Schoolworkout Maths

S1 Revision Notes: Numerical and Descriptive Statistics

Section 1: Mean and standard deviation

Recap:
The **mean** ($\bar{x}$) and the **variance** of a set of data are found using the formulae:

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

$$\text{variance} = \frac{\sum x_i^2}{n} - \bar{x}^2 \qquad \text{or variance} = \frac{\sum (x_i - \bar{x})^2}{n}.$$

where n is the number of values.

The **standard deviation** is the square root of the variance. The standard deviation measures how far the data tend to be from the mean value and so informs us of the spread of the data.

- **Example:**

a) An A level Chemistry class sat a test. The marks were

45%, 62%, 75%, 39%, 52%, 84%, 71%, 65%, 64%

Find the mean and the standard deviation of the marks.

b) Last year, the A level Chemistry group sat the same test. Their mean mark was 59% and their standard deviation was 9.34%. Compare the marks scored by this year's group with last years.

- **Solution:**

a) $\bar{x} = \frac{\sum x_i}{n} = \frac{45 + 62 + ... + 64}{9} = \frac{557}{9} = 61.\dot{8}\,\%$

$$\sum x_i^2 = 45^2 + 62^2 + ... + 64^2 = 36137$$

So $\quad \text{variance} = \frac{\sum x_i^2}{n} - \bar{x}^2 = \frac{36137}{9} - 61.\dot{8}^2 = 184.99$

So $\quad \text{s.d.} = \sqrt{184.99} = 13.6\,\%$ (3sf)

b) This year's Chemistry class obtained higher marks on average but their marks tended to be more spread out than last year's class.

N.B. *When you are asked to make a comparison, you should try to interpret both the mean and the standard deviation in the context of the question.*

You can work out the mean and the standard deviation using your calculator!:- Make sure that you can use your calculator buttons to find a mean and a standard deviation.

Finding the mean and the standard deviation from a table

Recap:
The formulae for the mean and variance can be modified to deal with data summarised in a frequency table:

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

$$\text{variance} = \frac{\sum f_i x_i^2}{\sum f_i} - \left(\frac{\sum f_i x_i}{\sum f_i}\right)^2.$$

where $\sum f_i = n$ is the total frequency.

- **Example:**

A factory employs 30 people. The table shows how many days the employees had off sick in the last month.

Number of days sick	Number of employees
0	17
1	6
2	4
3	2
4	1

Find the mean and the standard deviation.

- **Solution:**

We extend the table by adding columns for fx and fx^2:

Number of days sick, *x*	Number of employees, *f*	*fx*	fx^2
0	17	$0 \times 17 = 0$	$17 \times 0^2 = 0$
1	6	$1 \times 6 = 6$	$6 \times 1^2 = 6$
2	4	$2 \times 4 = 8$	$4 \times 2^2 = 16$
3	2	$3 \times 2 = 6$	$2 \times 32 = 18$
4	1	$4 \times 1 = 4$	$1 \times 42 = 16$
TOTALS	$\sum f = 30$	$\sum fx = 24$	$\sum fx^2 = 56$

We get:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{24}{30} = 0.8$$

and...

$$\text{variance} = \frac{\sum f_i x_i^2}{\sum f_i} - \left(\frac{\sum f_i x_i}{\sum f_i}\right)^2 = \frac{56}{30} - \left(\frac{24}{30}\right)^2 = 1.226667$$

i.e. s.d.= 1.11 (3SF)

- **Example**:

The table shows the lengths (in metres) of 250 vehicles on a cross-channel ferry.

Vehicle length (m)	Frequency
3.0-4.0	90
4.0-4.5	80
4.5-5.0	40
5.0-5.5	24
5.5-7.5	16

Estimate the mean and the variance of the lengths.

Note: We can only estimate the mean and the variance here since we do not know the exact lengths of the vehicles.

- **Solution**:

We base our calculations upon the mid-point of each interval...

Vehicle length (m)	Mid-point, x	Frequency, f	fx	fx^2
3.0-4.0	3.5	90	315	1102.5
4.0-4.5	4.25	80	340	1445
4.5-5.0	4.75	40	190	902.5
5.0-5.5	5.25	24	126	661.5
5.5-7.5	6.5	16	104	676
TOTALS		**250**	**1075**	**4787.5**

An estimate of the mean length is: $\frac{1075}{250} = 4.3$ m

The variance is... $\frac{4787.5}{250} - 4.3^2 = 0.66$

Notes:

- Make sure that the mean seems a sensible size. Does it lie roughly in the middle of the data?

