DATA HANDLING

Five-Number Summary

The five-number summary consists of the minimum and maximum values, the median, and the upper and lower quartiles.

- The minimum and the maximum are the smallest and greatest numbers that occur in the data set.
- The median is the middle most number when the data is arranged from smallest to greatest. (We say the data is ordered).
- The upper and lower quartiles are the median of the upper and lower halves of the data, respectively. Note that there are various methods used for determining the upper and lower quartiles of a set of data. The simplest of all the methods for finding the upper and lower quartiles is described below:
  - Arrange the data in order from smallest to greatest, and identify the median.
  - Identify the middle number of each half of the data on either side of the median.

Example

Example: Consider the set {12, 13, 18, 20, 22, 27, 29}.

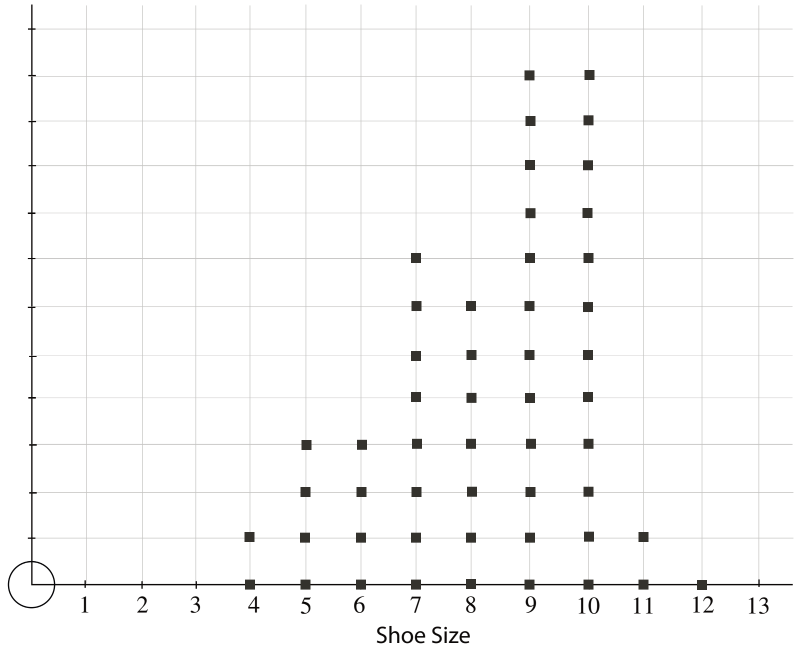
- The minimum is 12.
- The lower half is {12, 13, 18}, and the middle number of that half is 13. Therefore, the lower quartile is 13.
- The median is the middle term, 20.
- The upper half is {22, 27, 29}, and the middle number of that half is 27. Therefore, the upper quartile is 27.
- The maximum value is 29.

Example

Example: The shoe sizes of a group of 50 Grade 11 students were recorded and summarised in the table below:

Shoe size	4	5	6	7	8	9	10	11	12
Frequency	2	4	4	8	7	12	10	2	1

The above table can be illustrated graphically with a dot plot.



The 50 shoe sizes can be listed as follows:

{4;4;5;5;5;5;6;6;6;6;7;7;7;7;7;7;7;7;8;8;8;8;8;8;8;9;9;9;9;9;9;9;9;9;9;9;9;10;10;10;10;10;10;10;10;10;10;10;10;11;11;12}

The 5 number summary for the data is as follows:

- The **minimum** shoe size is 4
- The **maximum** shoe size is 12
- The **median** is the average of the middle two numbers, i.e. the 25th and the 26th numbers. The 25th number is 8 and the 26th number is 9. Therefore the median shoe size is $\frac{8+9}{2} = 8{,}5$

The median splits the data into two halves: the lower half and the upper half.

The lower graph is {4;4;5;5;5;5; 6;6;6;6;7;7;7;7;7;7;7;7;8;8;8;8;8;8;8}

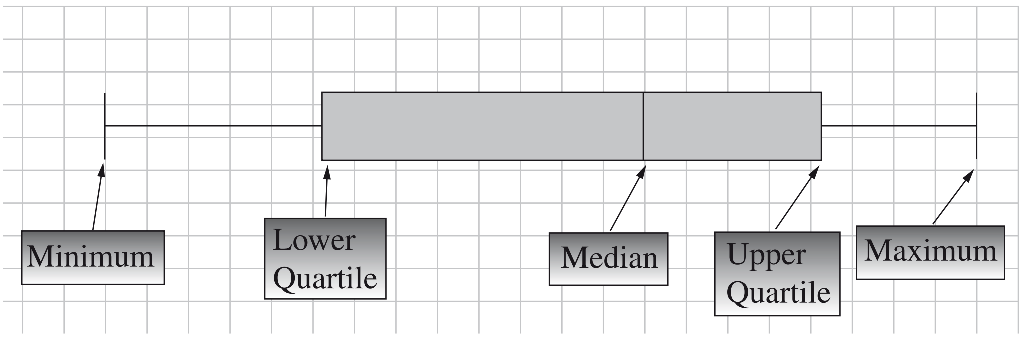
The upper half is {9;9;9;9;9;9;9;9;9;9;9;9;10;10;10;10;10;10;10;10;10;10;10;10;11;11 ;12}

- The **lower quartile** is the median of the lower half which is the 13th number in the list, i.e. 7
- The **upper quartile** is the median of the upper half which is the 38th number in the list, i.e. 10

Box and Whisker Plot

Box and Whisker Plots allow us to interpret the spread of the data more easily.

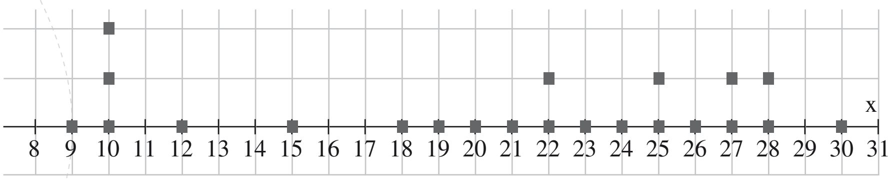
A typical box and whisker plot looks as follows:



The Box is the part from the lower quartile to the upper quartile and the whiskers are the lines on either end of the box. The end point of the whiskers give us the minimum and maximum values.

It is very important to note that the first 25% (first quarter) of results lies between the minimum and the lower quartile. The next 25% (second quarter) of results lies between the lower quartile and the median. The third quarter lies between the median and the upper quartile and the last quarter of data lies between the upper quartile and the maximum vale.

The diagram below shows a dot plot of a set of 22 values.



