30 \leq l < 60$	31
$60 \leq l < 90$	32
$90 \leq l < 120$	23
$120 \leq l < 240$	17
$240 \leq l < 360$	2

Use linear interpolation to estimate:

a) The lower quartile.
b) The upper quartile.
c) The 90th percentile.

Q5)

Distance (to the nearest mile)	Number of commuters
0 – 9	10
10 – 19	19
20 – 29	43
30 – 39	25
40 – 49	8
50 – 59	6
60 – 69	5
70 – 79	3
80 – 89	1

Find the interquartile range for the distance travelled by commuters.

Answers

1)

(c) $Q_1 = 10.5 + \frac{(50/50.25)}{62} \times 10 [= 18.56]$

(d) $Q_3 = 25.5$ (Use of $n + 1$ gives 25.734..)
IQR = 6.9 (Use of $n + 1$ gives 7.1)

2) $Q_2 = 50.5 + \frac{(67-23)}{53} \times 10 = 58.8$

$Q_1 = 52.48$; $Q_3 = 67.12$

3)

$$Q_1 = 20 + \left(\frac{39}{92} \times 20\right) = 28.5$$

$$Q_3 = 60 + \left(\frac{23}{39} \times 40\right) = 83.6$$

$$IQR = 55.1$$

4)

$$Q_1 = 30 + \left(\frac{15}{31} \times 30\right) = 44.5$$

$$Q_3 = 90 + \left(\frac{12}{23} \times 30\right) = 105.7$$

$$P_{90} = 120 + \left(\frac{7}{17} \times 120\right)$$

$$= 169.4$$

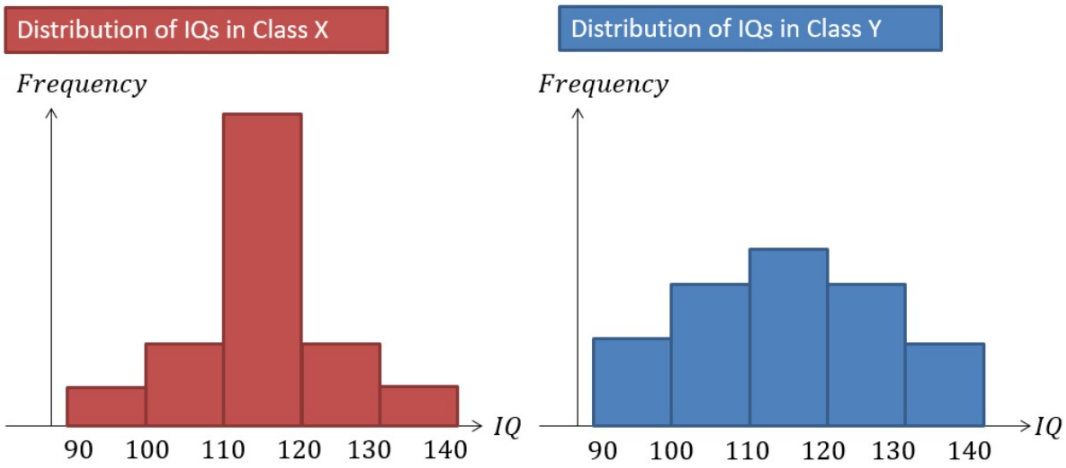
5)

$$Q_1 = 19.5 + \left(\frac{1}{43} \times 10\right) = 19.7$$

$$Q_3 = 29.5 + \left(\frac{18}{25} \times 10\right) = 36.7$$

$$IQR = 17.0$$

Variance



Here are the distribution of IQs in two classes of the same size.

The (estimated) mean IQ is the **same** for the two classes.

The (estimated) range is the **same** for the two classes.

The overall spread of values is **different** for the two classes.
It is **more** spread for the Class Y – the results vary more.
It is **less** spread for Class X - lots of students have similar IQs (between 110 and 120).