- If you use your calculator to find the mean and the standard deviation, make sure that you give enough significant figures in your answer. It is sensible to write down your full calculator display and then round it to 3 significant figures. You will not get any marks unless you show at least 3 significant figures in non-exact answers.

Section 2: Median and the Quartiles

To find the **median** of a set of n numbers:

- List the numbers carefully in order of size, smallest first;
- The median is the middle number, i.e. the number in position $\frac{n+1}{2}$.

There are several ways to find the **quartiles**. Two ways are illustrated in the example below. The different ways don't always give the same answer, so make sure your method is clear.

- **Example:**

A transport analyst recorded the speeds of 15 cars passing along a stretch of road during peak hours. Their speeds in mph were:

21, 19, 26, 14, 28, 26, 25, 34, 22, 22, 27, 34, 18, 23, 29.

She repeated her survey during off-peak hours and found that the speeds of a sample of 18 cars were:

34, 29, 30, 31, 32, 37, 42, 25, 48, 28, 34, 32, 35, 37, 31, 30, 28, 36

Find the median and the interquartile range for each set of data.
Compare the speeds of traffic during peak and off-peak hours.

- **Solution**:

Peak hours:

We begin by ordering the data:

14, 18, 19, 21, 22, 22, 23, **25**, 26, 26, 27, 28, 29, 34, 34

The median is in position (15 + 1)/2 = 8, i.e. the median is 25 mph.

The quartiles can be found in either of these ways:

The lower quartile is the median of the lower half of the data:

14, 18, 19, 21, 22, 22, 23

So the lower quartile (L.Q) is 21.

The upper quartile is the median of the upper half of the data:

26, 26, 27, 28, 29, 34, 34

So the upper quartile (U.Q.) is 28.

The IQR = U.Q. – L.Q. = 28 – 21 = 7

OR

The lower quartile is the value in position $\frac{n+1}{4}$ and the upper quartile is in position $\frac{3(n+1)}{4}$.

So, here the L.Q. is the $\frac{15+1}{4} = 4^{\text{th}}$ number

and the U.Q. is the $\frac{3(15+1)}{4} = 12^{\text{th}}$ number.

Therefore L.Q. = 21 and U.Q. = 28

The IQR is also 7

Off-Peak hours:

We begin by ordering the data:

25, 28, 28, 29, 30, 30, 31, 31, **32,** **32**, 34, 34, 35, 36, 37, 37, 42, 48

The median is in position (18 + 1)/2 = 9.5, i.e. the median is *half-way* between the 9th and the 10th numbers i.e. 32 mph.

The quartiles can be found in either of the two ways described earlier:

The lower quartile is the median of the lower half of the data:

25, 28, 28, 29, 30, 30, 31, 31, 32

So the lower quartile (L.Q) is 30.

The upper quartile is the median of the upper half of the data:

32, 34, 34, 35, 36, 37, 37, 42, 48

So the upper quartile (U.Q.) is 36.

The IQR = U.Q. – L.Q. = 36 – 30 = 6

OR

The L.Q. is the $\frac{18+1}{4} = 4.75^{th}$ number i.e. 5th number and the U.Q. is the $\frac{3(18+1)}{4} = 14.25^{th}$ number, i.e. 14th number.

Therefore U.Q. = 36 and L.Q. = 30

The IQR is 6

Comparison:

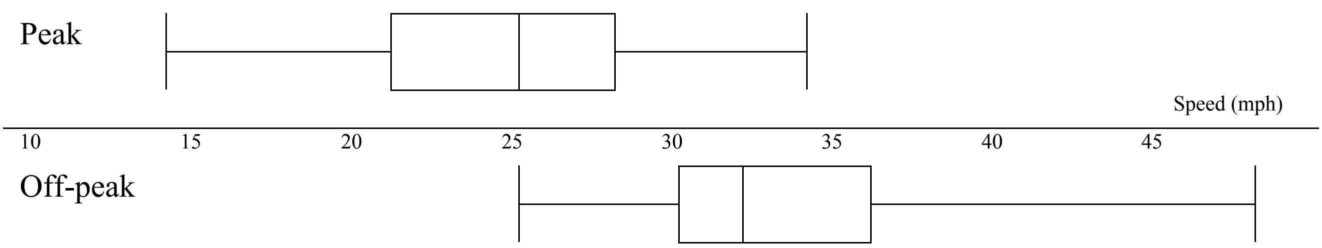
The medians show that traffic flows more quickly on average during off-peak times.
The interquartile ranges are fairly similar, but the speeds for the off-peak times were slightly more varied that those of peak times.