The individual results can be read from the dot plot. From smallest to largest:

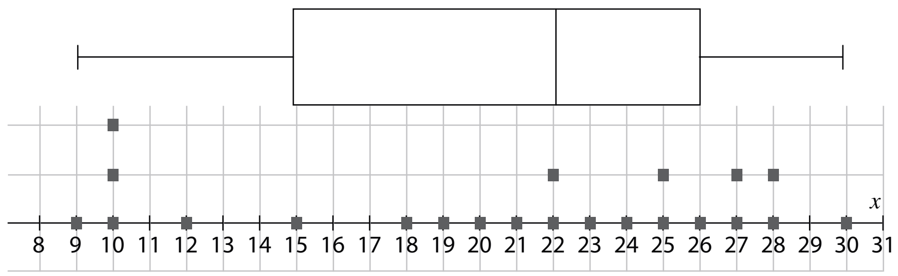
9, 10, 10, 10, 12, 15, 18, 19, 20, 21, 22, ¦ 22, 23, 24, 25, 25, 26, 27, 27, 28, 28, 30

half way

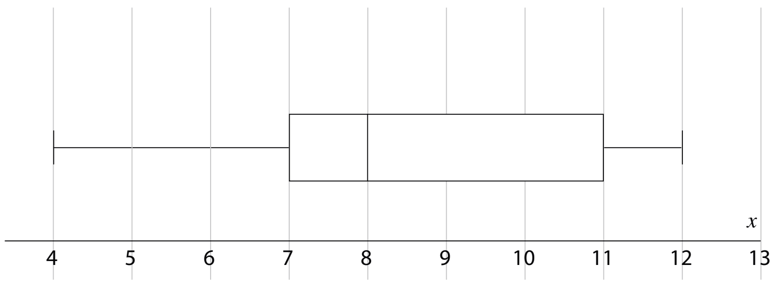
Consider the five number summary for this set of data:

- Minimum is 9.
- The lower quartile is the median of the lower half of the data, i.e. the median of 9, 10, 10, 10, 12, 15, 18, 19, 20, 21, 22. Therefore the lower quartile is the 6th number in the list which is 15
- The median is the middle most value which is the average of the two most middle numbers, i.e. $\frac{22+22}{2} = 22$
- The upper quartile is the median of the upper half of the data, i.e. the median of 22, 23, 24, 25, 25, 26, 27, 27, 28, 28, 30
  Therefore the upper quartile is the number 26.
- The maximum is 30

The Box and whisker plot would therefore look as follows:

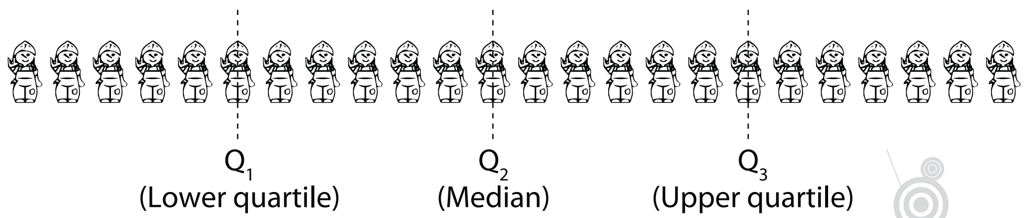


Example: Study the box and whisker plot below which summarises the shoe sizes for a class of 23 boys.



Supposing the boys were asked to stand in a straight line so that the boys with the smallest shoe size were on the left and the boys with the largest shoe size were on the right.

Look at the diagram depicting the ordered arrangement of the 23 boys.



The diagram shows the 12th boy is in the median position, the 6th boy is in the lower quartile position and the 6th boy from the right is in the uopper quartile position.

1. What was the shoe size of the boy in the middle?
   Answer: 8 (The median)

2. Tom is the boy standing 6th from the left. What is his shoe size?
   Answer: 7 (The lower quartile)

3. Sipho is the boy standing 6th from the right. What is his shoe size?
Answer: 11 (The upper Quartile)

4. What is the inter quartile range for this data?
Answer: 11-7 = 4 (IQR = $Q_3 - Q_1$)

(This is the range of values between the lower and upper quartile)

5. What is the range for the data?
Answer: 12 – 4 = 8 (Range = Max – Min)

6. Which part of the plot shows greater variability in the shoe sizes, the lower quarter or the upper quarter? Explain.

Answer: The lower quarter shows greater variability and/or spread. The whisker for this part is longer.

Ogive Curves (Cumulative Frequency curves)

In mathematics, the name ogive is applied to any continuous cumulative frequency polygon. (Its name is derived from its resemblance to the shape of architectural molding known as the ogee pattern.)

The ogive for the table below is sketched:

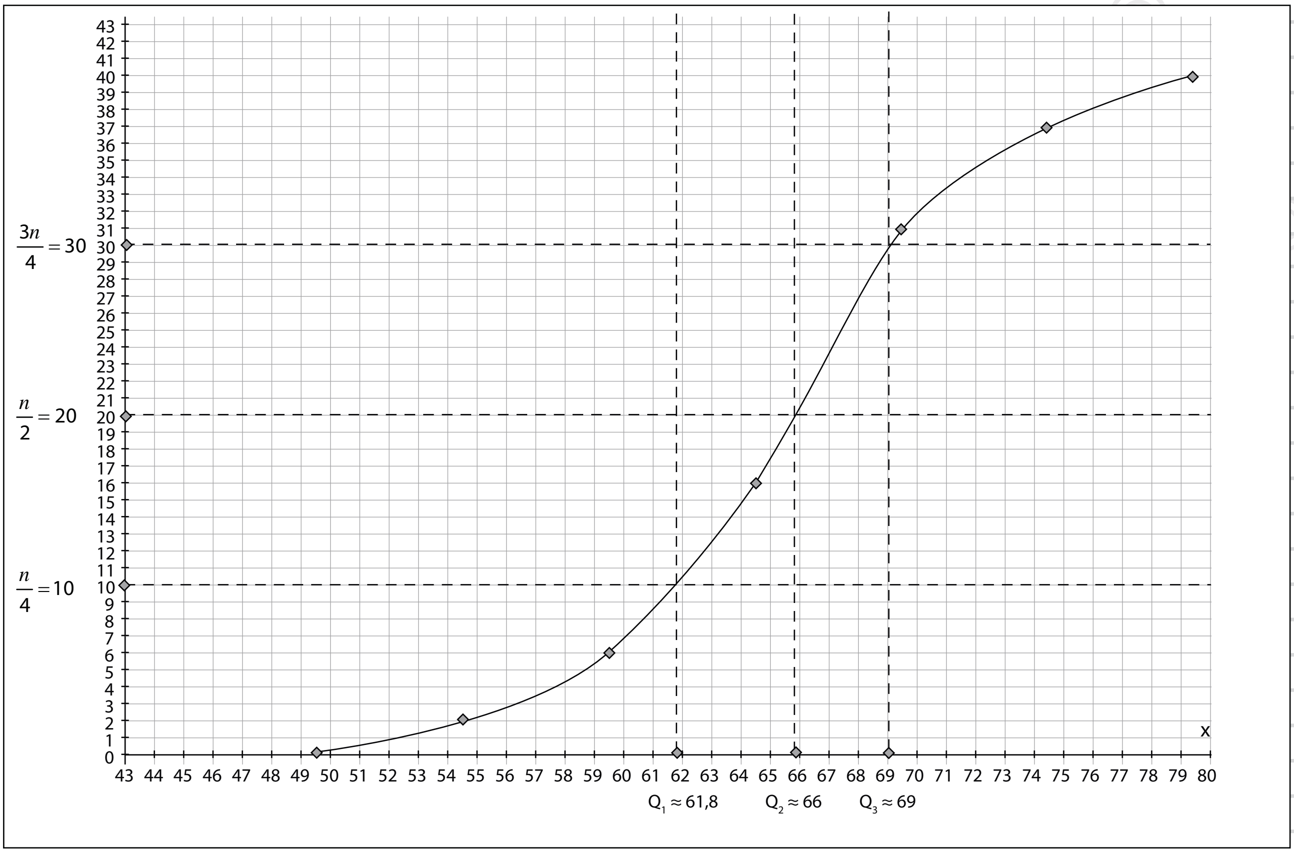
Weight	Cumulative Frequency
$49,5 < w \leq 54,5$	2
$54,5 < w \leq 59,5$	6
$59,5 < w \leq 64,5$	16
$64,5 < w \leq 69,5$	31
$69,5 < w \leq 74,5$	37
$74,5 < w \leq 79,5$	40