The interquartile range would convey this, but we want a method to measure spread that takes into account **all the values**, even the 'extremes' – this is called **variance**

Variance is a measure of spread that takes all values into account.
Variance, by definition, is the **average squared distance from the mean**.

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

Note: $S_{xx} = \sum(x-\bar{x})^2$

Hence we have $\sigma^2 = \frac{S_{xx}}{n}$

Distance from mean...

Squared distance from mean...

Average squared distance from mean...

Simpler formula for variance and standard deviation

NOTE:
$\Sigma x^2 \neq (\Sigma x)^2$

Variance

"The mean of the squares minus the square of the mean"
MSMSM

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

*** Proof: (certainly not in syllabus!)**
Note that $\bar{x}$ is constant for a fixed variable, and that in general, $\Sigma kf(x) = k\Sigma f(x)$ for a constant k, i.e. we can factor out constants out of a summation.

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

Standard Deviation

$$\sigma = \sqrt{Variance}$$

The standard deviation **can 'roughly' be thought of as the <u>average distance from the mean</u>.**

Find the variance and standard deviation

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

3, 11

2cm 3cm 3cm 5cm 7cm

Extending to frequency/grouped frequency tables

Tip: It's better to try and memorise MSMSM than the formula itself – you'll understand what's going on better.

Exam Note: In an exam, you will pretty much certainly be asked to find the standard deviation for grouped data, and not listed data.

$$Variance = \frac{\Sigma f x^2}{\Sigma f} - \bar{x}^2$$

NOTE:
$\Sigma f x^2 \neq (\Sigma f x)^2$

An agriculturalist is studying the yields, y kg, from tomato plants. The data from a random sample of 70 tomato plants are summarised below.

Yield (y kg)	**Frequency** (f)	**Yield midpoint** (x kg)
$0 \leq y < 5$	16	2.5
$5 \leq y < 10$	24	7.5
$10 \leq y < 15$	14	12.5
$15 \leq y < 25$	12	20
$25 \leq y < 35$	4	30

(You may use $\sum fx = 755$ and $\sum fx^2 = 12\,037.5$)

(c) Estimate the mean and the standard deviation of the yields of the tomato plants. **(4)**

We can use our STATS mode to work out the various summations needed (and "1-Variable Calc" will contain this amongst its list)

4. The following table summarises the times, t minutes to the nearest minute, recorded for a group of students to complete an exam.

Time (minutes) t	11 – 20	21 – 25	26 – 30	31 – 35	36 – 45	46 – 60
Number of students f	62	88	16	13	11	10

[You may use $\sum ft^2 = 134281.25$]

(*a*) Estimate the mean and standard deviation of these data. **(5)**

1. Sara is investigating the variation in daily maximum gust, t kn, for Camborne in June and July 1987.

She used the large data set to select a sample of size 20 from the June and July data for 1987. Sara selected the first value using a random number from 1 to 4 and then selected every third value after that.

(a) State the sampling technique Sara used.

(1)

(b) From your knowledge of the large data set, explain why this process may not generate a sample of size 20.

(1)

The data Sara collected are summarised as follows

$$n = 20 \qquad \sum t = 374 \qquad \sum t^2 = 7600$$

(c) Calculate the standard deviation.

(2)

Coding

Estimate the mean height of the members of our class.

What would happen to the **mean** height and the **variance** if we all suddenly grew by 50cm… perhaps by standing on our chairs?

Now suppose you're all **3 times** your original height.

What do you think happens to the **mean** of your heights?

What do you think happens to the **standard deviation** of your heights?

What do you think happens to the **variance** of your heights?

Rules of coding

What is coding?

Coding data means applying the same rules to the data so it is easier to process. For example, rather than using large numbers that are in their millions, it might be easier to divide them all by 1,000,000 so it is easier to perform calculations.

Suppose our original variable (e.g. heights in cm) was x. Then y would represent the heights with 10cm added on to each value.