Box-and-whisker plots

If the above data appeared in an examination, you might well be asked to draw a box-and-whisker plot to compare the off-peak and peak speeds.
To draw a box plot you need 5 quantities:

The lowest and highest values;
The lower and upper quartiles;
The median.

You should draw a scale that is *common* to both box plots – the scale should be *labelled*. The box plots here would look something like:



Notes:

- Make sure you draw a box and whisker plot on *graph* paper.
- The mean and standard deviation are most useful when the data are roughly symmetrical and contains no outliers (or anomalous results).
- The median and the interquartile range are typically used if the data are skewed or if there are outliers in the data.

Section 3:Histograms

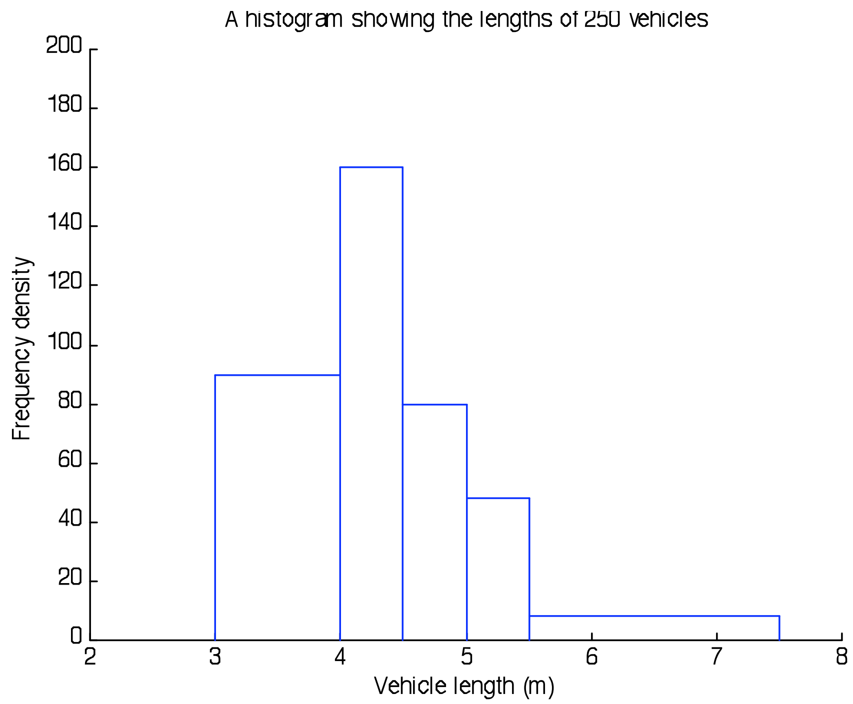
When you are asked to draw a histogram in a S1 examination, it is **essential** that you work out and plot the **FREQUENCY DENSITIES** on the y-axis, where

Frequency density = Frequency ÷ class width.

- **Example:**

The lengths (in metres) of 250 vehicles aboard a cross-channel ferry are summarised in the following table:

Vehicle length (m)	*Class width*	*Frequency*	*Frequency density = Frequency ÷ class width*
3.0-4.0	1	90	90
4.0-4.5	0.5	80	160
4.5-5.0	0.5	40	80
5.0-5.5	0.5	24	48
5.5-7.5	2	16	8



N.B. It is important for you to label each axis and to give your graph a title.

Sometimes you have to think carefully about the width of each interval. You have to do this if the upper endpoint of one interval does not appear to match the lower endpoint of the next interval.

- **Example (rounded data)**

A class of 30 Year 5 children took part in a running race. The teacher recorded how long each child took to complete the race to the nearest second. Their times are shown in the table.

Time interval (seconds)	Frequency
40 – 49	5
50 – 54	8
55 – 59	6
60 – 69	7
70 –	4

The intervals in this table do not appear to meet because the data has been recorded to the nearest second. The first interval actually includes all times from 39.5 seconds up to (but not including) 49.5 seconds; the second interval all times from 49.5 up to 54.5 etc.

Also, the last interval does not have an upper end point. In such circumstances it is conventional to assume that the last interval has a width that is twice that of the previous interval.

We therefore have this new table:

Time interval (seconds)	Class width	Frequency	Frequency density
39.5 – 49.5	10	5	0.5
49.5 – 54.5	5	8	1.6
54.5 – 59.5	5	6	1.2
59.5 – 69.5	10	7	0.7
69.5 – 89.5	20	4	0.2

A histogram can then be drawn.