Interval	Frequency	Cumulative Frequency
$49,5 < w \leq 54,5$	2	2
$54,5 < w \leq 59,5$	4	2+4
$59,5 < w \leq 64,5$	10	2+4+10
$64,5 < w \leq 69,5$	15	2+4+10+15
$69,5 < w \leq 74,5$	6	2+4+10+15+6
$74,5 < w \leq 79,5$	3	2+4+10+15+6+3

Note: The Cumulative Frequency is the sum of all the frequencies within a specific interval or boundary. Every interval always starts at the lower band.

The Cumulative Frequency table is obtained from the frequency table given on the right.

The sum of all the frequencies is always equal to the Cumulative Frequency value.



To plot this graph we plot the cumulative frequency value against the end point value(x-value) for each interval. Use a smooth, continuous curve.

Notice that the points plotted were (54,5 ; 2) (59,5 ; 6) (64,5 ; 16) (69,5 ; 31) (74,5 ; 37) (79,5 ; 40).

One extra point is obtained by plotting (49,5 ; 0) which is the lower boundary of the lowest class interval 49,5 – 54,5. This is done because all the values must lie above 49,5.

Finding the Lower Quartile, Median and Upper Quartile using an ogive curve.

Remember that each quartile represents 25% of the values in our data set. So to calculate the position of the lower quartile we must find $\frac{1}{4}$ of the total number of values(n).

- Q_1 (lower quartile) is at $\frac{n}{4}$th position. In the example above, $n = 40$ therefore, Q_1 is in line with the cumulative frequency of the 10th observation.

Draw a horizontal line through the 10 on the vertical axis. At the point where this horizontal line cuts the curve, draw a vertical line from the point to the horizontal axis. The lower quartile is the value on the horizontal axis. In the example above $Q_1 \approx 61,8$

- Q_2 (median) is at $\frac{n}{2}$th position. In the example above, $n = 40$ therefore, Q_2 is in line with the cumulative frequency of the 20th observation.

Draw a horizontal line through the 20 on the vertical axis. At the point where this horizontal line cuts the curve, draw a vertical line from the point to the horizontal axis. The median is the value on the horizontal axis. In the example above $Q_2 \approx 66$

- Q_3 (upper quartile) is at $\frac{3n}{4}$th position. In the example above, $n = 40$ therefore, Q_3 is in line with the cumulative frequency of the 30th observation.

Draw a horizontal line through the 30 on the vertical axis. At the point where this horizontal line cuts the curve, draw a vertical line from the point to the horizontal axis. The upper quartile is the value on the horizontal axis. In the example above $Q_3 \approx 69$

Cumulative frequency curves make it very simple to answer questions that involve " less than" or "more than".

In the example above read off the number of students that weigh less than 69,5 kg.

Answer: 31 (Read from 69,5 kg on x-axis to the curve, to the y-axis)

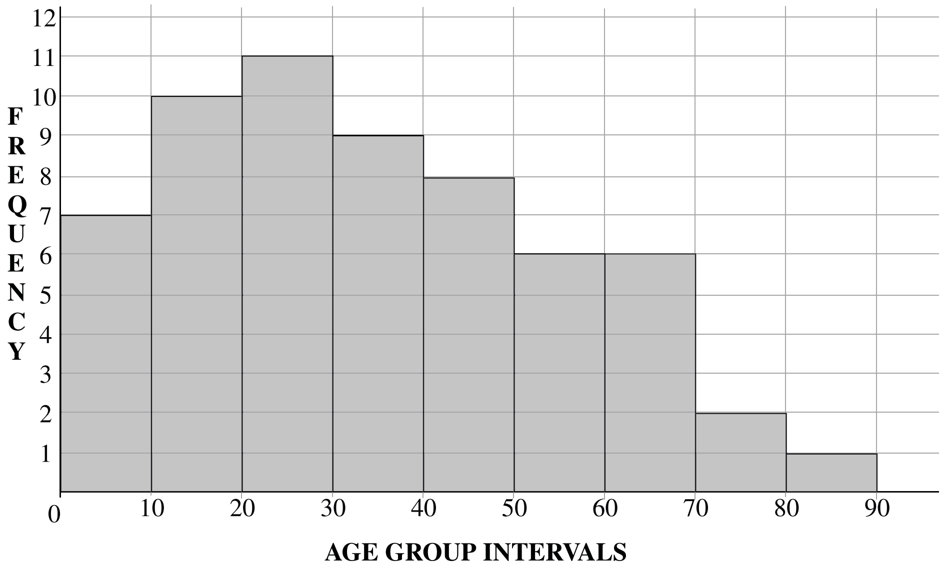
"More than " questions can also be answered but they are more difficult.

Use the curve above to read off how many students weigh more than 62,5 kg.

Answer : 40 – 12 = 28 (Read from x-axis 62,5 kg to curve, to y-axis to get 12)

Example

Example: Study the histogram given below and then answer the questions that follow:



Note: The bars 'touch' meaning that we are working with continuous data.

1. Complete the Cumulative frequency table below for the data given in the histogram above.

Age Group	Cumulative Frequency
0 < age ≤ 10	7
10 < age ≤ 20	17
20 < age ≤ 30	a
30 < age ≤ 40	b
40 < age ≤ 50	c
50 < age ≤ 60	d
60 < age ≤ 70	e
70 < age ≤ 80	59
80 < age ≤ 90	60

It might be easier for you to do this question if you include a frequency column in the table above as follows:(Each of the individual frequencies are obtained from reading off the heights of each bar).