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

$y = 3x$ $\quad \bar{y} =$ $\quad \sigma_y =$

$y = 2x - 5$ $\quad \bar{y} =$ $\quad \sigma_y =$

$y = ax + b$ $\quad \bar{y} =$ $\quad \sigma_y =$

Coding	Old mean $\bar{x}$	New mean $\bar{y}$	Old σ_x	New σ_y
$y = x - 20$	36		4	
$y = 2x$		72		16
$y = 3x - 20$	35		4	
$y = \frac{x}{2}$		20		$\frac{3}{2}$
$y = \frac{x + 10}{3}$	11		27	
$y = \frac{x - 100}{5}$		40		5

Summary:

Measures of **location** are affected by **all** parts of coding

Measures of **spread** are **only** affected by multiplicative parts of coding

Cost x of diamond ring (£)

£1010	£1020	£1030	£1040	£1050

We 'code' our variable using the following:

$$y = \frac{x - 1000}{10}$$

New values y:

£1	£2	£3	£4	£5

The following table summarises the times, t minutes to the nearest minute, recorded for a group of students to complete an exam.

Time (minutes) t	11 – 20	21 – 25	26 – 30	31 – 35	36 – 45	46 – 60
Number of students f	62	88	16	13	11	10

[You may use $\Sigma ft^2 = 134281.25$]

(*a*) Estimate the mean and standard deviation of these data. **(5)**
(*b*) Use linear interpolation to estimate the value of the median. **(2)**
(*c*) Show that the estimated value of the lower quartile is 18.6 to 3 significant figures. **(1)**
(*d*) Estimate the interquartile range of this distribution. **(2)**
(*e*) Give a reason why the mean and standard deviation are not the most appropriate summary statistics to use with these data. **(1)**

The person timing the exam made an error and each student actually took 5 minutes less than the times recorded above. The table below summarises the actual times.

Time (minutes) t	6 – 15	16 – 20	21 – 25	26 – 30	31 – 40	41 – 55
Number of students f	62	88	16	13	11	10

(*f*) Without further calculations, explain the effect this would have on each of the estimates found in parts (*a*), (*b*), (*c*) and (*d*). **(3)**

Suppose we've worked all these out already.

From the large data set, data on the maximum gust, g knots, is recorded in Leuchars during May and June 2015.

The data was coded using $h = \frac{g-5}{10}$ and the following statistics found:

$$S_{hh} = 43.58 \qquad \bar{h} = 2 \qquad n = 61$$

Calculate the mean and standard deviation of the maximum gust in knots.

Coded time (x minutes)	Frequency (f)	Coded time midpoint (y minutes)
$0 \leq x < 5$	3	2.5
$5 \leq x < 10$	15	7.5
$10 \leq x < 15$	2	12.5
$15 \leq x < 25$	9	20
$25 \leq x < 35$	1	30

(You may use $\sum fy = 355$ and $\sum fy^2 = 5675$)

1. A company manager is investigating the time taken, t minutes, to complete an aptitude test. The human resources manager produced the table below of coded times, x minutes, for a random sample of 30 applicants.

(a) Use linear interpolation to estimate the median of the coded times.

(2)

(b) Estimate the standard deviation of the coded times.

(2)

The company manager is told by the human resources manager that he subtracted 15 from each of the times and then divided by 2, to calculate the coded times.

(c) Calculate an estimate for the median and the standard deviation of t.

(3)

The following year, the company has 25 positions available. The company manager decides not to offer a position to any applicant who takes 35 minutes or more to complete the aptitude test.

The company has 60 applicants.

(d) Comment on whether or not the company manager's decision will result in the company being able to fill the 25 positions available from these 60 applicants. Give a reason for your answer.