- **Example (ages)**

The ages (in completed years) of 120 people voting at a polling station were:

Age (years)	Frequency
18 – 29	10
30 – 49	13
50 – 59	21
60 – 69	42
70 – 89	34

The intervals in this table also do not appear to meet because age is recorded in completed years. An adult whose age lies in the interval 18 – 29 can be anything from 18 up to (but not including) 30 years old. We therefore have this new table:

Age (years)	Class width	Frequency	Frequency density
18 up to 30	12	10	0.83
30 up to 50	20	13	0.65
50 up to 60	10	21	2.1
60 up to 70	10	42	4.2
70 up to 90	20	34	1.7

A histogram can then be drawn.

Section 4: Stem-and-leaf diagrams

Stem-and-leaf diagrams are used to represent data in its original form. Each piece of data is split into two parts.
The numbers in the leaves should be written in numerical order and you should include a key on your diagram.

- **Example**

The data below show the highest November temperature recorded in several European countries last year:

10, 7, 14, 17, 14, 9, 21, 22, 14, 19, 11, 20, 13, 18, 22.

Draw a stem-and-leaf diagram to illustrate the data.

- **Solution**

Each number has two parts to it – a tens digit and a units digit. We will write the tens digits on the stem and the unit digits as the leaves:

0 | 7 means 7˚C

0	7	9							
1	0	4	7	4	4	9	1	3	8
2	1	2	0	2					

Write stems in order:

0 | 7 means 7˚C

0	7	9							
1	0	**1**	3	4	4	**4**	7	8	9
2	**0**	1	2	2					

Note: The rows of a stem-and-leaf diagram are sometimes split so that one row contains low digits (0, 1, 2, 3, 4) and the next row contains high digits (5, 6, 7, 8 , 9). The following example illustrates this.

- **Example**

The heights (in cm) of some school children are measured:

145, 138, 132, 143, 142, 142, 149, 153, 135, 140, 134, 148, 146, 142, 151

13 | 2 means 132cm

13	2				
13	8	5	4		
14	3	2	2	0	2
14	5	9	8	6	
15	3	1			
15					

Write leaves in order

13 | 2 means 132cm

13	2				
13	4	5	8		
14	0	2	2	2	3
14	5	6	8	9	
15	1	3			
15					

Notes:

- It is common to be asked to find medians and quartiles from stem-and-leaf diagrams.
- A stem-and-leaf diagram has the advantage that it contains the accuracy of the original data.
- A box-and-whisker plot has the advantage that it can be easily interpreted and comparisons can easily be made.

Section 5: Cumulative frequency diagrams

- **Example:**

A secretary weighed a sample of letters to be posted.

This interval is short for 50 – 60 g

Mass (g)	20 -	30 -	40 -	50 -	60 -	70 -	80 – 90
Number of students	2	4	12	7	8	17	3

Draw a cumulative frequency graph for the data
Use your graph to find the median weight of a letter and the interquartile range of the weights.

Solution:
We first need to work out the *cumulative frequencies* – these are a *running total* of the frequencies.

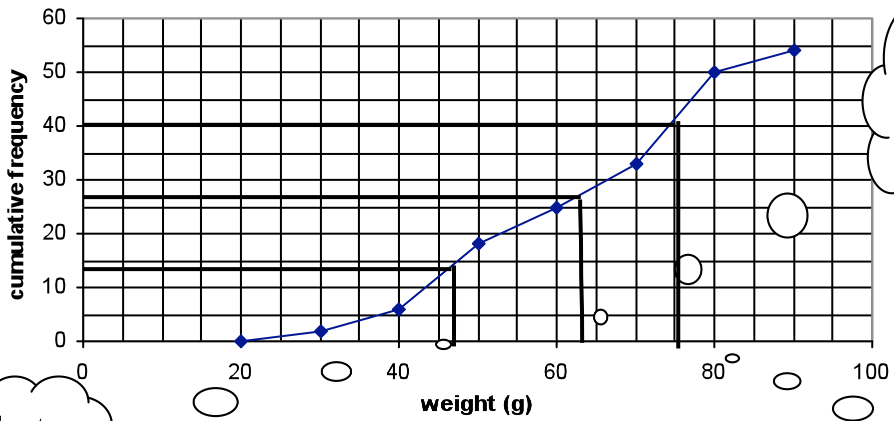
Mass (g)	Frequency	Cumulative frequency
20 – 30	2	2
30 – 40	4	6
40 – 50	12	18
50 – 60	7	25
60 – 70	8	33
70 – 80	17	50
80 – 90	4	54

There are 18 letters so far

There are 54 letters below 90g

We plot the cumulative frequency graph by plotting the cumulative frequencies on the vertical axis and the masses on the horizontal axis. *It is important that the cumulative frequencies are plotted above the* ***endpoint*** *of each interval.* So we plot the points (30, 2), (40, 6), (50, 18), …, (90, 54). As no letter weighed less than 20g, we can also plot the point (20, 0).