Age Group	Frequency	Cumulative Frequency
0 < age ≤ 10	7	7
10 < age ≤ 20	10	17
20 < age ≤ 30	11	a
30 < age ≤ 40	9	b
40 < age ≤ 50	8	c
50 < age ≤ 60	6	d
60 < age ≤ 70	6	e
70 < age ≤ 80	2	59
80 < age ≤ 90	1	60

Answer: $a = 28$ $b = 37$ $c = 45$ $d = 51$ $e = 57$.

Remember that the cumulative frequency is the running total of all the individual frequencies.

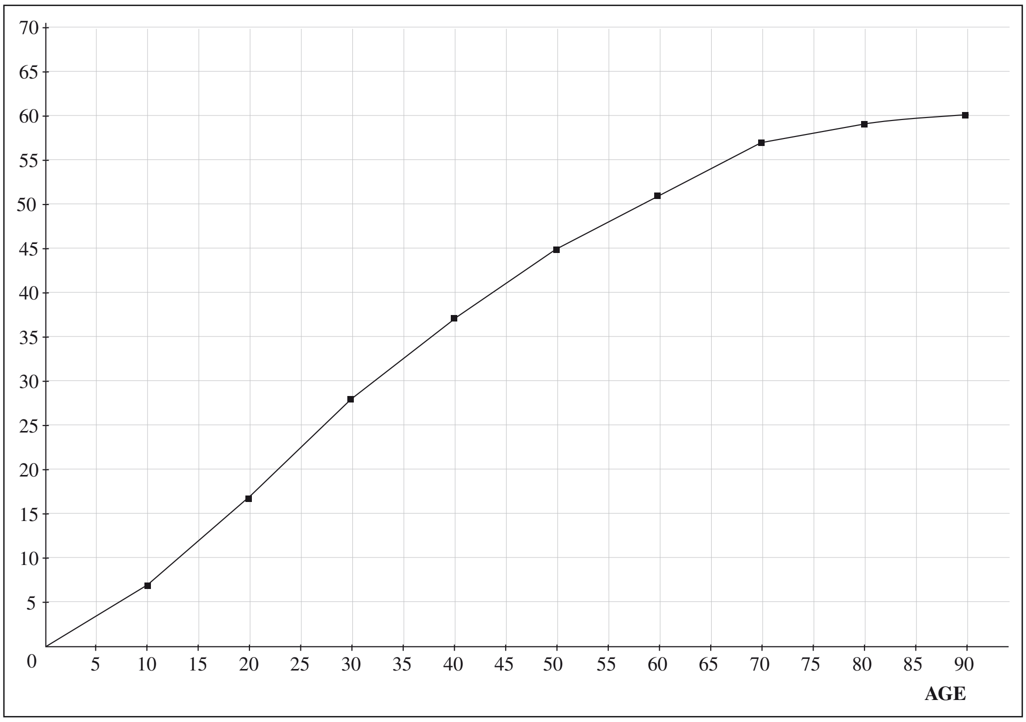
2. Draw an ogive curve for the data given above.

Answer: The following points are plotted: Note that we use the upper limits when plotting the points.

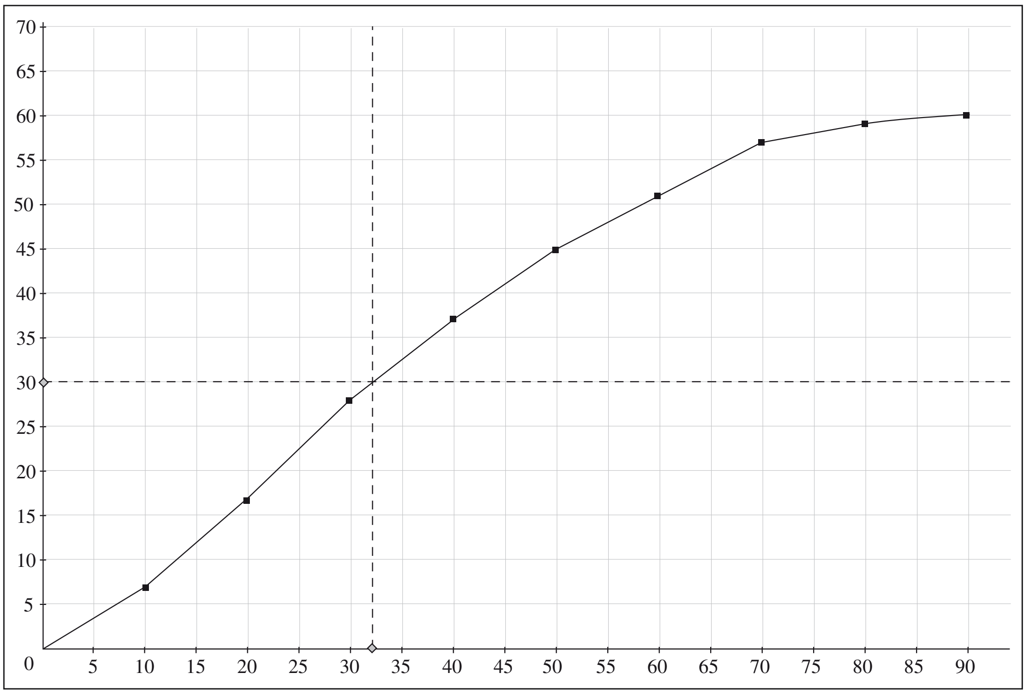
Age Group	Cumulative Frequency	Point to be plotted
0 < age ≤ 10	7	(10;7)
10 < age ≤ 20	17	(20;17)
20 < age ≤ 30	28	(30;28)
30 < age ≤ 40	37	(40;37)
40 < age ≤ 50	45	(50;45)
50 < age ≤ 60	51	(60;51)
60 < age ≤ 70	57	(70;57)
70 < age ≤ 80	59	(80;59)
80 < age ≤ 90	60	(90;60)

Once you have plotted the points you must join them with a smooth curve. Remember to include the point (0;0) which is the value of the lower limit of the first interval.

The Ogive Curve. Normally you will be given block paper in the exam to sketch the ogive curve.



3. Find the median age



The median age is obtained by drawing a horizontal line through 30 on the vertical axis(since $\frac{n}{2} = \frac{60}{2} = 30$. Then draw a vertical line down from the point where the horizontal line cuts the graph on the horizontal axis. Read the value of the median. In the example above the median value is approximately 32.

Standard deviation and Variance

The variance and the standard deviation are measures of how spread out a set of data is. In other words, they are measures of variability.

The variance is the average squared deviation of each number from the mean. For example, for the numbers 1, 2 and 3, the mean is 2 and the variance is:

$$\sigma^2 = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3} = \frac{2}{3} = 0{,}6667$$

$\bar{x}$, pronounced 'x bar' is the mathematical symbol for the mean.

The formula for the variance is given by $\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$

The standard deviation formula is very simple: it is the square root of the variance. It is the most commonly used measure of spread when we have large sets of numbers. (You do not need to memorise this formula, or the formula for variance, since they are given on the matric formula sheet.)

For example, for the numbers 1, 2 and 3, the standard deviation is $\sigma = \sqrt{0{,}667} = 0{,}8165$

We are also expected to compare the standard deviation of two sets of data. The larger the standard deviation, the greater the variability of the data (the greater the spread of the data).

It is often useful to use the following table when calculating the variance or standard deviation.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$

Manual calculation for finding σ – the standard deviation.

1. Find $\bar{x}$. (The mean average).
2. Subtract the mean from each of your values. (Column 2).
3. Square each of the results. (Column 3).
4. Add al the values in column 3, and divide by the total number of original values. i.e.: find the average of column 3. This answer is the variance. (σ^2).
5. To find the standard deviation, σ, square root the answer found in step 4.

Example

Example

Finding the standard deviation manually for 5 test scores:

62% ; 80% ; 71% ; 51% ; 86%

x ;	x; − $\bar{x}$	(x; −$\bar{x}$)^2
62	– 8	64
80	10	100
71	1	1
51	–19	361
86	16	256

$\bar{x} = 70$

$\therefore$ mean of column 3 $\Rightarrow \sigma^2 = \frac{64 + 100 + 1 + 361 + 256}{5}$

$\sigma^2 = 156{,}4$

$\sigma = \sqrt{156{,}4}$

$= 12{,}50599\ldots$

(Standard deviation)

The variance and/or standard deviation can be calculated easily with a calculator: Although you are encouraged to use a calculator to calculate the standard deviation, you must also be able to perform this calculation manually.

For the purposes of this discussion, we will use the CASIO fx – ES PLUS to demonstrate this:

Follow these steps to compute the standard deviation for 5 test scores: 62% ; 80% ; 71% ; 51% ; 86%.

Step 1: Press " SET UP" . Select 2: STAT

Step 2: Press 1: 1 – VAR

Step 3: Enter the numbers one by one followed by the equals after each number.

In this example, enter 62 then =; now enter 80 then =; now enter 71 then =, and so on. Remember to press = after the last entry.

Once you have completed entering all the data as described in step 3, press the AC button once.

Using the raw data

Step 4: Press the " shift" button and then the "1" button. (Notice that $\bar{x}$ is also an option here, so you can use your calculator to determine the mean.)

Select 4: VAR

Step 5: Select 3: $x\sigma n$ and then the "=" button.

The answer you get is 12,50599…

This answer is the standard deviation. If you need the variance simply square the result by pressing the "x^2 " button.

The variance is 156,4

Standard deviation / Variance of grouped data

The table used for a set of grouped data is slightly different as the frequency has to be taken into account now.

x_i	f_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$f_i\,(x_i - \bar{x})^2$

Note: This is slightly more complex calculation than before. You are strongly advised to learn hoe to peform this calculation using a calculator.

The formula for variance is now: $\sigma^2 = \dfrac{\sum_{i=1}^{n} f_i\,(x_i - \bar{x})2}{\sum_{i=1}^{n} fi}$

Note that when using the calculator be sure to put the frequency mode on.

On the CASIO fx – 82ES PLUS, this is done by pressing "SHIFT" then "SET UP". Scroll down and select 3:STAT. Then select 1:ON.

Example

Example

The shoe sizes of a group of 50 Grade 11 students were recorded and summarised in the table below:

Shoe size	4	5	6	7	8	9	10	11	12
Frequency	2	4	4	8	7	12	10	2	1

Calculate the standard deviation

1. using a table and
2. using a calculator

LIBERTY LIFE

Solution

1. First we must calculate the mean:

$$\bar{x} = \frac{(4 \times 2) + (5 \times 4) + (6 \times 4) + (7 \times 8) + (8 \times 7) + (9 \times 12) + (10 \times 10) + (11 \times 2) + (12 \times 11)}{50}$$

$= 8{,}12$

x_i	f_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$f_i (x_i - \bar{x})^2$
4	2	–4,12	16,9744	33,9488
5	4	–3,12	9,7344	38,9376
6	4	–2,12	4,4944	17,9776
7	8	–1,12	1,2544	10,0352
8	7	–0,12	0,0144	0,1008
9	12	0,88	0,7744	9,2928
10	10	1,88	3,5344	35,344
11	2	2,88	8,2944	16,5888
12	1	3,88	15,0544	15,0544
	$\sum_{i=1}^{9} f_i = 50$ (sum of this column)			$\sum_{i=1}^{9} f_i (x_i - \bar{x})^2 = 177{,}28$ (sum of this column)

Therefore the standard deviation $= \sigma = \sqrt{\frac{177{,}28}{50}} = 1{,}8830$

2. Using a calculator, it is much faster and easier.

Make sure you have put the frequency on.

Step 1: Press SET UP and select 2: STAT

Step 2: Select 1: 1- VAR

Step 3: Enter the data into the table as presented to you in the table above.

The shoe size under the "x" column and the frequency under the "FREQ" column.

Once you have completed entering the data correctly. PRESS the " AC" Button.

Step 4: Press " SHIFT" then the "1" button. Select 4: VAR
Step 5: Select 3: $x\sigma n$ and then the "=" button

The answer is $\sigma = 1{,}8822976367 \approx 1{,}8830$

Scatter Plots

In science, the scatter plot is widely used to present measurements of two or more related variables. It is particularly useful when the variables of the y-axis are thought to be dependent upon the values of the variable of the x-axis (usually an independent variable).

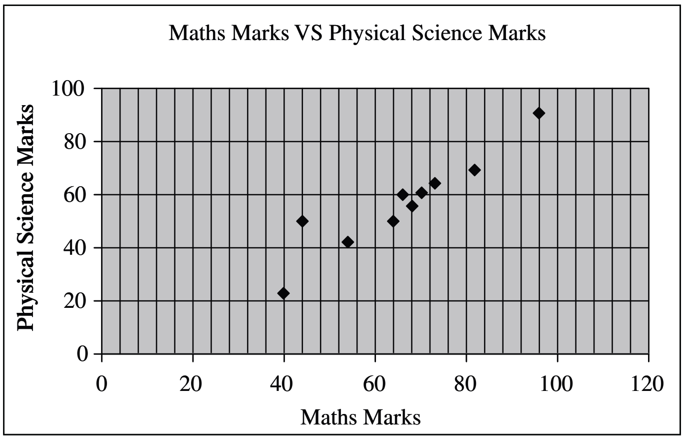
In a scatter plot, the data points are plotted but not joined The resulting pattern indicates the type and strength of the relationship between two or more variables.

You need to be able to intuitively state whether a trend is linear (Straight line), quadratic (parabola) or exponential.

Example

Example:

Learners' mathematics results were compared to their physical science results in the graph below.

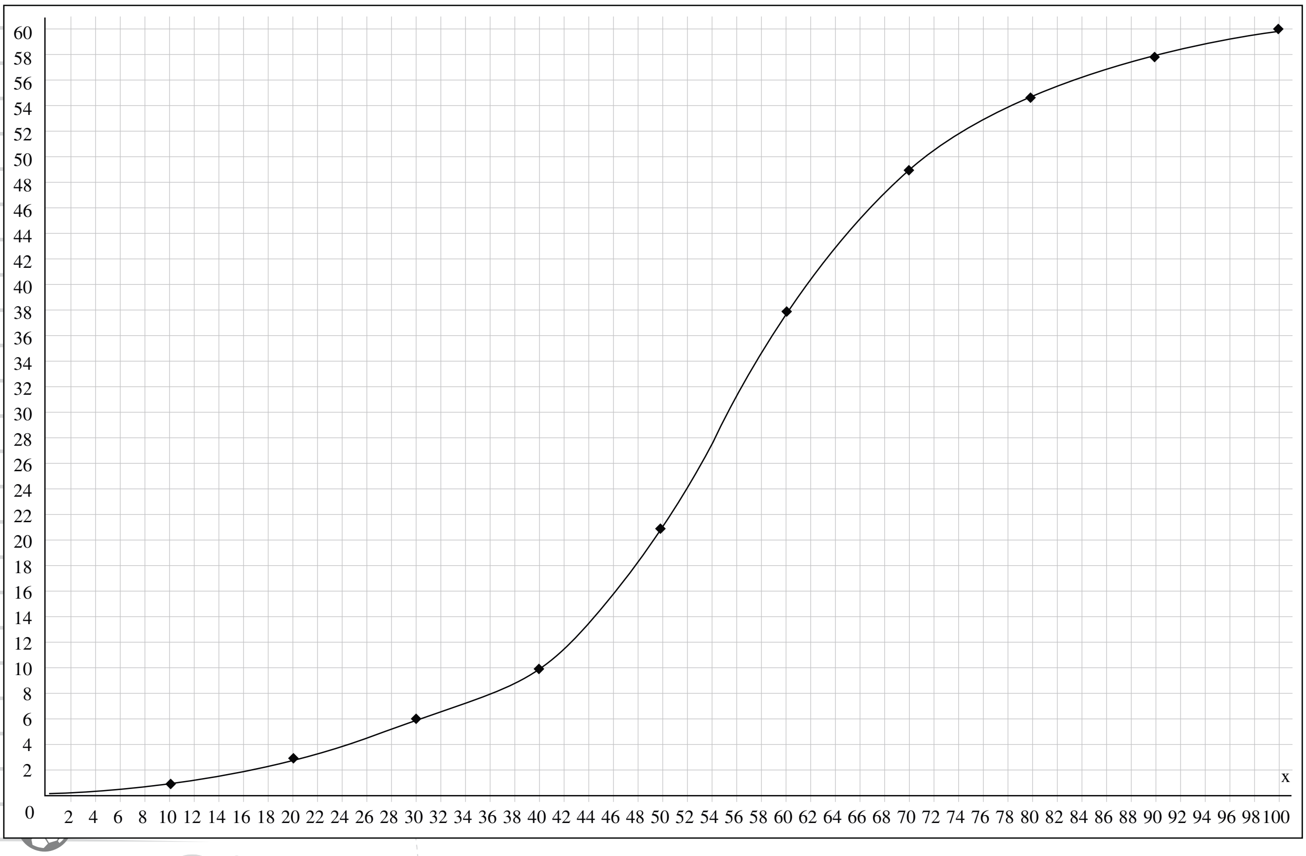


1. How many learner's results were compared?
   Answer: 10 (there are 10 data points on the plot).
2. Is the trend linear, quadratic or exponential?
   Answer: linear (A straight line would pass through most of the points)
3. Explain the trend.
   Answer: The greater their maths marks, the greater their science marks. We can also say that there is a positive correlation, since this gradient would be positive

Activity

Activity 1

1. Sixty candidates entered an examination in which the maximum mark was 100. The ogive curve was drawn from their marks.



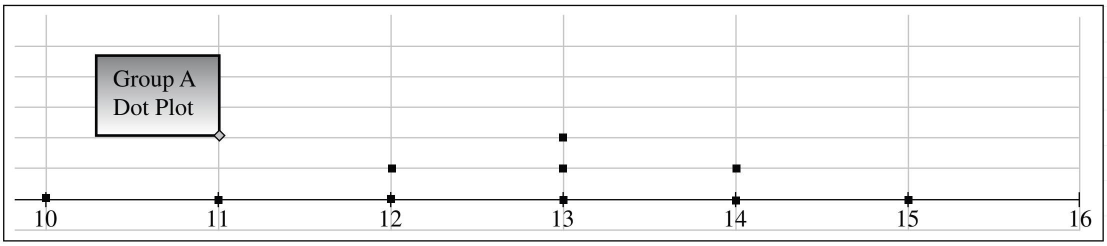
Use the graph to estimate

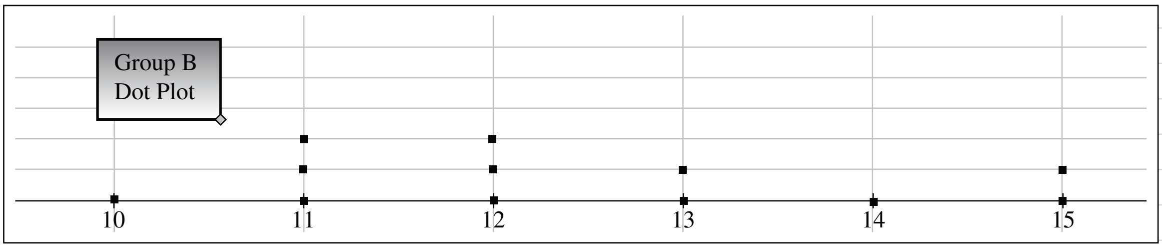
1.1 the median

1.2 the interquartile range

1.3 the number of candidates who scored 80 or more.

2.1 Find the standard deviations of the following age distributions of two groups of children. (Give the answer correct to 4 decimal figures). Use a table for Group A and the calculator for Group B.





2.2 Which group has a more uniform age distribution?

3. The number of children in 20 families is shown below:

Children	**Families**
0	1
1	4
2	9
3	3
4	2
5	1

3.1 Determine the five number summary for the data above.

3.2 Draw the box and whisker plot for the data above.

4. The table below shows the heights of 40 students in a class.

Height (cm)	$155<h\leq160$	$160<h\leq165$	$165<h\leq170$	$170<h\leq175$	$175<h\leq180$
Frequency	4	9	10	11	6

4.1 Determine an estimate for the mean

4.2 Determine an estimate for the standard deviation

4.3 Complete the cumulative frequency table below:

Height (cm)	$155<h\leq160$	$160<h\leq165$	$165<h\leq170$	$170<h\leq175$	$175<h\leq180$
Frequency	4	9	10	11	6
Cumulative Frequency					

4.4 Draw the cumulative frequency curve.

4.5 From the curve, estimate the interquartile range.

5. Two workers vacuum floors. Each vacuums 10 floors and the time it takes them to vacuum each floor is recorded, to the nearest minute:

Worker A: 3; 5; 2; 7; 10; 4; 5; 5; 4; 12

Worker B: 3; 4; 8; 6; 7; 8; 9; 10; 11; 9

5.1 For worker A's times, determine

5.1.1 the median time

5.1.2 the lower and upper quartiles

5.2 Draw two box and whisker plots to compare the times of each worker. Make a statement comparing their times.

5.3 Which worker would be best to employ? Give a reason for your answer.

6. The weights of a box of 20 chocolates are summarised as follows:

$$\sum_{i=1}^{20} w_i = 60\text{ g} \quad \sum_{i=1}^{20} (w_i - 3)^2 = 219\ g^2$$

6.1 Determine

6.1.1 the mean

6.1.2 the variance

6.1.3 the standard deviation

6.2 A second box containing 30 chocolates has a mean weight of 3 g and a standard deviation of 1. Which box contains chocolates that are more uniform in weight?

Solutions

1.1 55

1.2 Upper quartile = 66; lower quartile = 45

Therefore Interquartile range = 66 – 45 = 21

1.3 60 –55 = 5. Draw a vertical line through 80 on the horizontal axis. At the point where it cuts the graph draw a horizontal line so that it cuts the vertical axis. This number is 55.

2.1 From the dot plot Group A data is:

10, 11, 12, 12, 13, 13, 13, 14, 14, 15

Group A : $\bar{x} = \frac{10 + 11 + 12 + 12 + 13 + 13 + 13 + 14 + 14 + 15}{10} = 12{,}7$

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
10	–2,7	7,29
11	–1,7	2,89
12	–0,7	0,49
12	–0,7	0,49
13	0,3	0,09
13	0,3	0,09
13	0,3	0,09
14	1,3	1,69
14	1,3	1,69
15	2,3	5,29

$$\sum_{i=1}^{10} (x_i - \bar{x})^2 = 20{,}1 \quad \therefore \sigma = \sqrt{\frac{20{,}1}{10}} = 1{,}4177^{(4.d.p)}$$

From the dot plot Group B data is:

10, 11, 11, 11, 12, 12, 12, 13, 13, 14, 15, 15

Group B: Using CASIO fx –82ES PLUS

Step 1: Press Mode Select 2: Stat

Step 2: Select 1: 1-VAR

Step 3: Enter the data. Press "=" after every number entered. Press AC once all numbers are entered.

Step 4: Press "SHIFT" "1". Select 4: VAR

Step 5: Select 3: $x \sigma n$ and then press "="

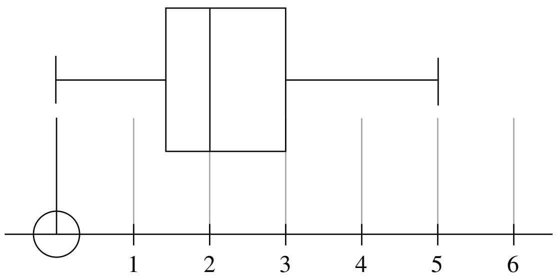
Using the calculator, we get $\sigma = 1{,}5523$

3.1 Five number summary:

Min = 0; Q1 = $\frac{1+2}{2}$ = 1,5; Q2 = $\frac{2+2}{2}$ = 2; Q3 = $\frac{3+3}{3}$ = 3; Max = 5

Since there are 20 families, The median is the average of the 10th and 11th families. The lower quartile is the median of the first half which is the average of the 5th and 6th families and the upper quartile is the median of the second half, which is the average of the 15th and 16th families.

3.2



4.1 mean = $\frac{157{,}5 \times 4 + 162{,}5 \times 9 + 167{,}5 \times 10 + 172{,}5 \times 11 + 177{,}5 \times 6}{40} = 168{,}25$

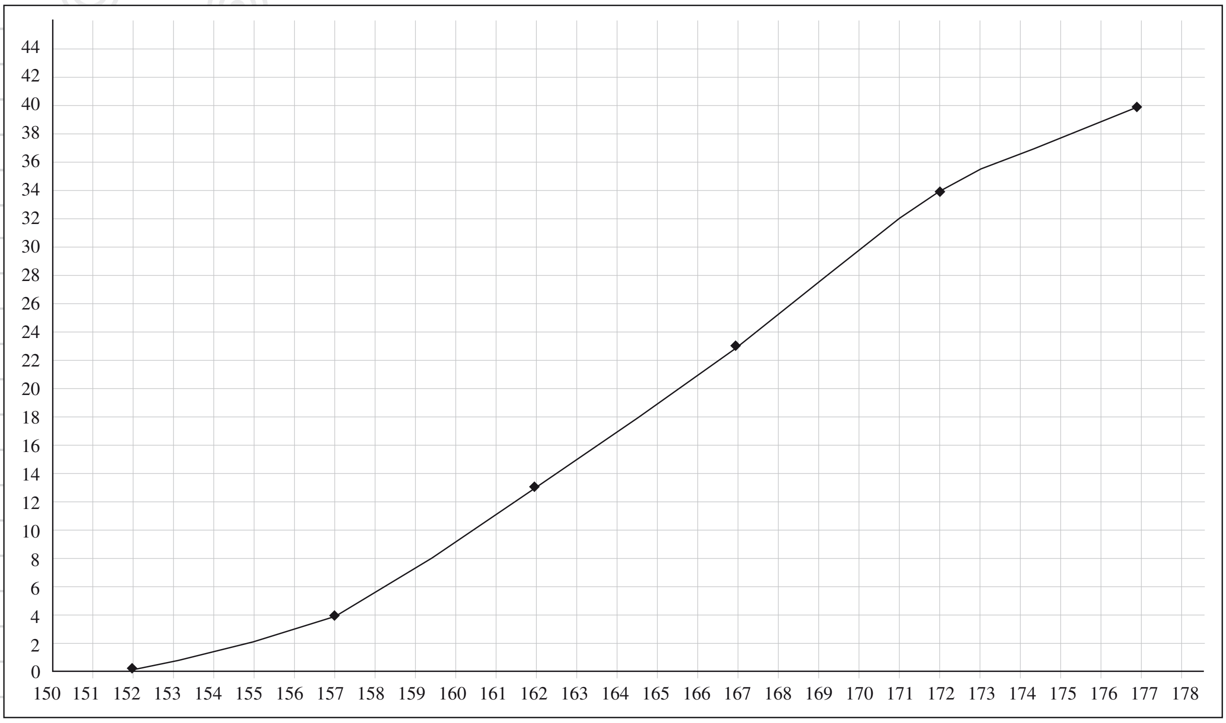
4.2 Using the calculator, $\sigma = 6{,}583$

Note: when we have grouped data we always use the midpoint of the internal as our x-value.

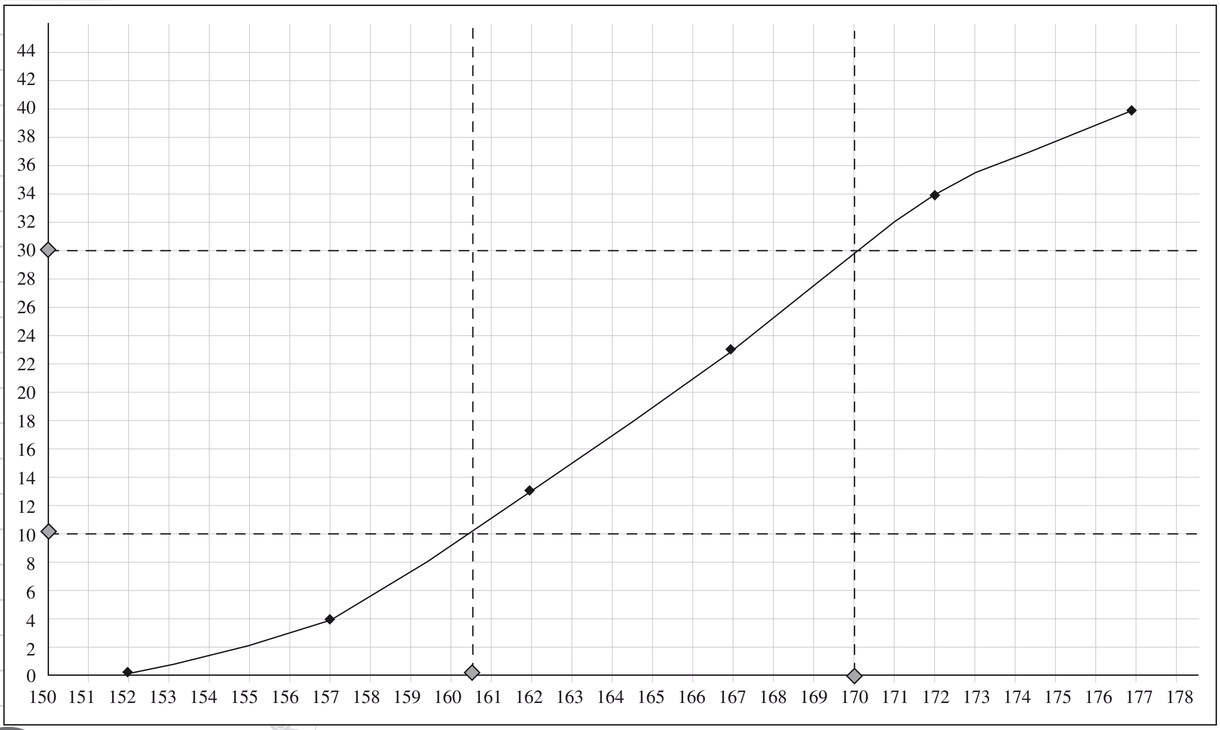
4.3

Height (cm)	$155<h\leq160$	$160<h\leq165$	$165<h\leq170$	$170<h\leq175$	$175<h\leq180$
Frequency	4	9	10	11	6
Cumulative Frequency	4	13	23	34	40

4.4



4.5 The dotted lines indicate where the lower quartile and upper quartile can be found.



The lower quartile = 163,5 The upper quartile = 173

Therefore the Interquartile range = 173 – 163,5 = 9,5

5. Worker A: 3; 5; 2; 7; 10; 4; 5; 5; 4; 12

Worker B: 3; 4; 8; 6; 7; 7; 9; 10; 11; 9

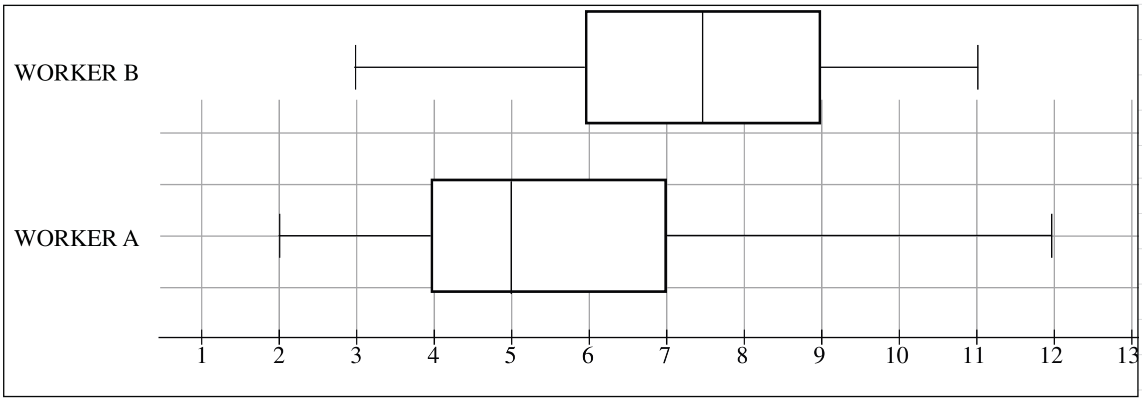
5.1 Worker A: In ascending order: 2; 3; 4; 4; 5; 5; 5; 7; 10; 12; Therefore

5.1.1 median = $\frac{5+5}{2} = 5$

5.1.2 Q1 = 4 and Q3 = 7

5.2 Worker B: In ascending order: 3; 4; 6; 7; 7; 8; 9; 9; 10; 11

Median = $\frac{7+8}{2}$ = 7,5   Q1 = 6 and Q3 = 9



The box and whiskers plot for worker A is longer than that for worker B.

5.3 On average, Worker A works faster, therefore worker A should be chosen.

Notice that the interquartile range is the same in both cases.

6.1.1 mean = $\frac{\sum_{i=1}^{20} w_i}{20} = \frac{60}{20} = 3$ g

6.1.2 variance = $\frac{\sum_{i=1}^{20} (w_i - 3)2}{20} = \frac{219}{20} = 10{,}95$

6.1.3 standard deviation = $\sqrt{10{,}95}$ = 3,309 (3d.p)

6.2 The second box has chocolates that are more uniform in weight as the standard deviation for that set is smaller.