(2)

Large Data Set Exam Questions

4. Joshua is investigating the daily total rainfall in Hurn for May to October 2015

Using the information from the large data set, Joshua wishes to calculate the mean of the daily total rainfall in Hurn for May to October 2015

(a) Using your knowledge of the large data set, explain why Joshua needs to clean the data before calculating the mean.

(1)

Using the information from the large data set, he produces the grouped frequency table below.

Daily total rainfall (r mm)	Frequency	Midpoint (x mm)
$0 \leqslant r < 0.5$	121	0.25
$0.5 \leqslant r < 1.0$	10	0.75
$1.0 \leqslant r < 5.0$	24	3.0
$5.0 \leqslant r < 10.0$	12	7.5
$10.0 \leqslant r < 30.0$	17	20.0

You may use $\sum fx = 539.75$ and $\sum fx^2 = 7704.1875$

(b) Use linear interpolation to calculate an estimate for the upper quartile of the daily total rainfall.

(2)

(c) Calculate an estimate for the standard deviation of the daily total rainfall in Hurn for May to October 2015

(2)

(d) (i) State the assumption involved with using class midpoints to calculate an estimate of a mean from a grouped frequency table.

(ii) Using your knowledge of the large data set, explain why this assumption does not hold in this case.

(iii) State, giving a reason, whether you would expect the actual mean daily total rainfall in Hurn for May to October 2015 to be larger than, smaller than or the same as an estimate based on the grouped frequency table.

(3)

1. The daily mean air temperatures from the large data set, x °C, for the month of June 2015 in Jacksonville are summarised in the table below.

Daily mean air temperature (°C)	$22 \leqslant x < 24$	$24 \leqslant x < 25$	$25 \leqslant x < 26$	$26 \leqslant x < 27$	$27 \leqslant x < 28$	$28 \leqslant x < 32$
Frequency	2	5	7	4	6	6

(a) Use your calculator to estimate the mean and the standard deviation of the daily mean air temperatures from the large data set, for the month of June 2015 in Jacksonville.

Give each of your answers to 3 significant figures.

(2)

The mean and standard deviation for the daily mean air temperatures from the large data set for Perth in June 2015 are 14.8 °C and 2.37 °C respectively.

The minimum daily mean air temperature in Perth in June 2015 was 8.8 °C and the maximum daily mean air temperature was 18.5 °C

(b) Using limits for outliers of

$$\text{mean} - 3 \times \text{standard deviation}$$
$$\text{mean} + 3 \times \text{standard deviation}$$

show that there are no outliers in the data for Perth in June 2015.

(2)

(c) (i) Assuming each location is typical of the hemisphere it is in, suggest what these means and standard deviations imply about the relative daily mean air temperature in June 2015 in each hemisphere.

Give reasons for your answers.

(2)

(ii) Comment on the validity of the assumption in (i)

(1)

1. The number of hours of sunshine each day, y, for the month of July at Heathrow are summarised in the table below.

Hours	$0 \leqslant y < 5$	$5 \leqslant y < 8$	$8 \leqslant y < 11$	$11 \leqslant y < 12$	$12 \leqslant y < 14$
Frequency	12	6	8	3	2

~~A histogram was drawn to represent these data. The $8 \leqslant y < 11$ group was represented by a bar of width 1.5 cm and height 8 cm.~~

~~(a) Find the width and the height of the $0 \leqslant y < 5$ group.~~

(3)

(b) Use your calculator to estimate the mean and the standard deviation of the number of hours of sunshine each day, for the month of July at Heathrow.
Give your answers to 3 significant figures.

(3)

The mean and standard deviation for the number of hours of daily sunshine for the same month in Hurn are 5.98 hours and 4.12 hours respectably.
Thomas believes that the further south you are the more consistent should be the number of hours of daily sunshine.

(c) State, giving a reason, whether or not the calculations in part (b) support Thomas' belief.

(2)

(d) Estimate the number of days in July at Heathrow where the number of hours of sunshine is more than 1 standard deviation above the mean.

(2)