Cumulative frequency diagram to show weights of letters



The median is about 63g

L.Q. is about 47g

U.Q. is about 75g

The total number of letters examined was 54. The median will be approximately the $54 \div 2 = 27^{th}$ letter. We draw a line across from 27 on the vertical axis and then find the median on the horizontal axis. We see that the median is about 63g. (Note: we could find the median a bit more accurately by drawing a line across at (n + 1)/2 = 55/2, i.e. at 27.5.)

The lower quartile will be the $\frac{1}{4} \times 54 = 13.5th$ value. From the horizontal scale we find that the lower quartile is about 47g.

The upper quartile is the $\frac{3}{4} \times 54 = 40.5th$ value. This is about 75g.

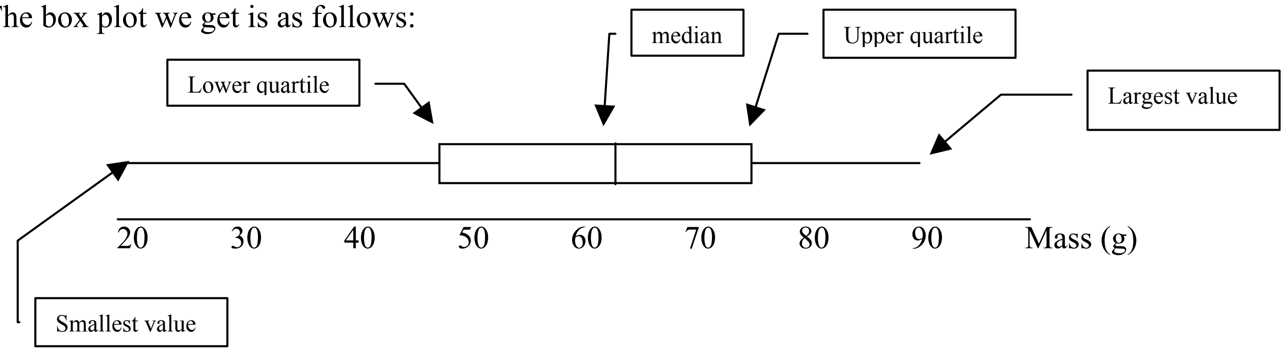
Therefore the interquartile range is U.Q – L.Q. = 75 – 47 = 28g.

Note: We can represent the data in the above example as a *box-and-whisker plot*. A box plot is based on 5 measurements:

- The lowest value
- The lower quartile
- The median
- The upper quartile
- The largest value.

In the example above we don't know the exact values of the lightest and heaviest letters. However we do know that no letter weighed less than 20g and no letter weighed more than 90g. So we take the lowest and largest values as 20g and 90g respectively.

The box plot we get is as follows:



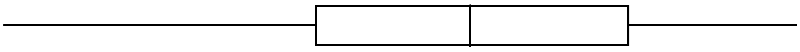
Notes:

- It is VITAL that you remember to plot the cumulative frequencies above the end-point of each interval.
- When finding the median and quartiles, you should draw in the vertical and horizontal lines on the graph. You must make sure that you read the median and quartiles off your scale as accurately as you can.
- Choose sensible scales on your axes. Draw axes that go up in 1's, 2's, 5's, 10's, 20's, 50's rather than axes numbered up in 3's, 6's, 7's, 15's etc.
- Don't draw your graph too small! Always use graph paper for drawing a cumulative frequency graph.
- At A level it is usual to join the points in a cumulative frequency diagram with straight lines (unless you are asked for a cumulative frequency curve).

Section 6: Skewness

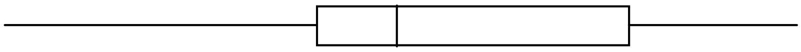
A distribution can sometimes be described in one of the following ways:

- Symmetrical



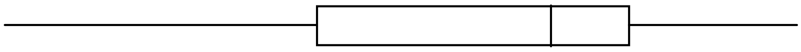
The median line lies in the middle of the box (i.e. UQ – median = median – LQ)

- Positively skewed



The median line lies closer to the L.Q. than the U.Q (i.e. UQ – median > median – LQ)

- Negatively skewed



The median line lies closer to the U.Q. than to the L.Q. (i.e. UQ – median < median – LQ).

We can also describe skewness from a histogram: