# 6. Statistics

## A. Back to Basics – Measures of Central Tendency

Statistics is used in many different areas of life to show relationships between certain things.

You get many different types of statistics of which you learn about one major segment. This is where you are given a "bunch" of data that tells you about a certain group of people or things – for example the different heights in your class.

For example – Your teacher measured 20 people in your class and came out with the following data:

145 cm, 176 cm, 136 cm, 178cm, 159cm, 166cm, 176cm, 159cm, 143cm, 189cm, 155cm, 176cm, 162cm, 155cm, 156cm, 178cm, 122cm, 157cm, 165cm, 148cm

To input that into you calculator press:

MODE 1 To get into your statistics mode on the Sharp EL-W535.

Now press 0 to get to the single data mode.

To enter data simply type the first number, e.g. 145 into the calculator and then press DATA CD CHANGE to enter the data into your calculator, then enter the rest of the data into your calculator.

> Remember to make sure that your memory for the previous data is cleared before you enter your new data. To do this press:
>
> 2ndF ALPHA 1 0
>
> Remember that your calculator remembers all your data until you purposefully clear the memory.

Check that your data set = 20.

Your next step is to arrange your data in ascending order:

122; 136; 143; 145; 148; 155; 155; 156; 157; 159; 159; 162; 165; 166; 176; 176; 176; 178; 178; 189.

- The Mode:
  The mode is the value in your ungrouped data that appears the **MOST.**
- The Median:
  The median is the value that is in the **MIDDLE** of your data.
- The Mean:
  The mean is the average of the data because it is **NASTY** to work out ☺

To find the mode look at your data and look for the value that appears the most times: in our example this would be: 176.

In this case we can see that the mode does not accurately represent the data that we have.

To find the median use this formula: position = $\frac{(n+1)}{2}$

In our example then the position of our median would be:

$$p = \frac{20+1}{2}$$

$$\therefore p = 10{,}5$$

This means that our median lies between two values: 159 and 159. It is easy to see that our median is therefore 159.

To find the mean there are two methods – the long and the short way:

- Method 1 – The long way
  - $mean = \frac{the\ sum\ of\ all\ scores}{the\ number\ of\ scores} = \frac{\sum x}{n} = \bar{x}$

$$mean = \frac{145+176+136+178+159+166+176+159+143+189+155+176+162+155+156+178+122+157+165+148}{20}$$

$$\therefore mean = \frac{3\ 201}{20}$$

$$\therefore mean = 160,05cm$$

To find the sum with your calculator simply press:

ALPHA DRG▸ Σx • = ENTER

- Method 2 – The short Method:
  - Press ALPHA $n!$ $\bar{x}$ 4 = ENTER (Just remember to show your working out).

*Grouped data:*

Grouped data is data that has already been organised in some form: for example:

Your class writes a maths test and gets the following marks:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 56; | 78; | 34; | 89; | 67; | 45; | 65; | 67; | 69; | 21; |
| 49; | 35; | 67; | 72; | 78; | 83; | 75; | 48; | 63; | 58; |
| 63; | 46; | 90; | 56; | 57; | 82; | 67; | 69; | 54; | 77 |

Your teacher groups the data in order to give herself an idea of what the spread of the data looks like.

| Mark | Frequency | Mark | Frequency |
|---|---|---|---|
| 0 – 9 | 0 | 50 – 59 | 5 |
| 10 – 19 | 0 | 60 – 69 | 9 |
| 20 – 29 | 1 | 70 – 79 | 5 |
| 30 – 39 | 2 | 80 – 89 | 3 |
| 40 – 49 | 4 | 90 – 99 | 1 |

- To find the Mode:
  - Look for the group with the highest frequency (the frequency that is the most) → this is the group 60 – 69.

- To find the median: you need to add a cumulative frequency to the table:

| Mark | Freq. | Cum. Freq. | Mark | Freq. | Cum. Freq. |
|---|---|---|---|---|---|
| 0 – 9 | 0 | 0 | 50 – 59 | 5 | 12 |
| 10 – 19 | 0 | 0 | 60 – 69 | 9 | 21 |
| 20 – 29 | 1 | 1 | 70 – 79 | 5 | 26 |
| 30 – 39 | 2 | 3 | 80 – 89 | 3 | 29 |
| 40 – 49 | 4 | 7 | 90 – 99 | 1 | 30 |

Now take 50% of your cumulative frequency total:
$0.5 \times 30 = 15$
So then look for where the cumulative frequency would equal 15 and that is the median group, in this case 60 – 69.

- To find the average mean:
  - Find the midpoint of each group:

| Mark | Freq. | C. F. | Mid | Mark | Freq. | C.F. | Mid |
|---|---|---|---|---|---|---|---|
| 0 – 9 | 0 | 0 | 4,5 | 50 – 59 | 5 | 12 | 54,5 |
| 10 – 19 | 0 | 0 | 14, 5 | 60 – 69 | 9 | 21 | 64,5 |
| 20 – 29 | 1 | 1 | 24,5 | 70 – 79 | 5 | 26 | 74,5 |
| 30 – 39 | 2 | 3 | 34,5 | 80 – 89 | 3 | 29 | 84,5 |
| 40 – 49 | 4 | 7 | 44,5 | 90 – 99 | 1 | 30 | 94,5 |

Again there are two methods:

- Method 1 – the long way:
  - $mean = \frac{the\ sum\ of\ all\ the\ midpoints\ times\ their\ frequencies}{the\ total\ number\ of\ scores}$
  - $mean = \frac{4{,}5\times0+14{,}5\times0+24{,}5\times1+34{,}5\times2+44{,}5\times4+54{,}5\times5+64{,}5\times9+74{,}5\times5+84{,}5\times3+94{,}5\times1}{30}$
  - $mean = \frac{0+0+24{,}5+69+178+272{,}5+580{,}5+372{,}5+253{,}5+94{,}5}{30}$
  - $mean = \frac{1\,845}{30}$
  - $mean = 61{,}5$

- Method 2 – the short (and easy way)
  - Put your calculator into Statistics mode by pressing:

    MODE 1
  - Then choose the “SD” (Single data) function by pressing:

    0

  Make sure you clear your previous data first ☺
  - To enter your data enter your mid-point first, 4.5, then press (x,y) and then enter your frequency, 0 and then press DATA CD CHANGE.
  - Now enter the rest of your data in the same way until your data set is = 10
  - To find the mean press:

    ALPHA 4 ($n!$ $\bar{x}$) = ENTER and the answer is.... 61.5 ☺

## B. Measures of Dispersion

*Range:*

The range gives the “width” of your data, for example you have a range of shoes – or all different types of shoes.

To find the range you say: Range = Maximum – minimum

or biggest – smallest

Example: Look at the following data and determine the range:

24; 36; 6; 46; 40; 21; 88; 23; 87; 36; 17

First find the biggest and smallest value:

Maximum = 88

Minimum = 6

$\therefore Range = 88 - 6 = 82$

*Quartiles:*

Just like the Median finds the middle position of the data or half of the data the quartiles find the quarters of the data. Before you can find the quartiles you need to make sure that your data is arranged in ascending order (i.e. from smallest to biggest).

The first quartile position $= \frac{n+1}{4}$

And the third quartile position $= \frac{3(n+1)}{4}$

So looking at the previous example, find the first and third quartile:

6; 17; 21; 23; 24; 36; 36; 40; 46; 87; 88

First quartile position = $\frac{11+1}{4} = 3$

Therefore $Q_1$ = 21

Third Quartile position = $\frac{3(11+1)}{4} = 9$

Therefore $Q_3$ = 46

*Interquartile range:*

The interquartile is the range between quartile 1 and quartile 3:

IQR = $Q_3 - Q_1$

From the example: IQR = 46 – 21

= 25

*The Semi-Interquartile range:*

Is half of the interquartile range in other words: SQR $= \frac{Q_3 - Q_1}{2}$

From the previous example: SQR = $\frac{46-21}{2} = 12\frac{1}{2}$

*Percentiles:*

Percentiles divide the data into 100 equal parts – in other words you need to give the actual data point for the percentage mentioned.

Percentile position = $\frac{\%(n+1)}{100}$

Find the 65th percentile for the previous example:

Percentile position = $\frac{65(11+1)}{100} = 7{,}8 \approx 8$

Therefore the 65th percentile is 40.

It is important to remember that your sample size affects your mean and the prediction of your mean using the median and mode.
The bigger your sample size the closer together your median and mean will be and the more able you will be able to predict your mean from your median.
The smaller your sample size the less you will be able to predict your mean from the median.

In the same way: If your sample size is bigger your standard deviation will be more accurate and will give a better indication of the spread of data, whereas the smaller your sample size the less accurate your standard deviation is.

*Standard Deviation*

There are two ways to find the standard deviation – the long way and the short way.

First some theory:

The formula for the standard deviation is $\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$

To find the variance simply square your standard deviation answer. Standard deviation tells us how spread out the data is – the bigger the value of the standard deviation the more spread out the data is and vice versa.

If the data is not skewed and from a relatively large sample then we can say the following things about the data:

Approximately 99,5% of the data lies within 3 standard deviations of the mean. While approximately 95% of the data lies within 2 standard deviations of the mean. Finally, approximately 66% (or two thirds) of the data lies within one standard deviation of the mean.

To find the Standard Deviation the Long Way:

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

What this formula means is: the sum of the *x-data point* minus the average (mean) and then squared, after which it is divided by the number of observations.

The easiest way to work out standard deviation is to use a table.

Looking at our example:

6; 17; 21; 23; 24; 36; 36; 40; 46; 87; 88

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

$\therefore\ \bar{x} = \frac{6+17+21+23+24+36+36+40+46+87+88}{11}$

$\therefore\ \bar{x} = \frac{424}{11}$

$\therefore\ \bar{x} = 38.54545455$

Now put your information in a table:

| $x$ | $x - \bar{x}$ | $(x - \bar{x})^2$ | ***answer*** |
|---|---|---|---|
| 6 | 6 – 38.54 | $(-32.54)^2$ | 1 059.206612 |
| 17 | 17 – 38.54 | $(-21.54)^2$ | 464.2066116 |
| 21 | 21 – 38.54 | $(-17.54)^2$ | 307.8429752 |
| 23 | 23 – 38.54 | $(-15.54)^2$ | 241.661157 |
| 24 | 24 – 38.54 | $(-14.54)^2$ | 211.5702479 |
| 36 | 36 – 38.54 | $(-2.54)^2$ | 6.479338843 |
| 36 | 36 – 38.54 | $(-2.54)^2$ | 6.479338843 |
| 40 | 40 – 38.54 | $(1.46)^2$ | 2.115702479 |
| 46 | 46 – 38.54 | $(7.46)^2$ | 55.57024793 |
| 87 | 87 – 38.54 | $(48.46)^2$ | 2 347.842975 |
| 88 | 88 – 38.54 | $(49.46)^2$ | 2 445.752066 |
| **Total or** $\sum(x - \bar{x})^2$ | | | 7 093.157025 |

Now we put this back into our formula:

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

$$\sigma = \sqrt{\frac{7\,093.157025}{11}}$$

$$\sigma = \sqrt{644.8324568}$$

$\sigma = 25.39355148$ or $\sigma \approx 25.39$

This means that the average distance between two points in our set of data is 25.39.

*The Short Way (or Using your Calculator – EL-W535HHT)*

Press MODE 1

Then press 0 for Single Data

(Make sure your data is cleared by pressing 2ndF ALPHA 1 0)

To input your data from the example press the first value (in this case 6) and then DATA CD CHANGE your calculator should say DATA SET = 1.

Now do this with all your other data from the example.

To find the standard deviation press ALPHA 6 = ENTER

∴ Standard Deviation = 25.49

Your two answers will be a little bit different because of the accuracy of your answers and your table.

Activity 1

1. A research company compares the wages of different part-time jobs of students in a university class. They receive the following information from 30 students:

1 140; 1 800; 900; 1 760; 1 800; 1 900;

1 080; 1 740; 1 220; 1 720; 1 800; 1 840;

780; 1 600; 1 400; 840; 1 120; 1 680;

1 860; 1 740; 1 600; 1 480; 1 640; 1 640;

1 520; 1 080; 1 560; 1 900; 880; 1 500

a) What is the mean of the data?

b) What is the median of the data?

c) Determine the mode of the data.

d) Determine quartile 1 and quartile 3.

e) Determine both the range and the interquartile range.

f) Find the standard deviation of the data.

g) What percentage of students earn below R 1 000 in wages?

h) What is the top 10% of wages that the students earn?

2. A scientist is measuring the growth of different bacteria in mm$^2$. He gets the following results for his 20 experiments:

2.71; 3.21; 3.84; 5.11; 3.92; 5.32;

4.06; 1.00; 4.33; 3.79; 4.57; 1.32;

3.79; 3.70; 4.11; 3.78; 2.02; 2.65;

1.78; 2.97

a) What is the mean of the data?

b) What is the median of the data?

c) Determine the mode of the data.

d) Determine quartile 1 and quartile 3.

e) Determine both the range and the interquartile range.

f) Find the standard deviation of the data.

g) What percentage of bacteria grew to less than 2mm$^2$.

h) How many cultures (bacterial experiments) were in the top 15% for growth and how big would the smallest of this group have been?

i) How would the mean be affected if we added the following 5 results to the data? Show your working out.

8.29; 7.02; 4.85; 6.90; 5.31

j) Would the standard deviation also be affected by these changes (from question (i) and by how much would it change by?

## C. Box and Whisker Plots

A Box and whisker plot is a diagram of what the data looks like. There are 5 important things (often called the 5-number summary) that you need to remember in order to draw the plot.

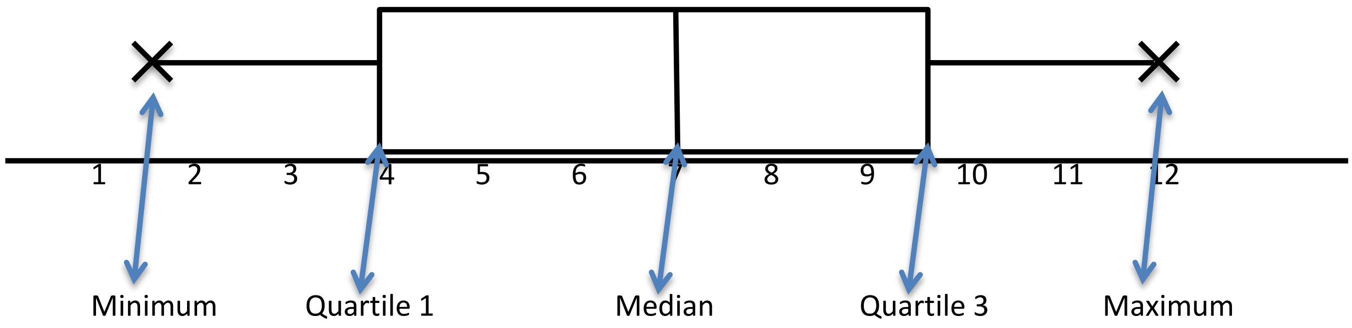
*5-number summary:*

Minimum Maximum

Quartile 1 Quartile 2 Quartile 3

a.k.a. → the median

The graph looks like this:



Example: A coffee shop counts the number of cappuccinos that they sell on any one day for two weeks. These are their results:

34; 44; 99; 39; 10; 56; 71; 71; 41; 93; 89; 11;
77; 68

Draw a box and whisker plot for the above example.

The First step is to order your data in ascending order:

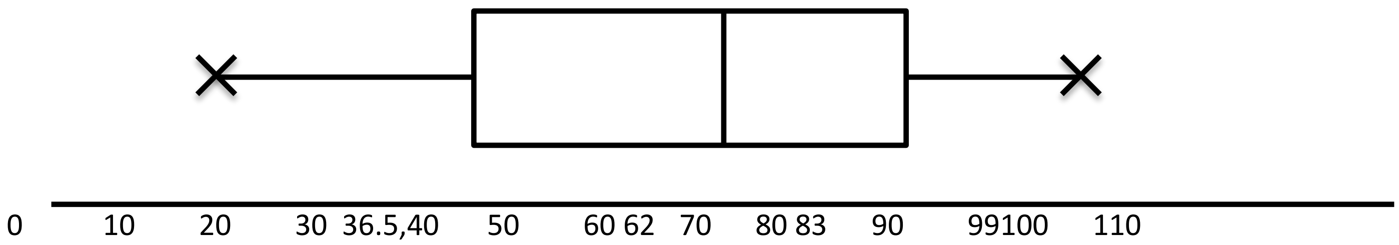
10; 11; 34; 39; 41; 44; 56; 68; 71; 71; 77; 89; 93; 99

The second step is to write down your 5-number summary:

Minimum → 10 Maximum → 99

Quartile 1 → $\frac{34+39}{2} = 36\frac{1}{2}$ Median → $\frac{56+68}{2} = 62$ Quartile 3 → $\frac{77+89}{2} = 83$

The last step is to use this information to draw the graph:



*Analysing the Box and Whisker Plot*

Every Quartile represents 25% of the data; in other words:

- From the minimum to quartile 1 → 25% of the data
- From Quartile 1 to median → 25% of the data
- From the median to quartile 3 → 25% of the data
- From quartile 3 to the maximum → 25% of the data
- From the minimum to the median → 50% of the data
- From the median to the maximum → 50% of the data
- From the minimum to quartile 3 → 75% of the data
- From quartile 1 to the maximum → 75% of the data

Activity 2

1. Your teacher measures the heights of your classmates in meters and gets the following results in your class of 18:

| | | | | | |
|---|---|---|---|---|---|
| 1.70; | 1.96; | 1.41; | 1.78; | 1.67; | 1.19; |
| 1.75; | 1.49; | 1.16; | 1.39; | 1.55; | 1.77; |
| 1.26; | 1.51; | 1.32; | 1.56; | 1.39; | 1.83 |

a) Give the five-number summary of the above data

b) Draw a box and whisker diagram of the above information

c) What percentage of students' heights lies above 1.53m?

2. Bob collects two different sets of data about the number of times girls go to movies in a year and the number of times boys go to movies in a year. Here is Bob's information:

Girls: 14; 32; 14; 28; 21; 5; 13; 29; 12; 9

Boys: 14; 29; 15; 24; 14; 1; 28; 18; 23; 30

a) Give the five-number summary for both sets of data.

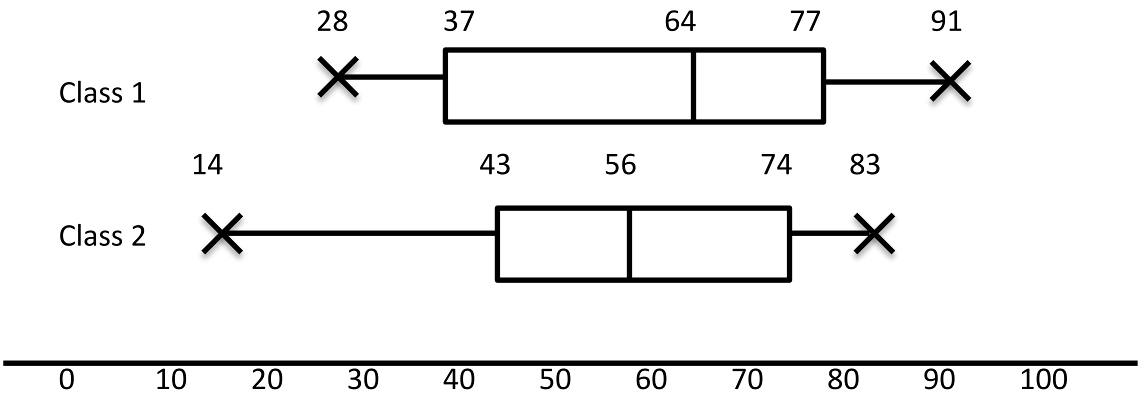
b) Draw box and whisker plots for both sets of data and use them to answer the questions that follow.

c) Briefly describe the distribution of each set of data.

d) Did more boys or more girls watch less than 14 movies?

e) What percentage of girls watched between 14 and 28 movies and what percentage of boys watched between 14 and 28 movies? Therefore were there more girls that watched between 14 and 28 movies, than boys or vice versa?

3. Below are two box and whisker diagrams for the geography marks of two different classes. Study them carefully before answering the questions that follow.



a) Which class has the greater range of marks?

b) In your opinion, which class did better in the test? Give a reason for your answer.

c) What is the interquartile range for each class? Do you think that the interquartile range gives a better indication of the spread of the data than the range? Give a reason for your answer.

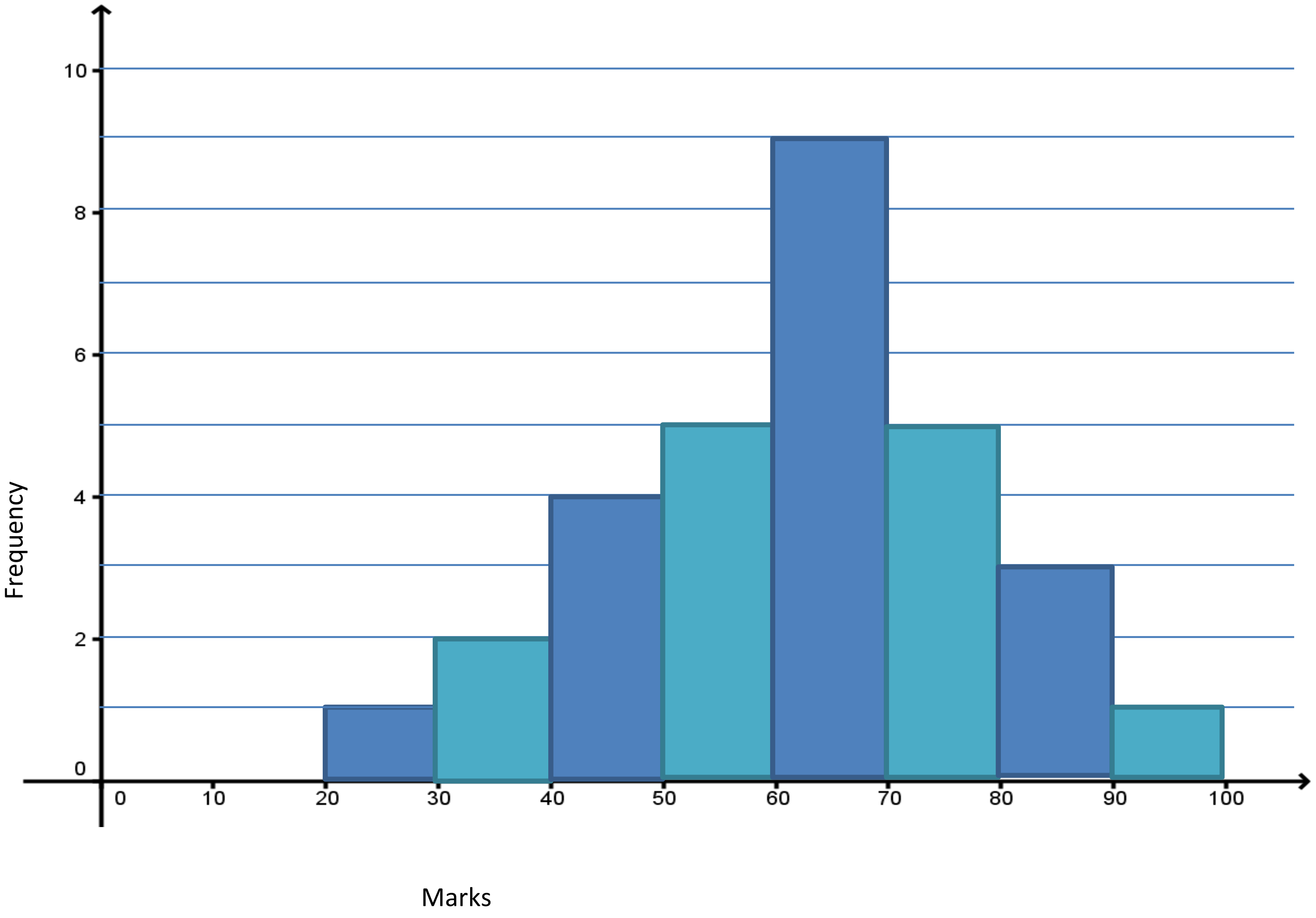
## D. Histograms, Frequency Polygons and Ogives

A histogram represents the distribution of data for grouped data (or the frequency of data against the type of observations) – like a bar graph but with NO gaps.

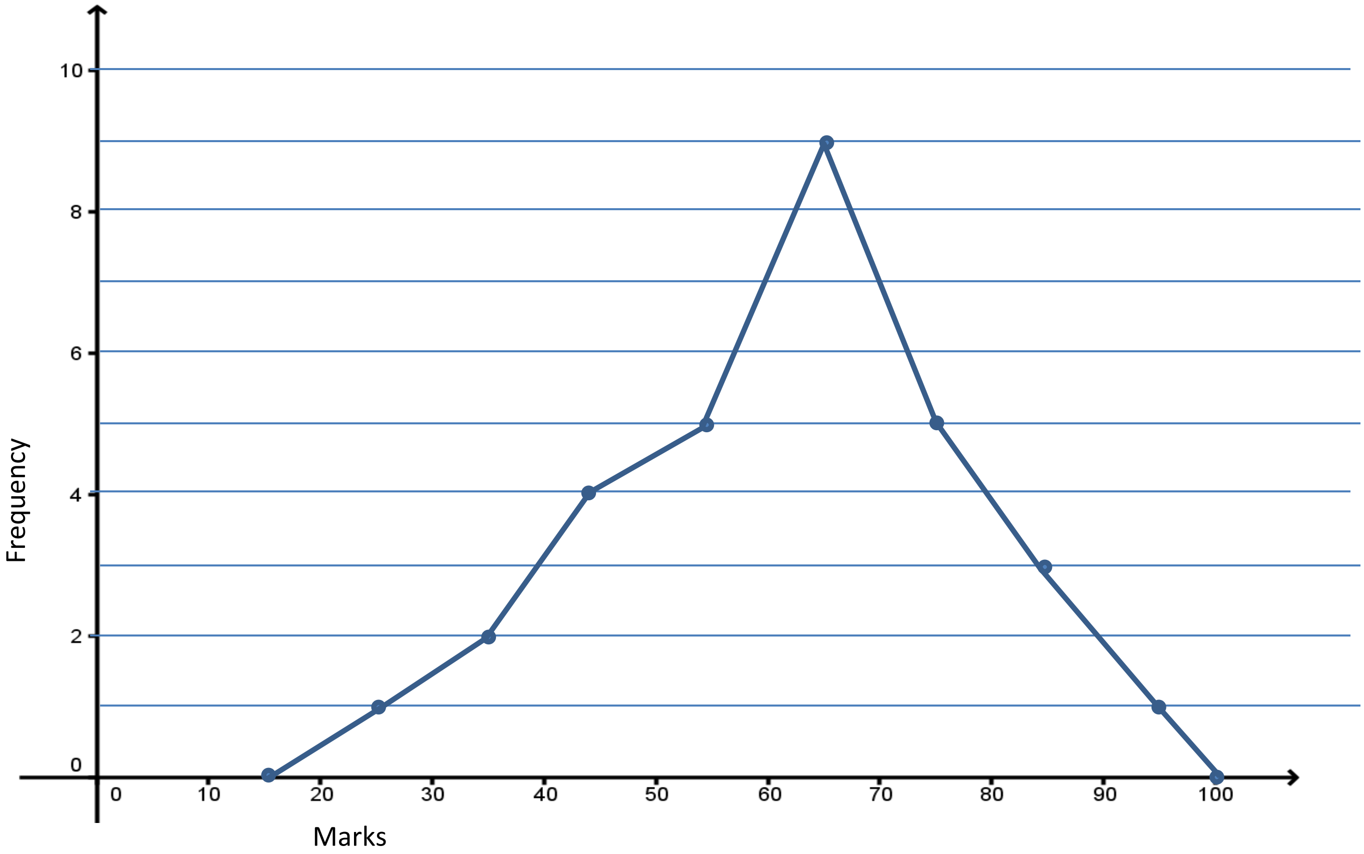
From our previous example:

| Mark | Freq. | Cum. Freq. | Mark | Freq. | Cum. Freq. |
|---|---|---|---|---|---|
| 0 – 9 | 0 | 0 | 50 – 59 | 5 | 12 |
| 10 – 19 | 0 | 0 | 60 – 69 | 9 | 21 |
| 20 – 29 | 1 | 1 | 70 – 79 | 5 | 26 |
| 30 – 39 | 2 | 3 | 80 – 89 | 3 | 29 |
| 40 – 49 | 4 | 7 | 90 – 99 | 1 | 30 |

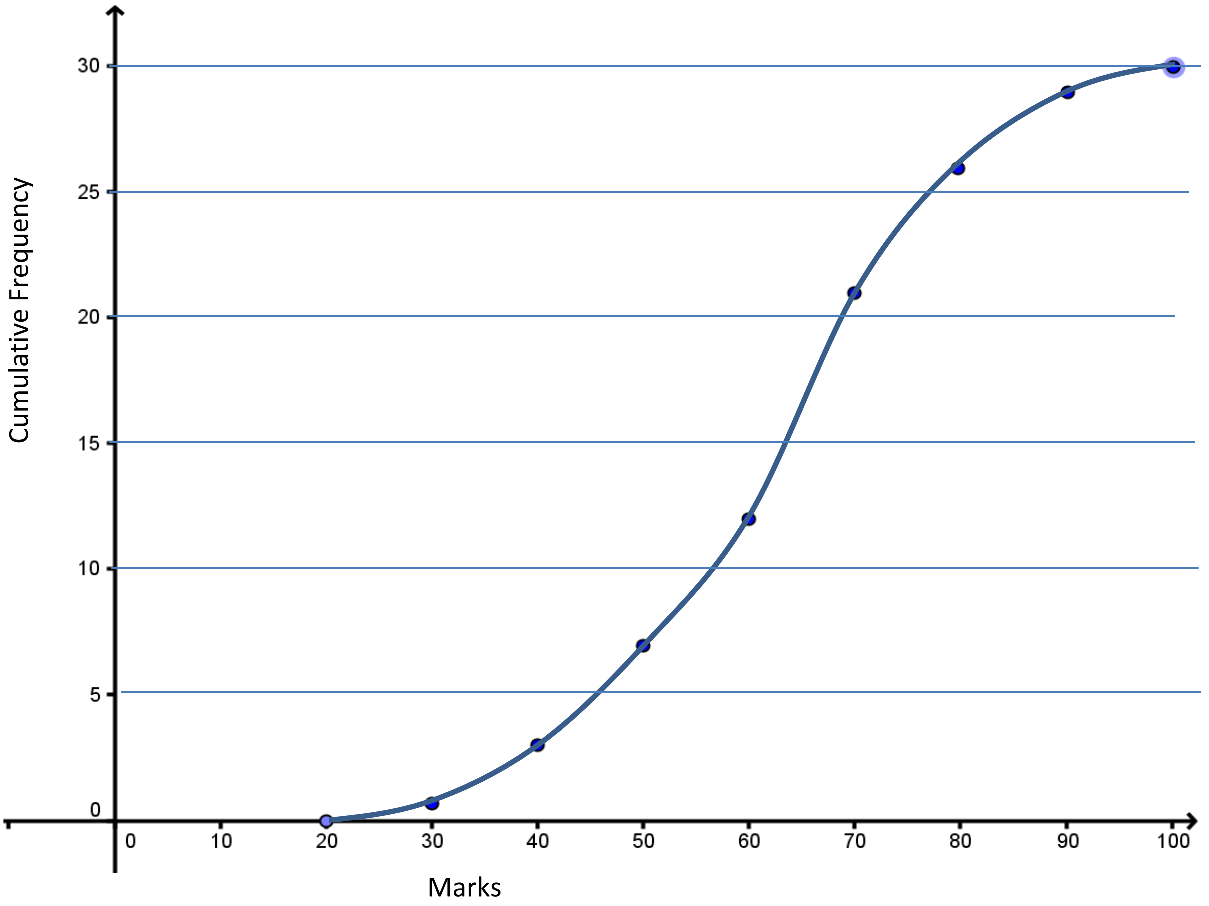
We can draw a histogram with the marks on the $x$-axis and the frequency on the $y$-axis.



A frequency polygon is a line graph of the frequency on the $y$- axis and the groups on the $x$-axis. The frequency is plotted at the mid-point of each group.



An ogive is a cumulative frequency straight line graph. The cumulative frequency is plotted at the end of each corresponding group either in frequency (most commonly) or percentage format. From the table our ogive should look like this:



You can use your ogive to find your median and your first and third quartiles as well.

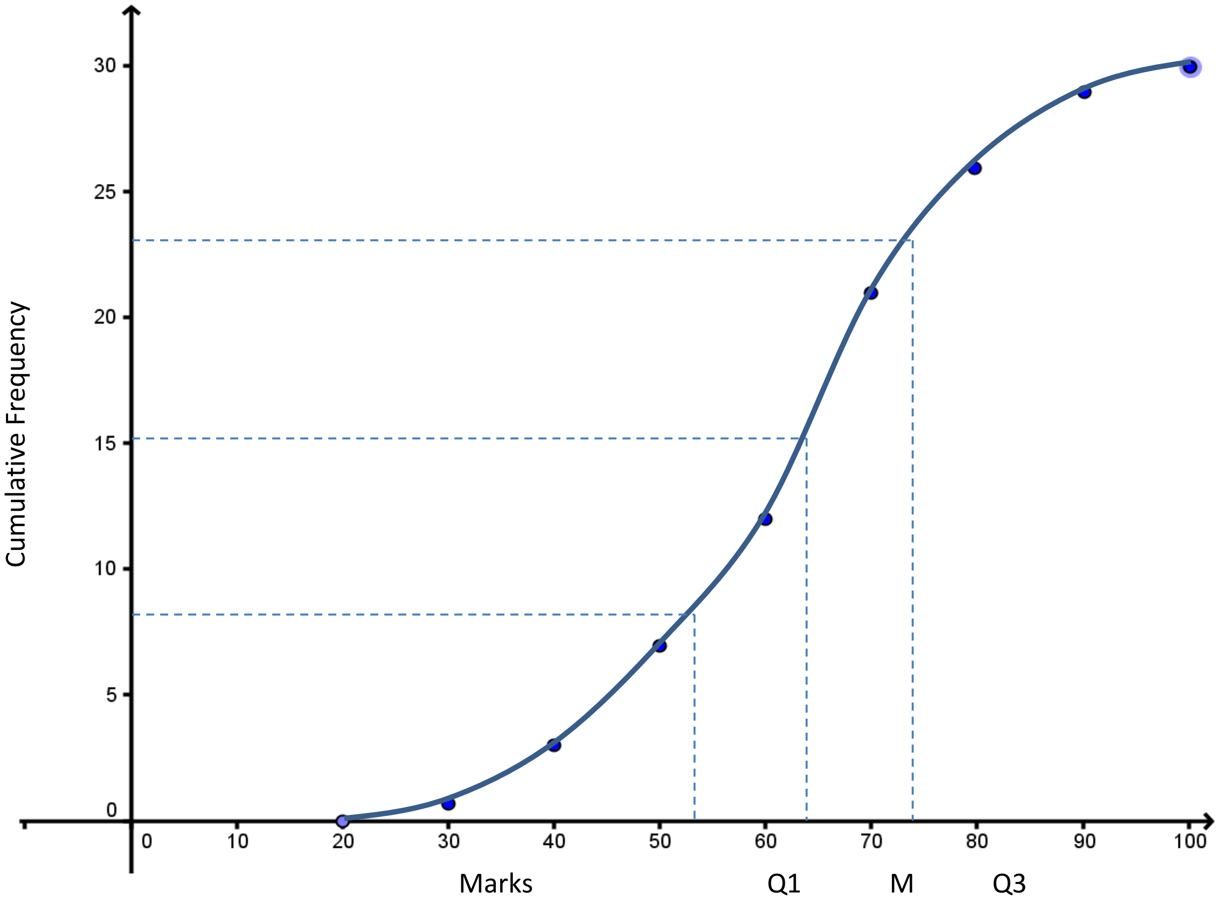
To find the median $\frac{n+1}{2} = \frac{30+1}{2} = 15.5.$

Find 15.5 on your cumulative frequency axis draw a line across until it meets the graph and then draw a line down to the $x$-axis to find the corresponding median.

To find the first and third quartile follow the same procedure:

First Quartile: $\frac{n+1}{4} = \frac{30+1}{4} = 7.75 \approx 8$

Third Quartile:$\frac{3(n+1)}{4} = \frac{3(30+1)}{4} = 23.25 \ \approx 23$



Thus the first quartile is approximately 53.

The median is approximately 63.

The third quartile is approximately 74.

Activity 3

1. The department of road-works needs to decide whether a certain road needs another lane. They decide to count the number of cars that travel on a certain part of the road everyday at a certain time. They do this for 1 month. This is the data that they found:

1 250; 1 750; 2 800; 3 250; 2 550; 2 600;

4 500; 2 700; 900; 2 300; 2 600; 2 950;

2 850; 800; 300; 1 900; 2 500; 3 950;

3 250; 3 250; 1 200; 1 800; 4 800; 700;

2 450; 1 050; 3 400; 3 950; 1 200; 4 350

a) Draw up a table containing the above information and group the data in groups of 500 along with their frequency and cumulative frequency.

b) Draw a histogram of the grouped data.

c) From the histogram determine the mode of the data.

d) Draw a frequency polygon on the same set of axes as (b).

e) Draw an ogive of the data on a separate set of axes.

f) From the ogive determine:

i) The median

ii) The first quartile

iii) The third quartile.

g) Give a possible reason for the day that there were only 300 cars.

2. The tuck-shop at school decides look at the number of fizzas children at school eat everyday. Below is the table of their findings:

| ***Number of Fizzas per child*** | ***Number of Children*** |
|---|---|
| $\leq 1$ | 5 |
| $\leq 2$ | 10 |
| $\leq 3$ | 12 |
| $\leq 4$ | 14 |
| $\leq 5$ | 17 |
| $\leq 6$ | 19 |
| $\leq 7$ | 8 |
| $\leq 8$ | 6 |
| $\leq 9$ | 5 |
| $\leq 10$ | 4 |

a) Determine the modal group of the data.

b) Draw a histogram of the data.

c) Draw a frequency polygon of the data

d) Draw an ogive of the data.

e) From the ogive determine:

i) The median

ii) The first quartile

iii) The third quartile.

3. A car company interviewed 700 university students to determine the age at which they got their licences. Below are their findings:

| ***Age that Student Got Licence*** | ***Frequency*** |
|---|---|
| $18 \leq x < 20$ | 142 |
| $20 \leq x < 22$ | 282 |
| $22 \leq x < 24$ | 195 |
| $24 \leq x < 26$ | 81 |

a) Determine the modal group of the data.

b) Draw a histogram of the data

c) Draw a frequency polygon of the data.

d) Draw an ogive of the data.

e) From the ogive determine:

i) The first quartile

ii) The median

iii) The third quartile.

4. A certain school decides to find out the average age that a student from their school gets their first full-time job after leaving school. This is a summary of what they found:

| ***Average Age of First Job*** | ***Frequency*** |
|---|---|
| $\leq 16$ | 6 |
| $\leq 17$ | 74 |
| $\leq 18$ | 23 |
| $\leq 19$ | 86 |
| $\leq 20$ | 43 |
| $\leq 21$ | 78 |
| $\leq 22$ | 201 |
| $\leq 23$ | 273 |
| $\leq 24$ | 258 |
| $\leq 25$ | 96 |
| $\leq 26$ | 24 |

a) Draw a histogram of the data.

b) Draw an ogive of the data

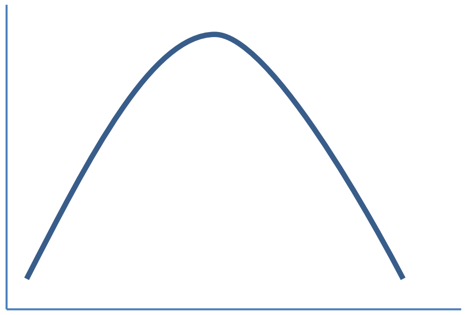
c) From the ogive determine the median, first quartile, and third quartile.

## E. The Normal Distribution and Skewness of Data

*The Normal Distribution*

A normal distribution is a frequency polygon that is symmetrical and where the mean is equal to the median which is equal to the mode.

Examples of normal distributions would be the graphs of the height of a population or the intelligence of a population and so on.



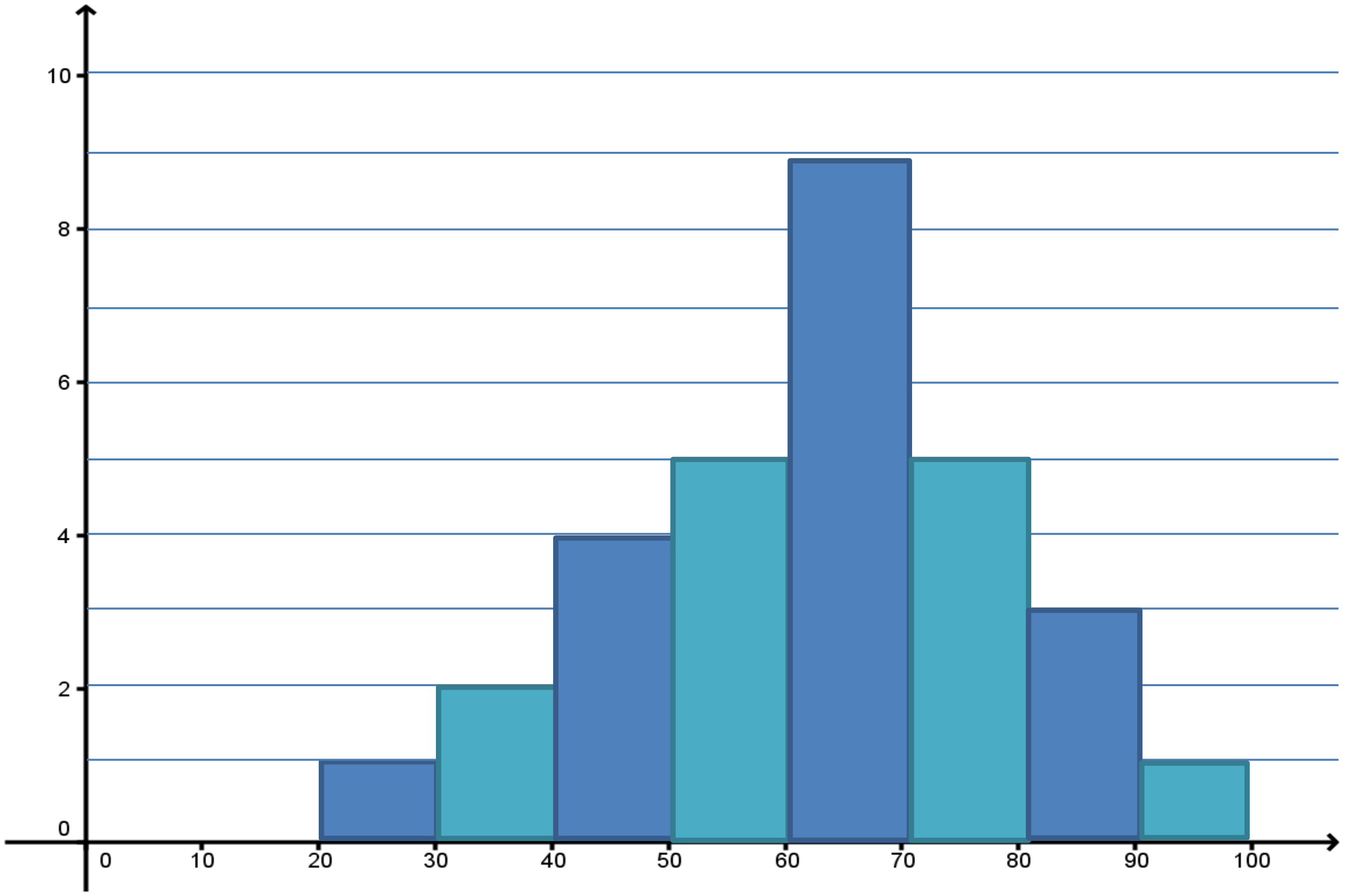
In a normal distribution approximately 67% of the data lies within the first standard deviation on either side of the mean. Approximately 95% of the data lies within two standard deviations on either side of the mean, and approximately 99% of the data lies within three standard deviations on either side of the mean.

*Skewness*

The skewness of your data tells you how your data looks and which side of the graph your data leans towards.

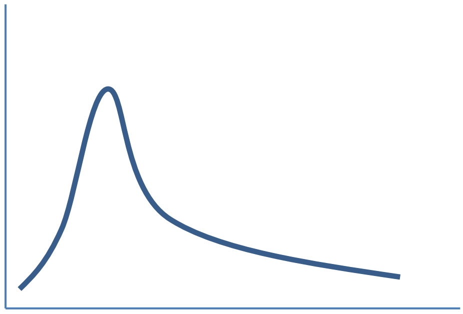
In a normal distribution, your data is unimodal – i.e. it has only one mode or one peak. Sometimes you will get a bimodal distribution which will have two modes or two peaks.

If your frequency polygon or histogram is symmetrical it means that the peak is approximately in the middle of the graph and the two ends look approximately the same. On the following page is an example:

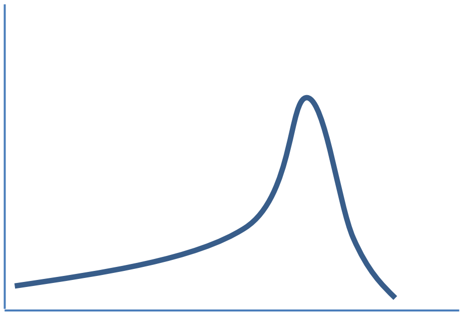


If the mode or tallest column is towards the left then the data tails off to the right and is said to be skewed to the right or positively skewed.

If the mode or tallest column is towards the right then the data tails off to the left and is said to be skewed to the left or negatively skewed.



Skewed to the right
Positively skewed



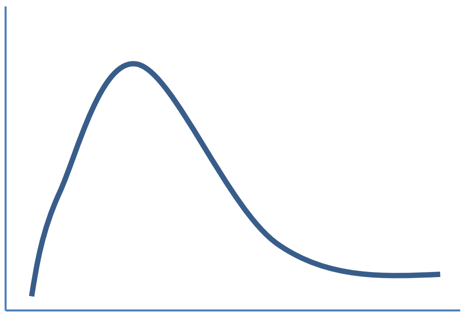
Skewed to the left
Negatively skewed

In positively skewed data the mode <median < mean.
In negatively skewed data the mode > median > mean.
In symmetrical data the mean =median = mode (approximately).
You can also determine the skewness of a box-and-whisker plot by looking at how far apart the whiskers, quartiles and the median are from each other.

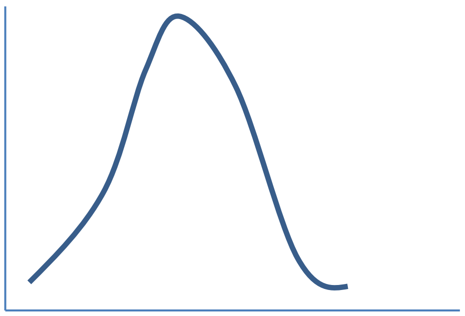
Activity 4

1. Determine the skewness of the following:

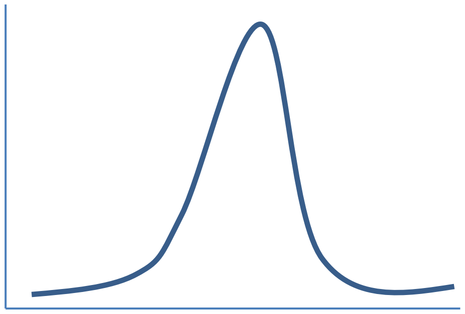
a)



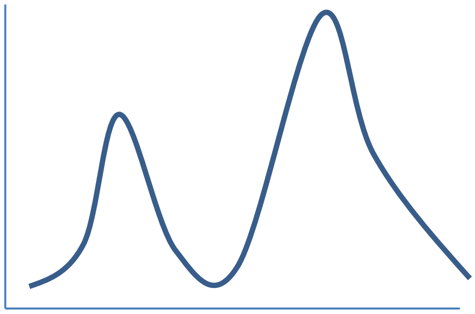
b)



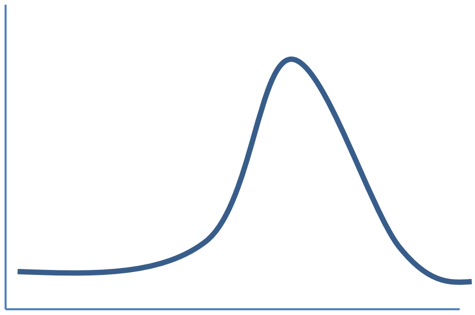
c)



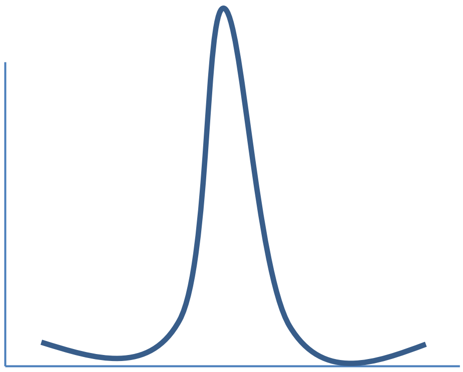
d)



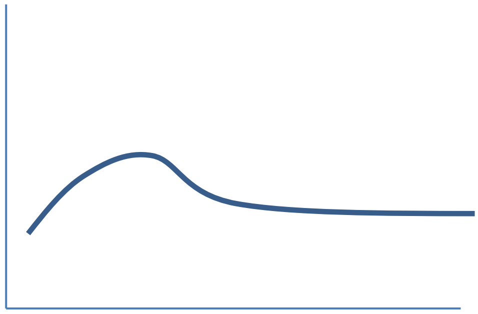
e)



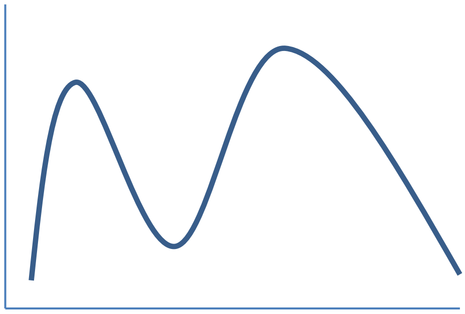
f)



g)

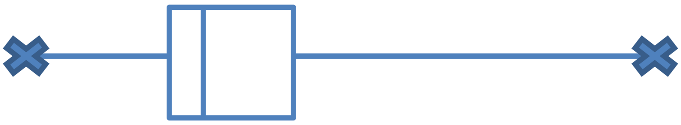


h)

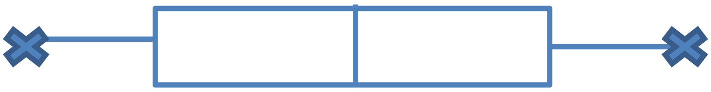


2. Look at the following box-and-whisker plots and determine their skewness.

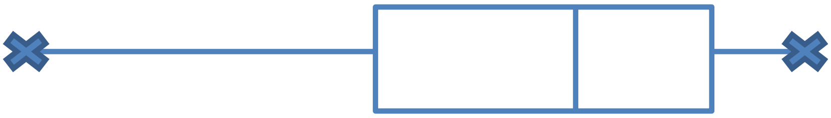
a)



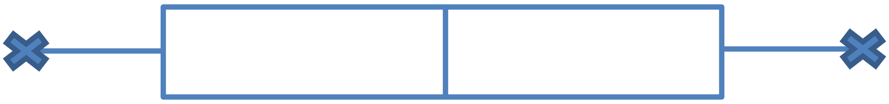
b)



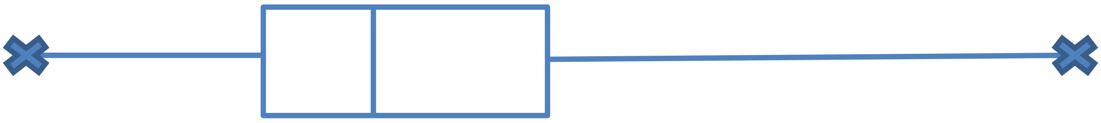
c)



d)



e)



## F. Scatter Plots, Lines of Best Fit and The Regression Line

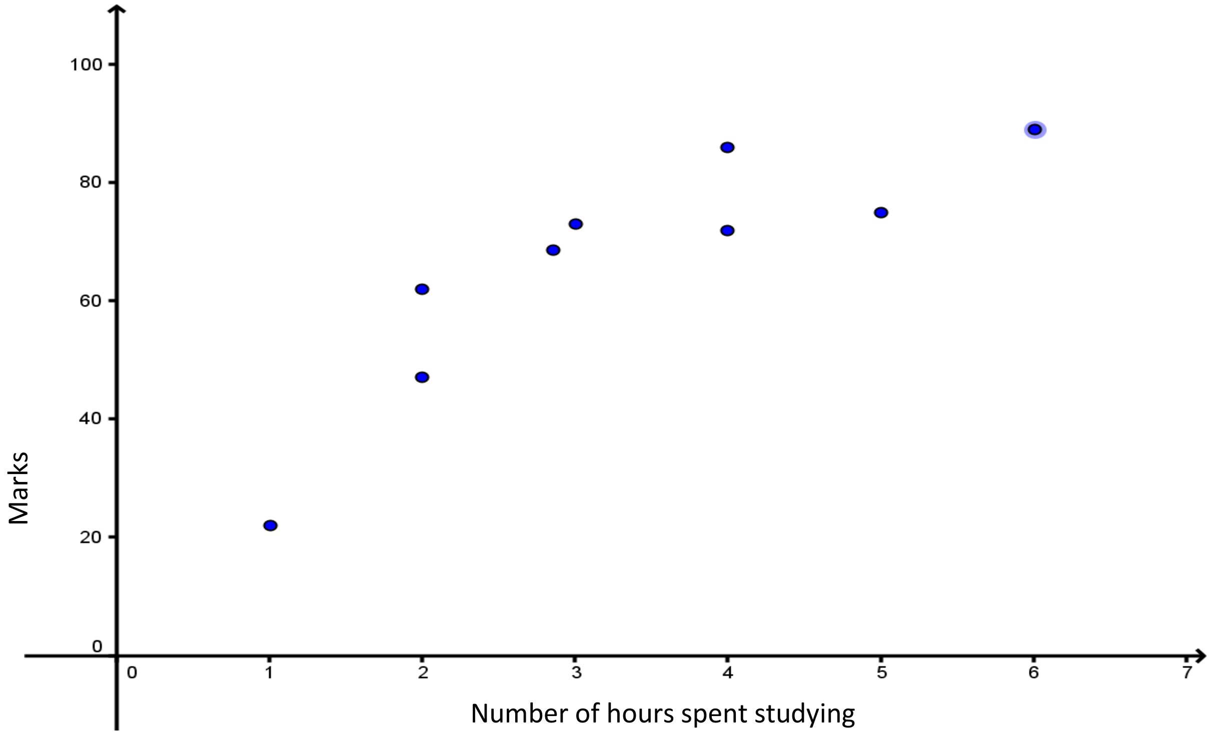
Scatter plots are graphs with two variables, $x$ and $y$, for example the number of hours spent studying vs the mark achieved for that test:

| ***Hours*** | 2 | 5 | 3 | 3 | 4 | 1 | 2 | 4 | 6 |
|---|---|---|---|---|---|---|---|---|---|
| ***Mark*** | 62 | 75 | 73 | 69 | 86 | 22 | 47 | 72 | 89 |

Now we use the number of hours as our $x$-axis because the number of hours spent studying affects the mark achieved.

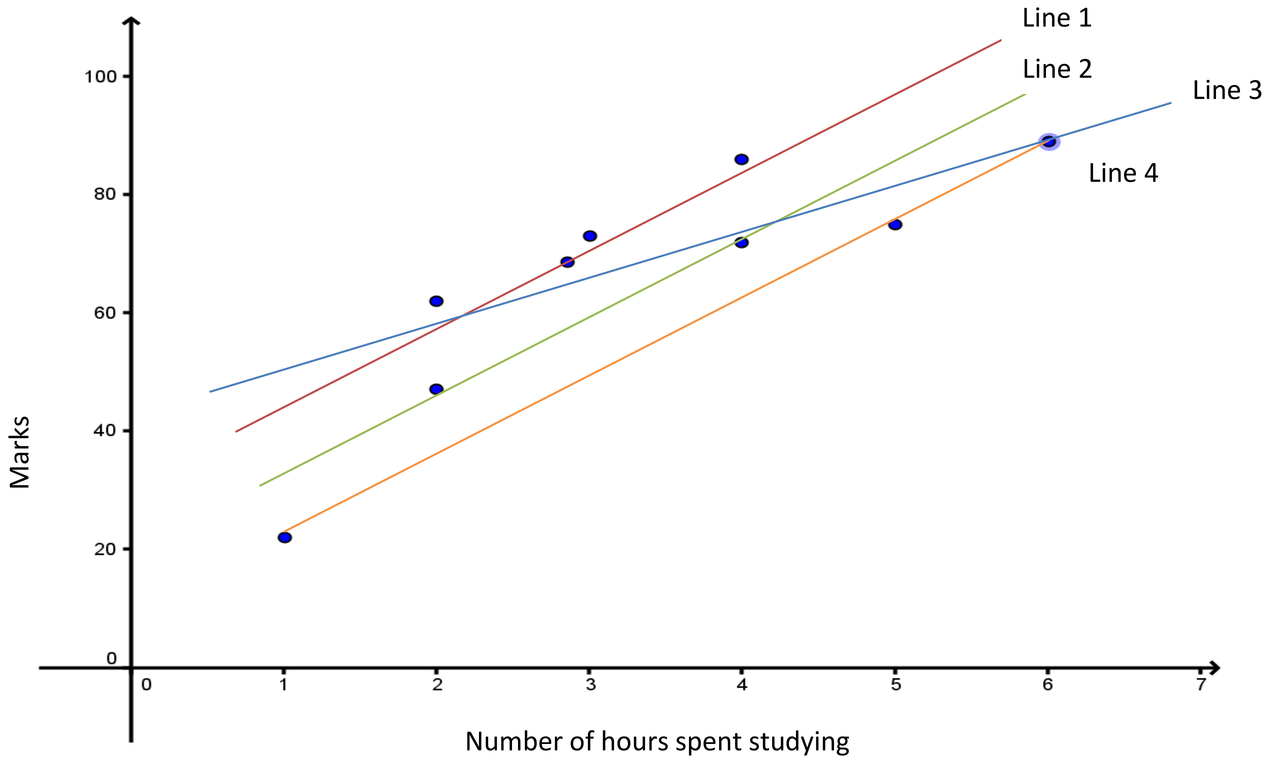
> Sometimes you cannot tell which variable affects the other variable – in which case you can choose which variable goes on which axis

And we use the marks achieved as the $y$-axis because the marks are affected by the number of hours spent studying.



From the scatter plot you can see that there is a relationship between the number of hours spent studying and the marks achieved. If you tried to draw a line through the dots or points plotted it would look like a straight line.

The line of best fit is the line that works the best with all the plotted data. For example we can see that the line that fits best with the above data is a straight line shown on the next page:



Look at the lines drawn above,

Line 1, 2 and 4 all have the same gradient. Line 4 passes through 3 points, however the other points are all above line 1. This means that the line does not represent the data accurately and so is not the line of best fit for this information. Line 1 passes through 1 point with three point above the line and very close to the line and 5 points below the line and much further from the line. So this is also not a good representation of the data. Line 2 passes through 2 points and has 4 points above the line and 3 points below the line. This is a much better representation of the information given. Line 3 has a different gradient to the other lines and only passes through 1 point, however there are 4 points on either side of the line that are approximately equal distances away from the line. This means that Line 3 fits the data best.

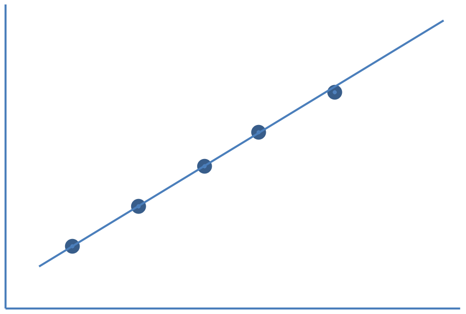
Remember that data can also have a line that is exponential in shape, logarithmic, parabolic or even hyperbolic in shape. Be aware of this for when your teacher asks you what kind of shape or line best fits the given data. Also be aware that you need to be able to interpret the shape.

*Relationships of data*

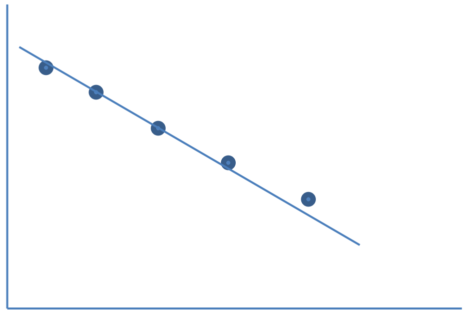
Look at the scatter plot above – the gradient is positive. This means that there is a positive relationship between the variables.
If the gradient is negative there is a negative relationship between the variables.

*Strength*

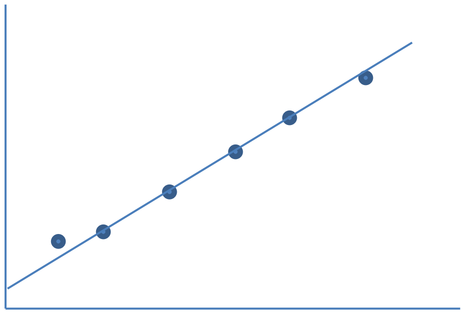
When the line of best fit lies perfectly on all the plotted points it means that the relationship is ideal or perfect – the one variable directly affects the other variable in proportion.
When the points are spread a little bit from the line, or not all points lie on the line the relationship is less strong.
When the points are spread very far away from the line it means that the relationship is very weak.
If you cannot draw a line of best fit it means that there is no relationship between the variable.



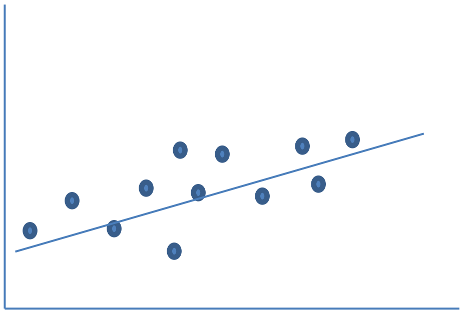
Positive Relationship



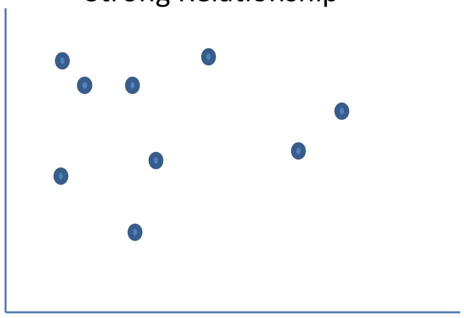
Negative Relationship



Strong Relationship



Weak Relationship



No Relationship

Activity 5

1. A driving company does a survey on the number of years spent driving vs the number of accidents occurring in the last year for that driver.

| ***Number of years driving*** | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|
| ***Number of accidents*** | 17 | 15 | 12 | 7 | 10 | 4 | 5 | 3 | 1 | 1 |

a) Draw a scatter plot of the above information.
b) What kind of line best fits this data?
c) Draw the line of best fit onto the scatter plot.
d) What kind of relationship does this data have?

2. Die-gogo uses a particular pesticide. They test the pesticides on a bug to determine the resistance that a bug would build up to the pesticide over time.

| ***Number of Sprays*** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| ***Number of Bugs still alive*** | 70 | 42 | 26 | 2 | 8 | 14 | 46 | 53 | 56 |

a) Draw a scatter plot of the above information.
b) What kind of line best fits this data?
c) From the graph, determine the optimum number of sprays.

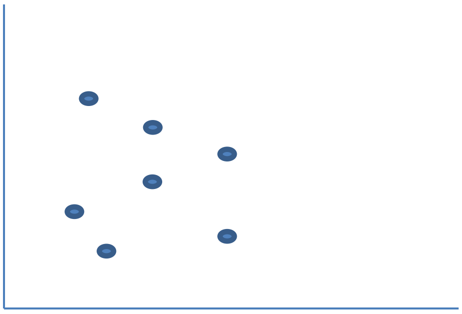
3. A scientist wants to try to determine whether there is a relationship between the heat of the day and how many murders are committed that day.

| ***Temperature*** | 25 | 27 | 23 | 32 | 35 | 19 | 28 | 29 | 26 |
|---|---|---|---|---|---|---|---|---|---|
| ***Number of Murders*** | 58 | 7 | 21 | 51 | 24 | 14 | 36 | 22 | 13 |

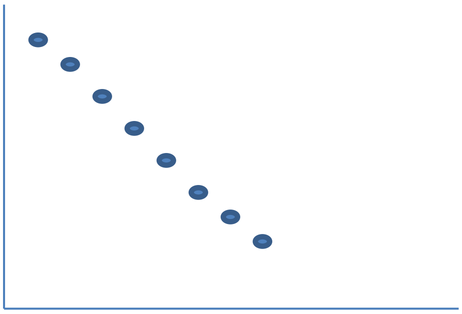
a) Draw a scatter plot of the above information.
b) Is there a relationship between the temperature and the number of murders per day?

4. State whether the following have a negative or positive relationship, and whether it is a strong or weak relationship or no relationship at all.

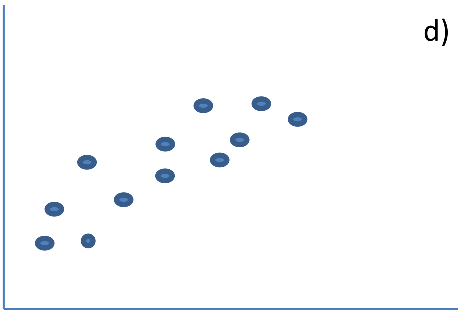
a)



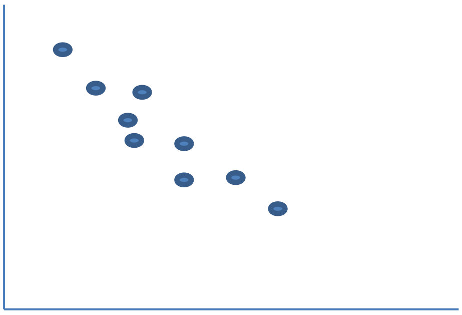
b)



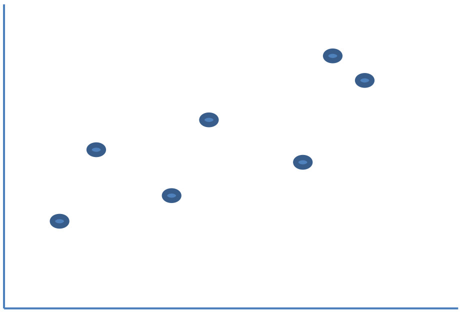
c)



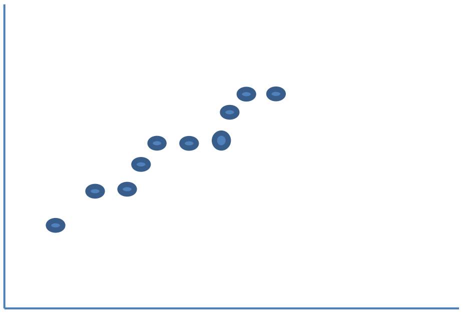
d)



e)



f)



*Regression Lines*

A regression line is calculated line of best fit. It is given by the formula:

$$y = bx + a$$

Where *b* is the gradient of the line and *"a"* is the y-intercept.

From the previous example:

| ***Hours*** | 2 | 5 | 3 | 3 | 4 | 1 | 2 | 4 | 6 |
|---|---|---|---|---|---|---|---|---|---|
| ***Mark*** | 62 | 75 | 73 | 69 | 86 | 22 | 47 | 72 | 89 |

It is very difficult to calculate the regression line because it measures the distance from the line of each point so that those distances cancel each other out, so we can use a calculator for this section.

Press: MODE 1 and then press 1 for Line data.

To enter data put the $x$-variable in first and then press (x,y) and the put in the $y$-variable and then press CHANGE.

From our example you would press: 2 (x,y) 62 CHANGE then the next set: 5 (x,y) 75 CHANGE and so on until all the data is entered.

Clear your screen first by pressing ON/C

To find the gradient, *b*, press ALPHA ) =

and you should get the answer 11.183

To find the $y$-intercept, *a*, press ALPHA ( =

And you should get the answer 28.83

So your regression line equation should look like this: $y = 11.183x + 28.83$

## Activity 6

1. From activity 5 determine the regression line for question number 1.

2. A chef wants to determine the type of relationship there is between the number of customers at his restaurant and the number of eggs he uses on that day. Below is a table of his information.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ***Number of Customers*** | 94 | 29 | 49 | 64 | 69 | 44 | 98 |
| ***Number of Eggs*** | 384 | 66 | 126 | 276 | 192 | 132 | 390 |

a) Draw a scatter plot of the data.
b) What kind of relationship does the data show?
c) Determine the regression line for the data.

3. Zookeepers around the world tried to determine the number of bananas a troupe of monkeys would eat against the number of monkeys in the troupe.
Below is the data they gathered:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ***Number of Monkeys*** | 6 | 7 | 12 | 3 | 5 | 1 | 7 | 8 | 2 |
| ***Number of Bananas*** | 75 | 53 | 175 | 22 | 38 | 21 | 84 | 96 | 18 |

a) Draw a scatter plot of the above data.
b) What kind of relationship does this data have?
c) Determine the regression line for this data.

4. Andy wanted to look at the relationship between the increase in fuel prices and the amount of kilometres he travelled in a month.

This is the data he collected:

| **Month** | ***Jan 2011*** | ***Feb 2011*** | ***Mar 2011*** | ***Apr 2011*** | ***May 2011*** | ***June 2011*** | ***July 2011*** | ***Aug 2011*** | ***Sept 2011*** |
|---|---|---|---|---|---|---|---|---|---|
| **Fuel Price** | 8,58 | 8,84 | 9,27 | 9,80 | 10,09 | 10,07 | 9,74 | 9,91 | 10,00 |
| **No. Of Km** | 405 | 384 | 348 | 336 | 312 | 306 | 304 | 300 | 300 |
| **Month** | ***Oct 2011*** | ***Nov 2011*** | ***Dec 2011*** | ***Jan 2012*** | ***Feb 2012*** | ***Mar 2012*** | ***Apr 2012*** | | |
| **Fuel Price** | 10,37 | 10,60 | 10,49 | 10,43 | 10,77 | 11,05 | 11,77 | | |
| **No. Of Km** | 252 | 240 | 228 | 204 | 192 | 116 | 64 | | |

Data from: http://www.aa.co.za/content/59/fuel-pricing/

a) Draw a scatter plot of the above data.

b) According to the line of best fit what kind of relationship does this data have?

c) Determine the regression line.

d) Find the correlation coefficient.

## G. Misleading Statistics

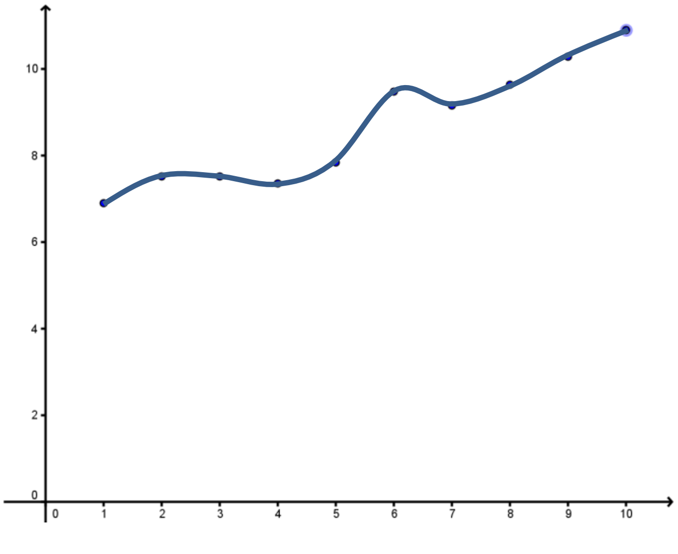
Look at the following questions:

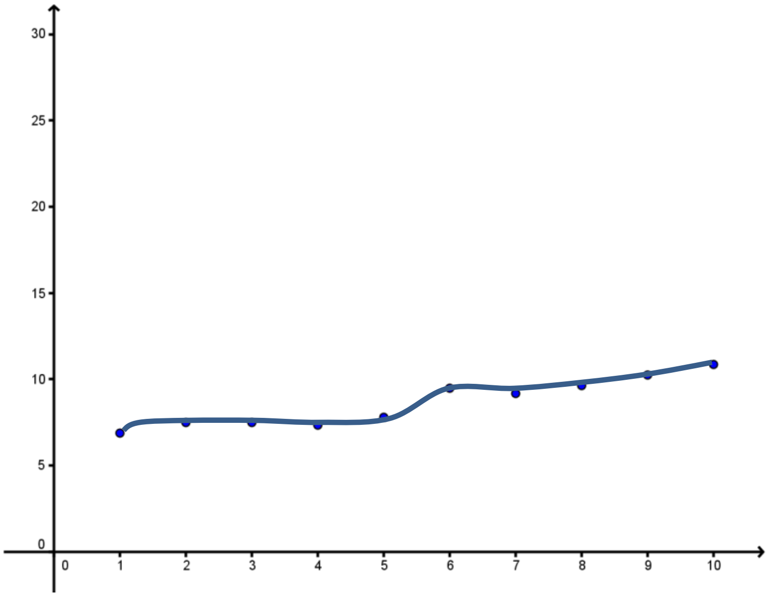
- Do you like hot chocolate?
- What kind of hot drink do you like?

The first question only allows a yes or no answer while the second question allows a person to express their own opinion.

The first question can be used to present skewed information, whereas the second question will give a more accurate idea of what hot drinks are popular.

This same technique can be used with graphs as well:
Look at the following two graphs that represent the same information – the price of diesel over the last 2 years.





If you were a travel company and you wanted to show the dramatic increase in fuel prices which graph would you use?
If you were the government and wanted to show that diesel prices had not increased very much which graph would you use?

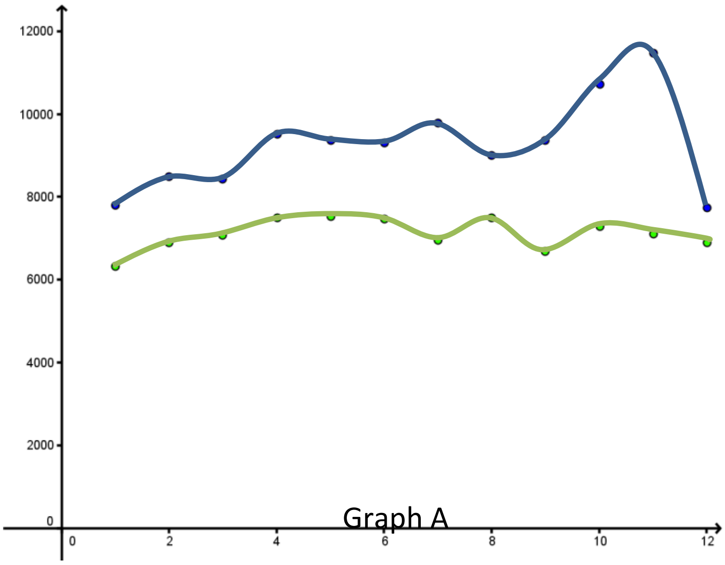
The only difference between these two graphs is the scale they are drawn with. The first has a scale of 0 to 10 and goes up in two's while the second graph has a scale that goes from 0 to 30 and goes up in five's. Scale makes a big difference.
You need to look out for a couple of things:

- Who does the research / statistics benefit?
- What is being compared? And why are they comparing it?

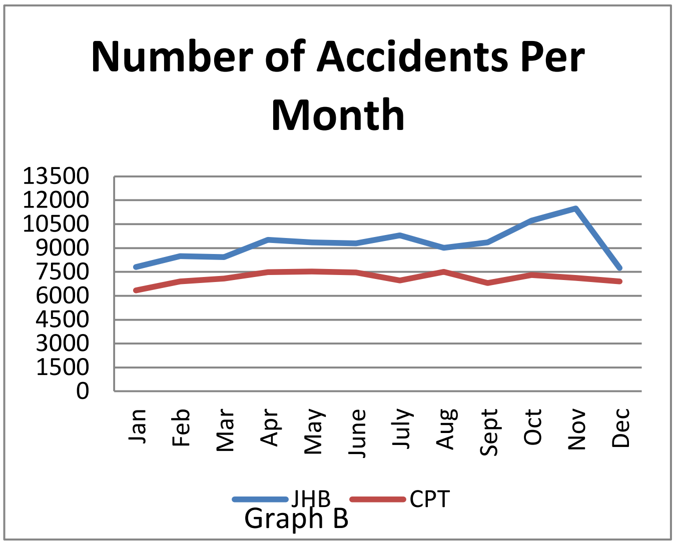
- Look at the size of the sample → statistics for a small sample is much easier to manipulate simply by changing one or two values.
- Look at the scale of the graph – think about why they would use that scale.

Activity 7

1. Look at the graphs below with regards to the number of accidents per month in a certain area which represent the same data:



Graph A



Graph B

a) Which graph in your opinion gives more information and why?

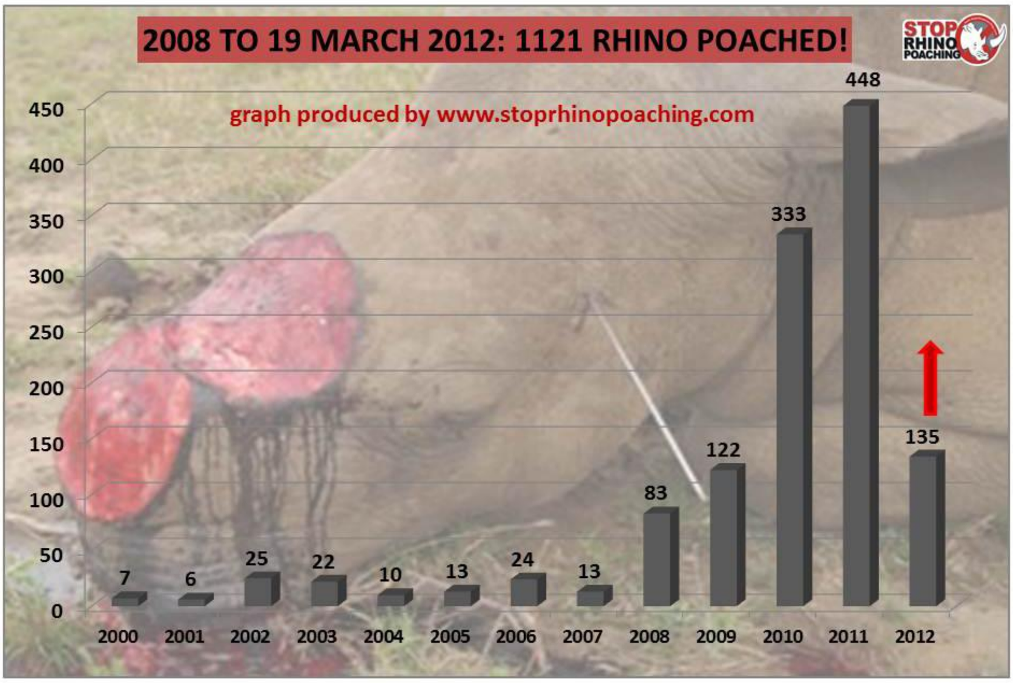
b) Which graph is easier to read values from?

c) If someone wanted to say that Johannesburg is much less safe on the road than Cape Town which graph would they use and why?

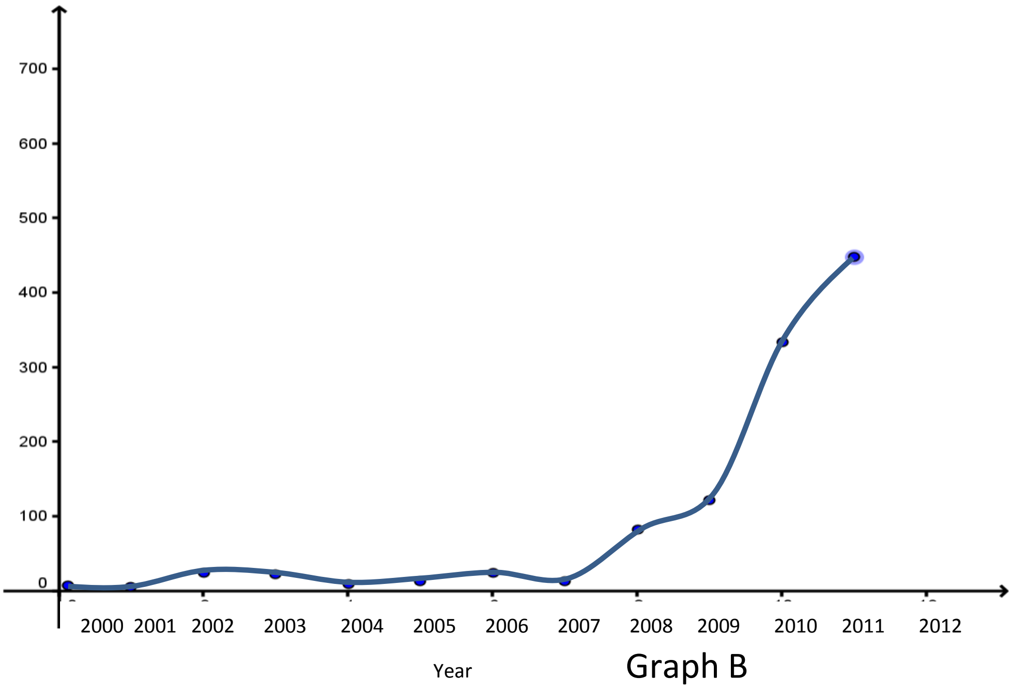
d) If someone wanted to say that there wasn't much difference between the number of accidents in Cape Town and Johannesburg which graph would they use and why?

e) What is the difference between the two graphs and why does it make such a difference.

2. Look at the following graphs about rhino poaching that use the same data carefully:



Graph A



Graph B

a) Which graph in your opinion more effectively conveys the plight of the rhinos?

b) Looking at both graphs, what general trend is occurring with regards to the number of deaths of the rhino per year?

c) Which graph would you use if you were trying to raise funds to raise awareness about the rhino's plight?

# Answers to Activities:

## Activity 1:

1. a) $\bar{x} = \frac{\sum x}{n}$

$\bar{x} = \frac{44\ 520}{30}$

$\bar{x} = 1\ 484$

b) To find the median, first arrange the data in ascending order:

| | | | | | |
|---|---|---|---|---|---|
| 780 | 840 | 880 | 900 | 1 080 | 1 080 |
| 1 120 | 1 140 | 1 220 | 1 400 | 1 480 | 1 500 |
| 1 520 | 1 560 | 1 600 | 1 600 | 1 640 | 1 640 |
| 1 680 | 1 720 | 1 740 | 1 740 | 1 760 | 1 800 |
| 1 800 | 1 800 | 1 840 | 1 860 | 1 900 | 1 900 |

$\text{Median Position} = \frac{n+1}{2}$

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

$= 15{,}5$

$\therefore \text{Median} = \frac{1\ 600+1\ 600}{2}$

$= 1\ 600$

c) 1 800

d) $\text{Quartile 1 position} = \frac{n+1}{4}$

$= \frac{30+1}{4}$

$= 7{,}75$

$\therefore \text{Quartile 1} = \frac{1\ 120+1\ 140}{2}$

$= 1\ 130$

$\text{Quartile 3 position} = \frac{3(n+1)}{4}$

$= \frac{3(30+1)}{4}$

$= 23{,}25$

$\therefore \text{Quartile 3} = \frac{1\ 760+1\ 800}{2}$

$= 1\ 780$

e) Range = Maximum – minimum

= 1 900 – 780

= 1 120

Interquartile Range = $Q_3 - Q_1$

= 1 780 – 1 130

= 650

f) Note: You won’t be expected to use the long method for more than 15 data values.

$\therefore\ \sigma = 344,88$

g) No. of students that earn less than R1000 = 4

∴ Percentage of students that earn less than R1000 = $\frac{4}{30} \times 100 = 13{,}3\dot{3}\%$

h) Top 10% of students = 3

∴ the top 3 wages are 1 860, 1 900, and 1 900.

2. a) $\bar{x} = \frac{\sum x}{n}$

$\bar{x} = \frac{67{,}98}{20}$

$\bar{x} = 3{,}399$

b) To find the median, we first need to arrange the data in ascending order:

1,00 1,32 1,78 2,02 2,63 2,71
2,97 3,21 3,70 3,78 3,79 3,79
3,84 3,92 4,06 4,11 4,33 4,57
5,11 5,32

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

$= \frac{20+1}{2}$

$= 10{,}5$

∴ Median $= \frac{3{,}78+3{,}79}{2}$

$= 3{,}785$

c) mode = 3,79

d) Quartile 1 position $= \frac{n+1}{4}$

$= \frac{20+1}{4}$

$= 5{,}25$

∴ Quartile 1 $= \frac{2{,}63+2{,}71}{2}$

$= 2{,}67$

Quartile 3 position $= \frac{3(n+1)}{4}$

$= \frac{3(20+1)}{4}$

$= 15{,}75$

∴ Quartile 3 $= \frac{4{,}06+4{,}11}{2}$

$= 4{,}085$

e) Range = maximum – minimum

= 5,32 – 1,00

= 4,32

Interquartile Range = $Q_3 - Q_1$

= 4,085 – 2,67

= 1,415

f) $\sigma = 1{,}155$

g) no. of bacteria that grew to less than 2mm$^2$ = 3

$\therefore$ Percentage of bacteria that grew to less than 2mm$^2$ = $\frac{3}{20} \times 100 = 15\%$

h) 15% of 20 = 3

$\therefore$ the third biggest growth of bacteria = 4,57

i) New mean $= \frac{67{,}98+8{,}29+7{,}02+4{,}85+6{,}90+5{,}31}{20+5}$

$= \frac{100{,}35}{25}$

$= 4{,}014$ $\therefore$ the mean would increase

j) new $\sigma = 1{,}7$

$\therefore$ yes the standard deviation would be affected – it would increase by 0,545.

Activity 2

1. a) Before you can give a 5-number summary, you need to put the data in ascending order:

1,16 1,19 1,26 1,32 1,39 1,39
1,41 1,49 1,51 1,55 1,56 1,67
1,70 1,75 1,77 1,78 1,83 1,96

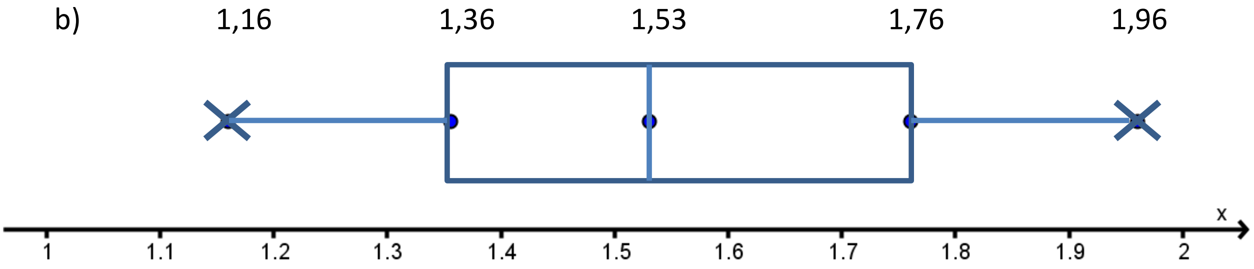
Minimum = 1,16 Quartile 1 Position $= \frac{n+1}{4} = \frac{18+1}{4} = 4{,}75$

Maximum = 1,96 $\therefore$ Quartile 1 $= \frac{1{,}32+1{,}39}{2} = 1{,}355$

Quartile 3 Position $= \frac{3(n+1)}{4} = \frac{3(18+1)}{4} = 14{,}25$ $\therefore$ Quartile 3 = $\frac{1{,}75+1{,}77}{2} = 1{,}76$

Median position $= \frac{n+1}{2} = \frac{18+1}{2} = 9{,}5$ $\therefore$ Median $= \frac{1{,}51+1{,}55}{2} = 1{,}53$

b)



c) 50%

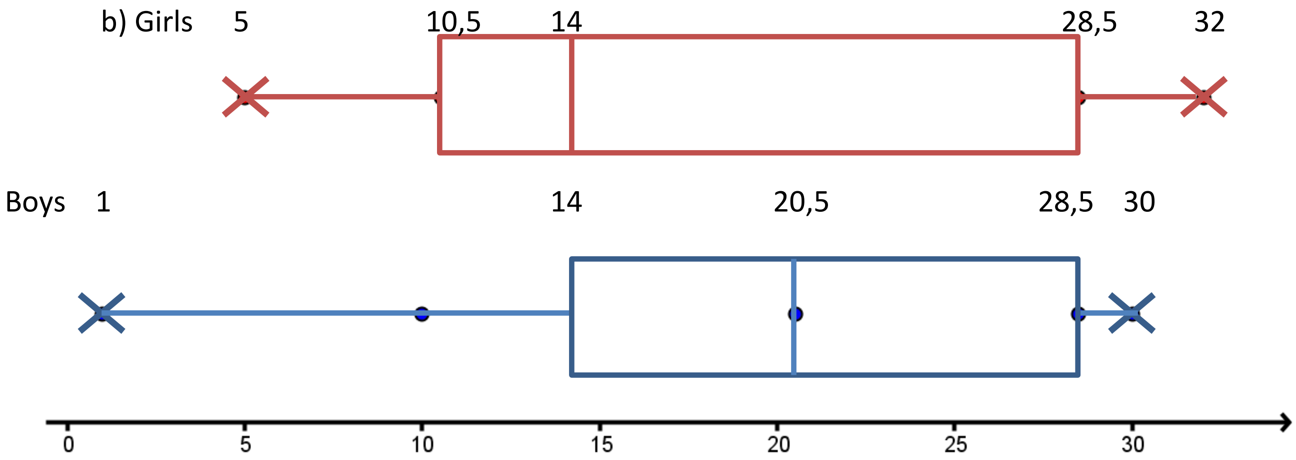
2. a) Girls: 5 9 12 13 14 14 21 28 29 30

| | | | |
|---|---|---|---|
| Minimum: | 5 | Maximum: | 32 |
| Quartile 1: | 10,5 | Quartile 3: | 28,5 |
| Median: | 14 | | |

Boys: 1 14 14 15 18 23 24 28 29 30

| | | | |
|---|---|---|---|
| Minimum: | 1 | Maximum: | 30 |
| Quartile 1: | 14 | Quartile 3: | 28,5 |
| Median: | 20,5 | | |

b)



c) Girls: The data is very spread out on the right side of the box and whisker plot, while the left side of the graph is very “clumped” together.

Boys: The data is more spread out on the left than on the right and 25% of the values fall between 28,5 and 30.

d) More girls watched less than 14 movies (50% of the girls, 50% of the boys watched less than 20,5 movies).

e) 25% of the girls watched between 14 and 28 movies.

While 50% of the boys watched between 14 and 28 movies.

Therefore there more boys watched between 14 and 28 movies.

3. a) Class 1 Range = 91 – 28 = 63

Class 2 Range = 83 – 14 = 69

Therefore Class 2 has the greater range of marks.

b) the first class did better as the lowest mark was 28% and 50% of the data falls between 37% and 77%, finally 25% of the marks fall between 77% and 91% which is better than the marks for class one whose lowest mark is 14% and 50% of the marks fall between 43% and 73%, the top 25% of the class only ranged from 74% to 83% which is much lower than the first class.

c) IQR class 1 = 77 – 37 = 40

IQR class 2 = 74 – 43 = 31

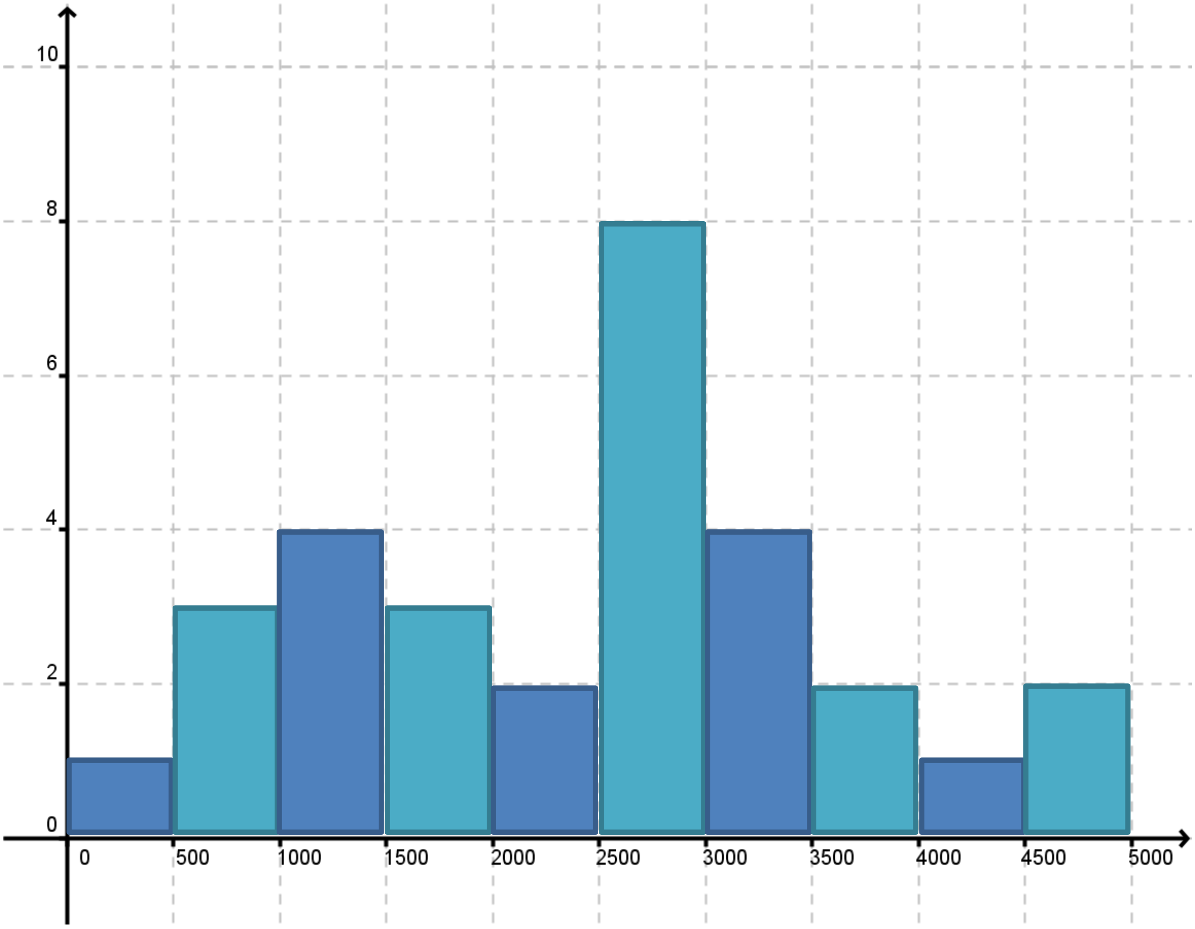
Yes – the interquartile range gives a much better indication of how close the values are to each other.

Activity 3

1. a)

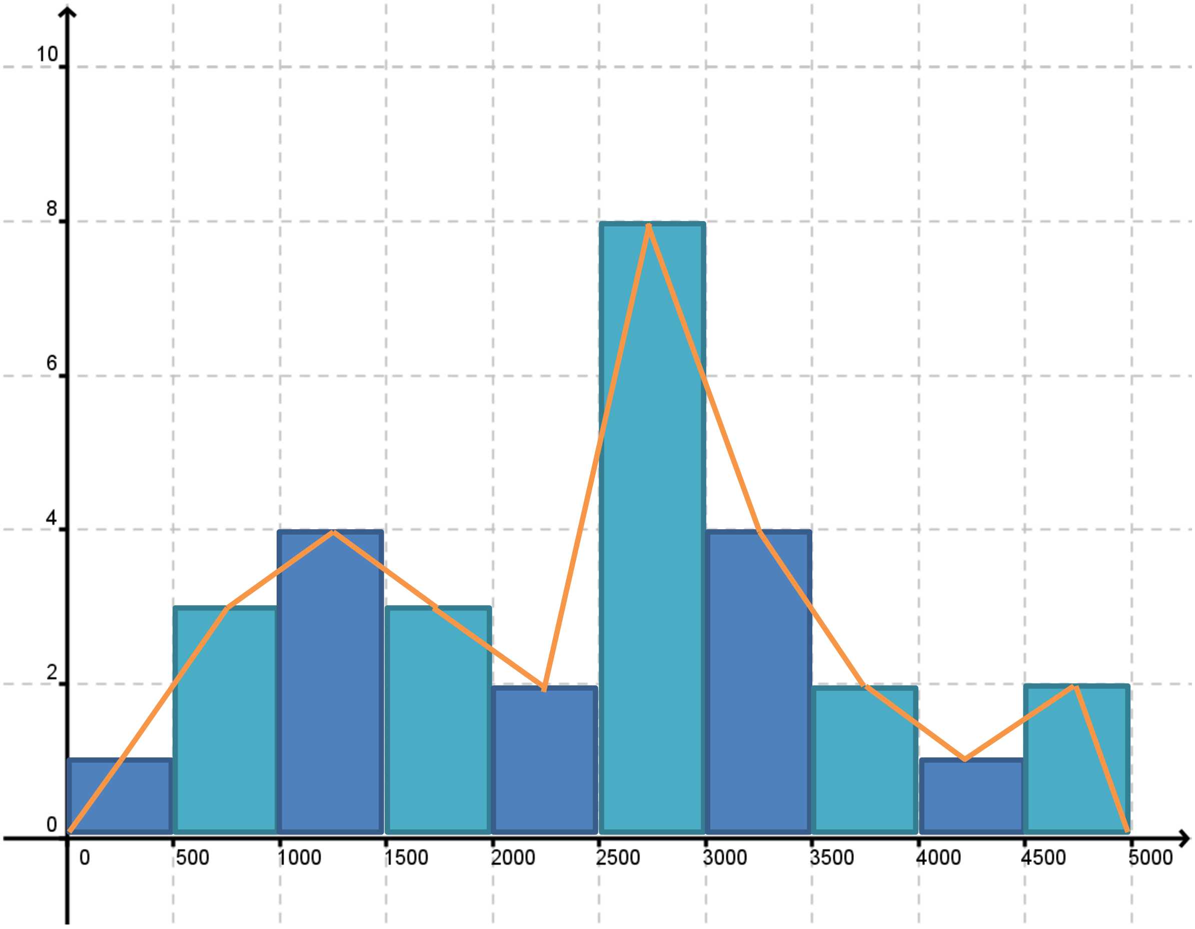
| ***Number of Cars*** | ***Frequency*** | ***Cumulative Frequency*** |
|---|---|---|
| $0 \leq cars < 500$ | 1 | 1 |
| $500 \leq cars < 1\,000$ | 3 | 4 |
| $1\,000 \leq cars < 1\,500$ | 4 | 8 |
| $1\,500 \leq cars < 2\,000$ | 3 | 11 |
| $2\,000 \leq cars < 2\,500$ | 2 | 13 |
| $2\,500 \leq cars < 3\,000$ | 8 | 21 |
| $3\,000 \leq cars < 3\,500$ | 4 | 25 |
| $3\,500 \leq cars < 4\,000$ | 2 | 27 |
| $4\,000 \leq cars < 4\,500$ | 1 | 28 |
| $4\,500 \leq cars < 5\,000$ | 2 | 30 |
| **Total** | **30** | |

b)

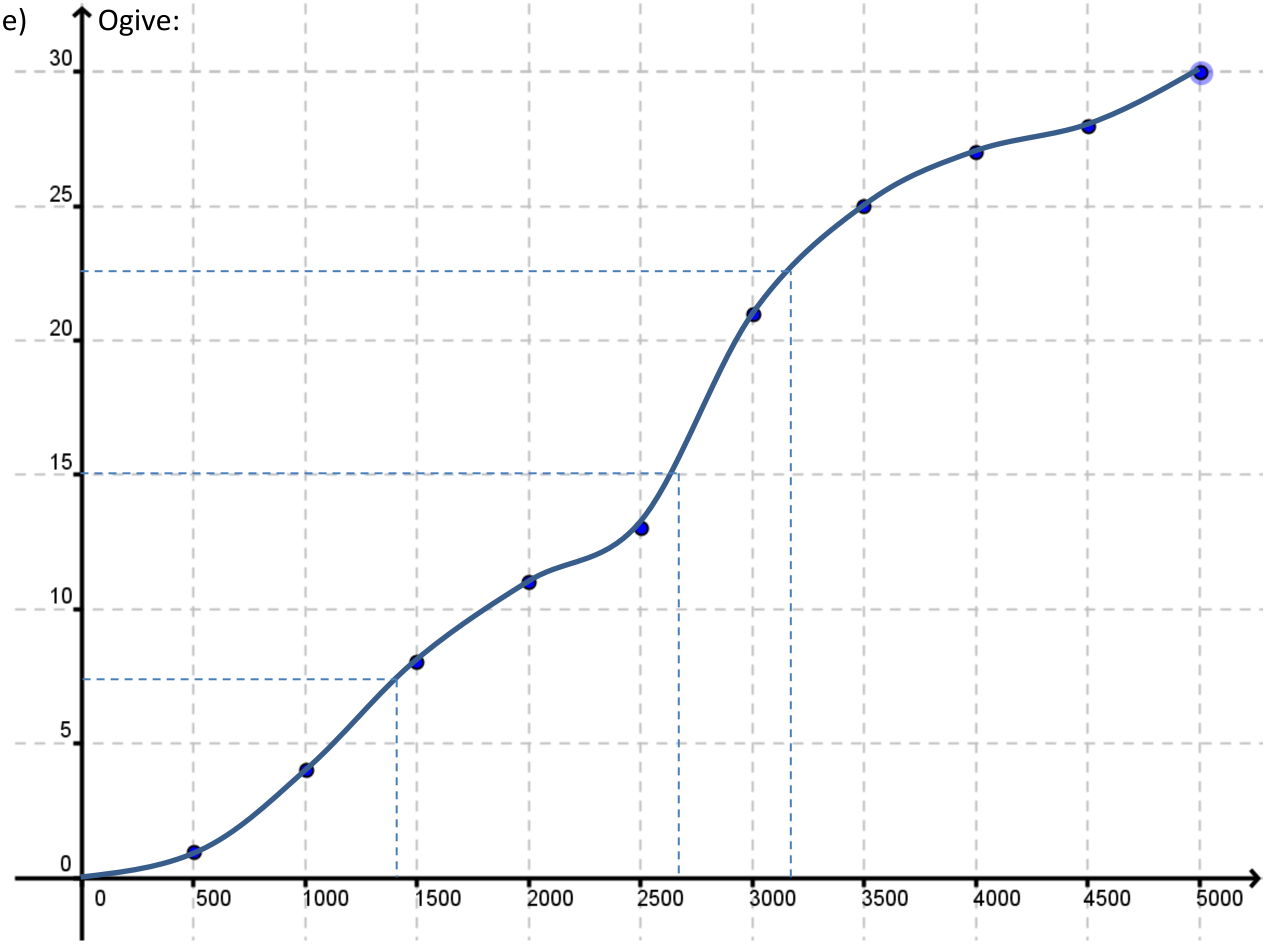


c) 2 500 – 3 000

d)



e) Ogive:



f) i) Median → 2 500 – 3 000

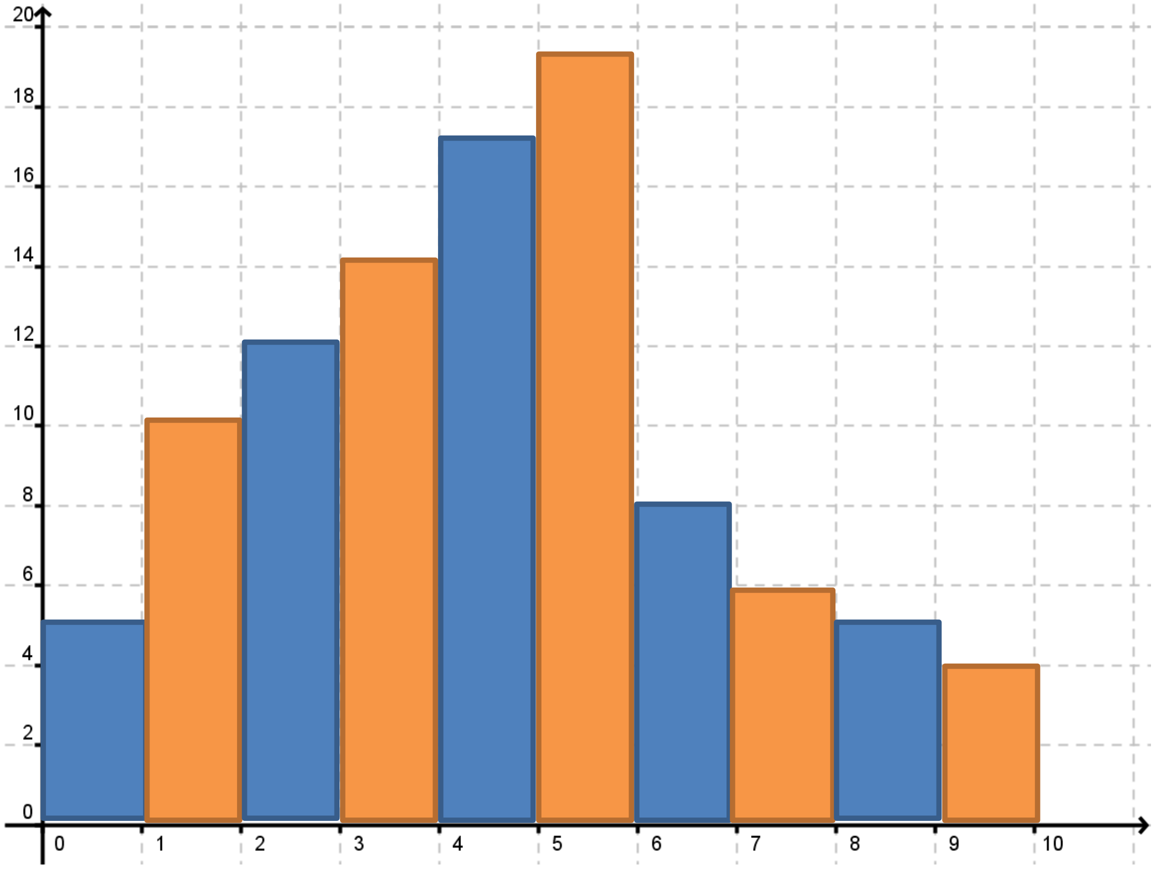
ii) Quartile 1 → 1 000 – 1 500

iii) Quartile 3 → 3 000 – 3 500

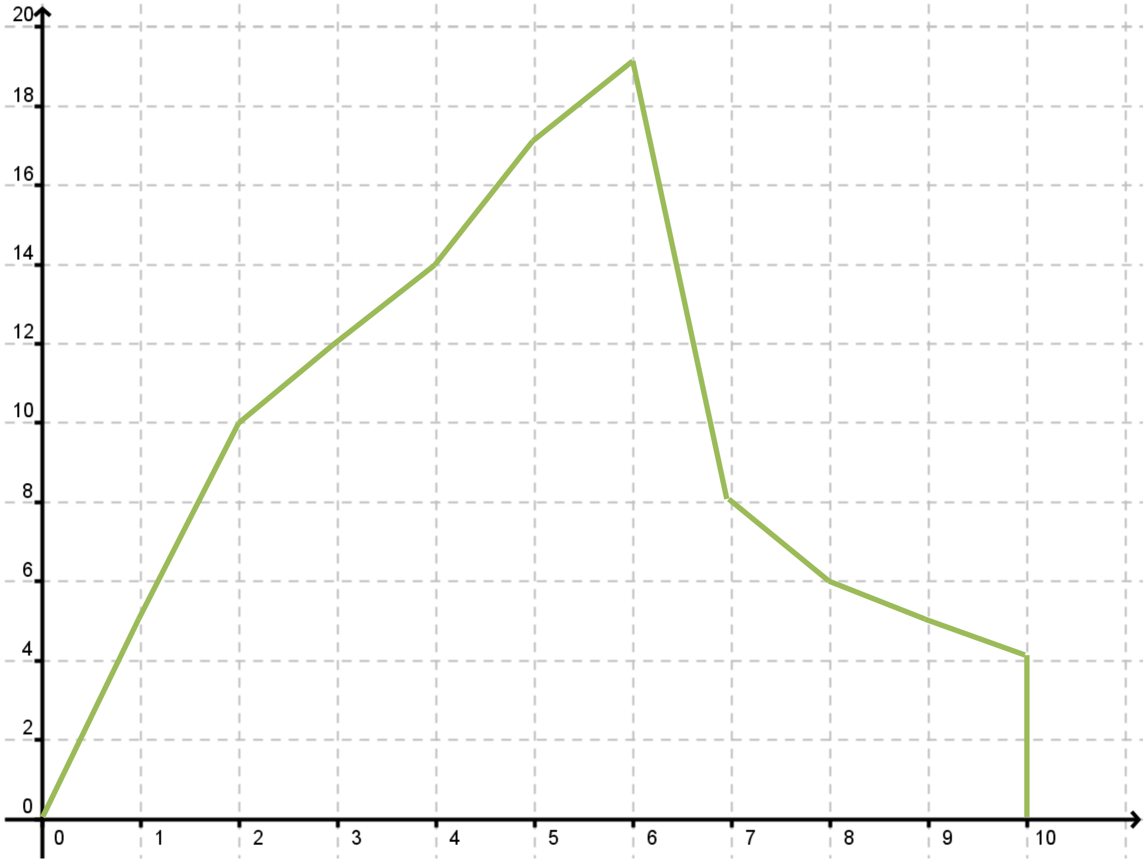
g) Perhaps the road was closed part of the day.
There was a riot
There was an accident and people avoided that road
There was a public holiday and people didn't drive.
And so… any valid reason – as long as it makes sense and is relevant to the question.

2. a) $\leq 6$

b)

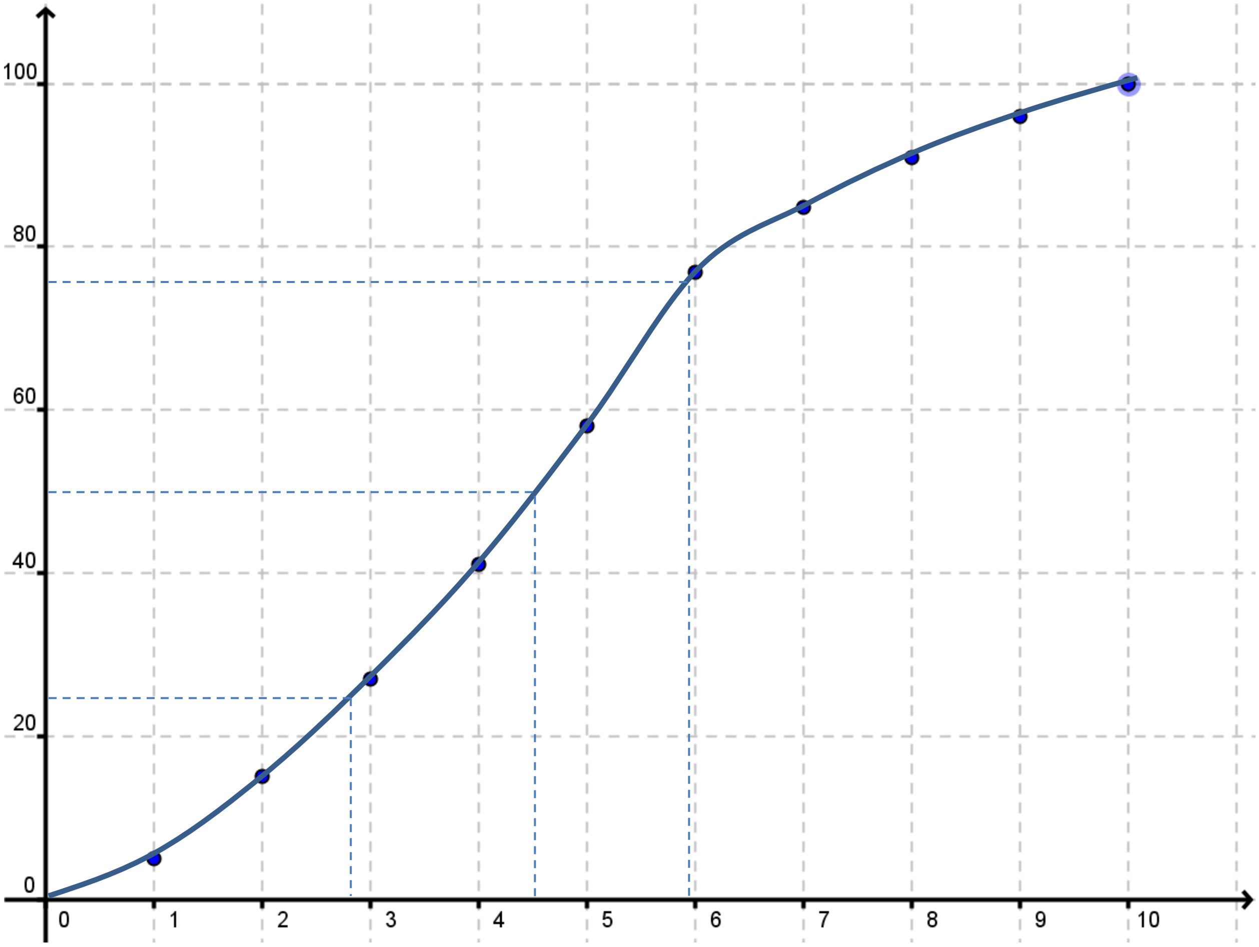


c)



d) To draw an ogive you need to find the cumulative frequency first:

| ***Number of Fizzas per child*** | ***Number of Children*** | ***Cumulative Frequency*** |
|---|---|---|
| ≤ 1 | 5 | 5 |
| ≤ 2 | 10 | 15 |
| ≤ 3 | 12 | 27 |
| ≤ 4 | 14 | 41 |
| ≤ 5 | 17 | 58 |
| ≤ 6 | 19 | 77 |
| ≤ 7 | 8 | 85 |
| ≤ 8 | 6 | 91 |
| ≤ 9 | 5 | 96 |
| ≤ 10 | 4 | 100 |



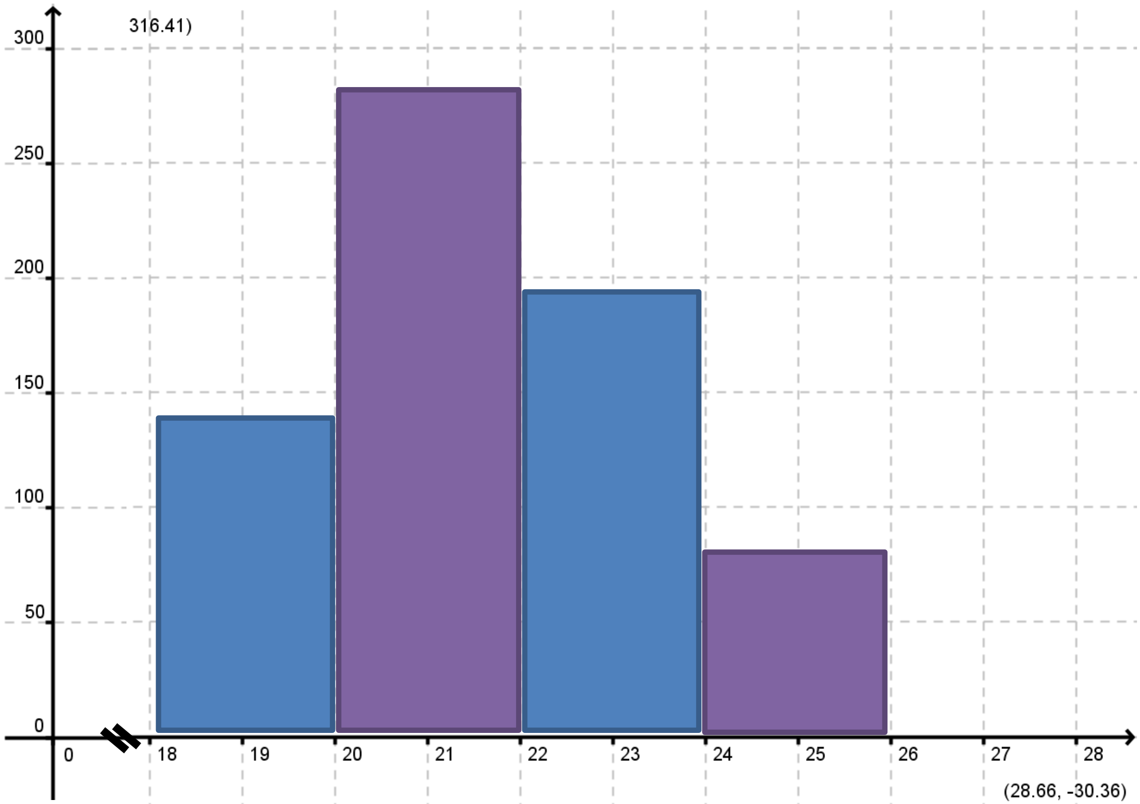
e) i) Median = $\leq 5$

ii) Quartile 1 = $\leq 3$

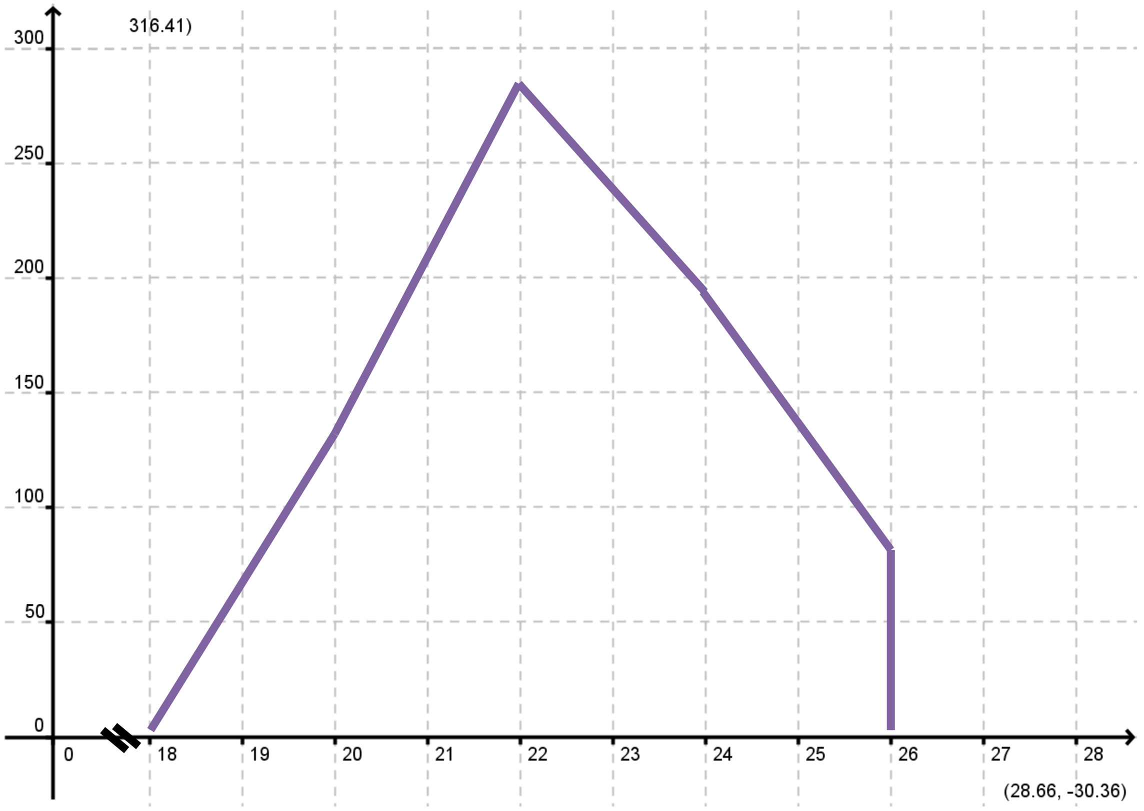
iii) Quartile 3 = $\leq 6$

3. a) $20 \leq x < 22$

b)

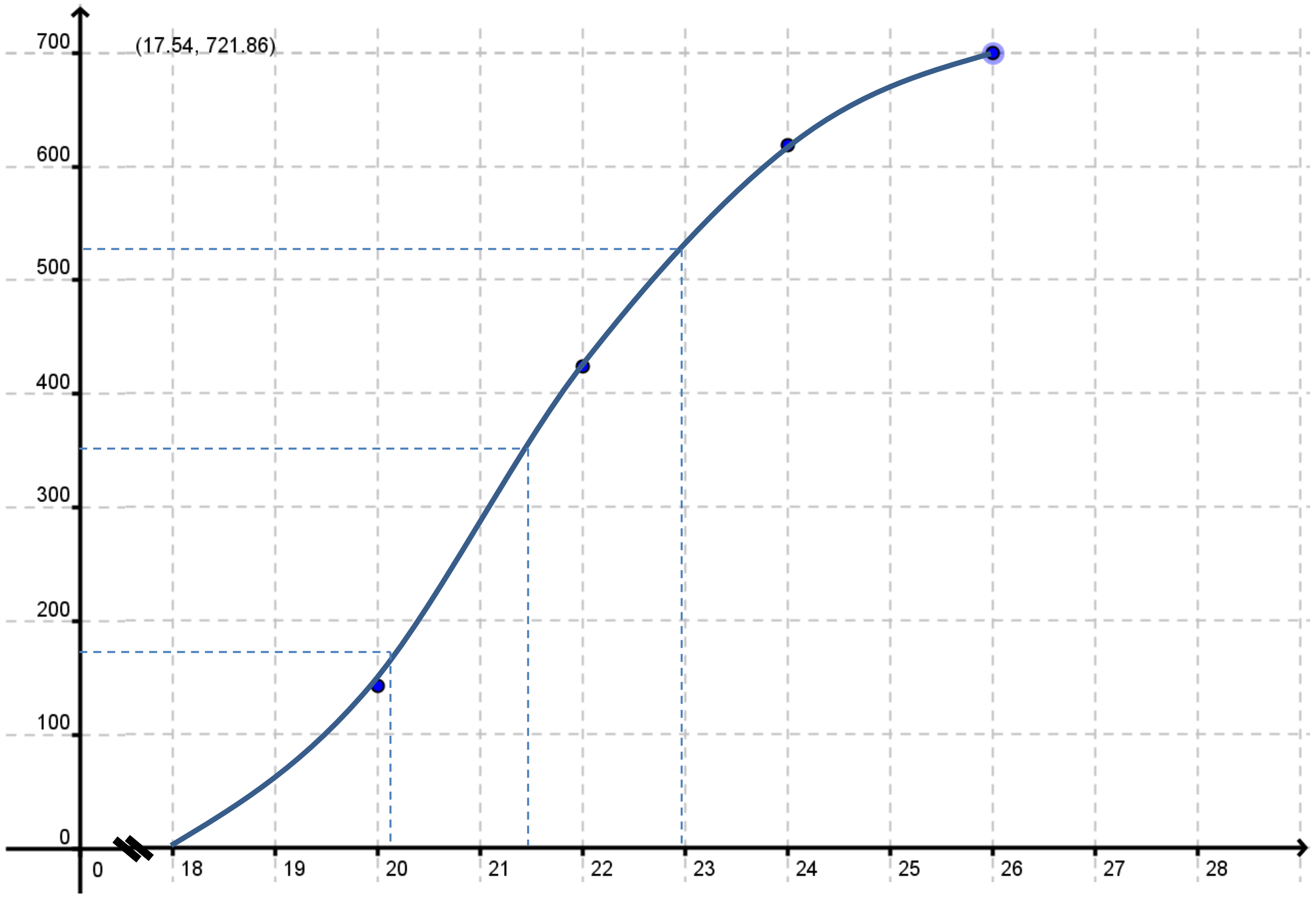


c)



d) To draw an ogive you need to find the cumulative frequency first:

| *Age that Student Got Licence* | *Frequency* | *Cumulative Frequency* |
|---|---|---|
| $18 \leq x < 20$ | 142 | 142 |
| $20 \leq x < 22$ | 282 | 424 |
| $22 \leq x < 24$ | 195 | 619 |
| $24 \leq x < 26$ | 81 | 700 |

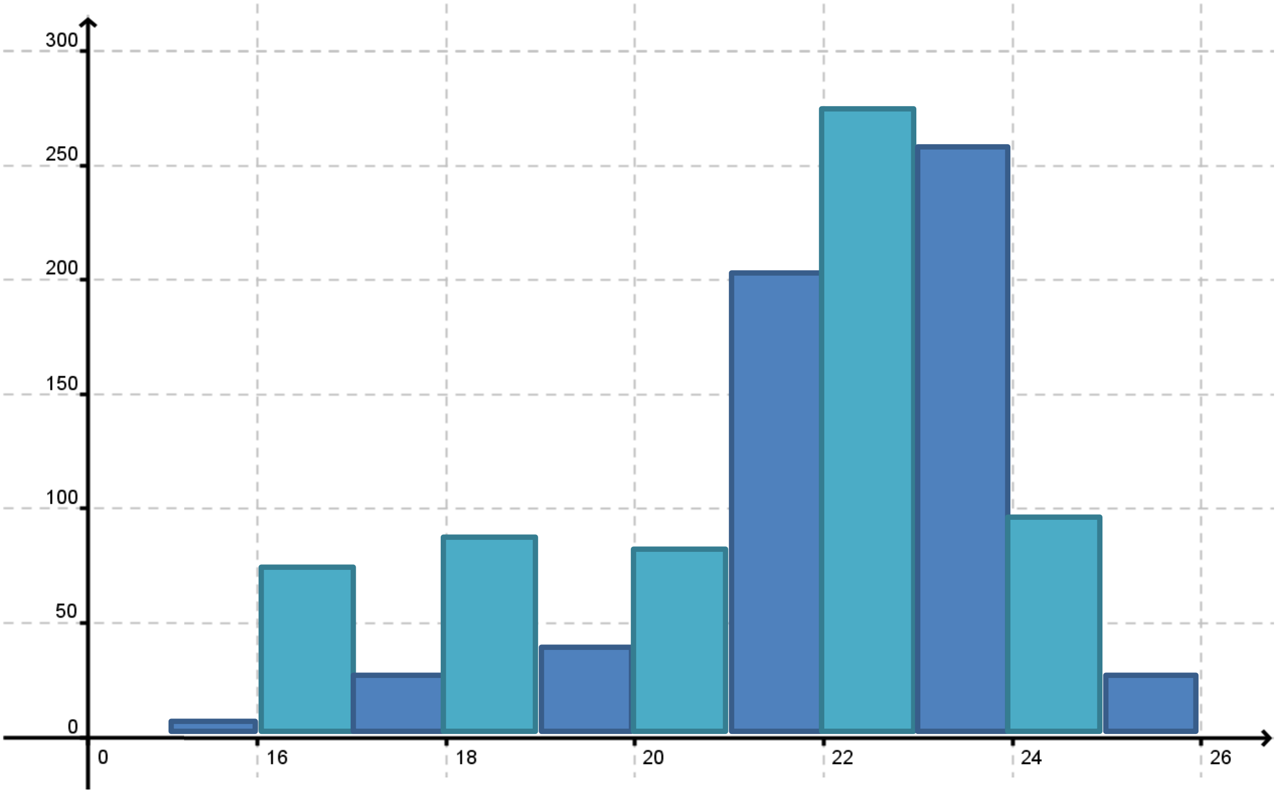


e) i) Median = $20 \leq x < 22$

ii) Quartile 1 = $20 \leq x < 22$

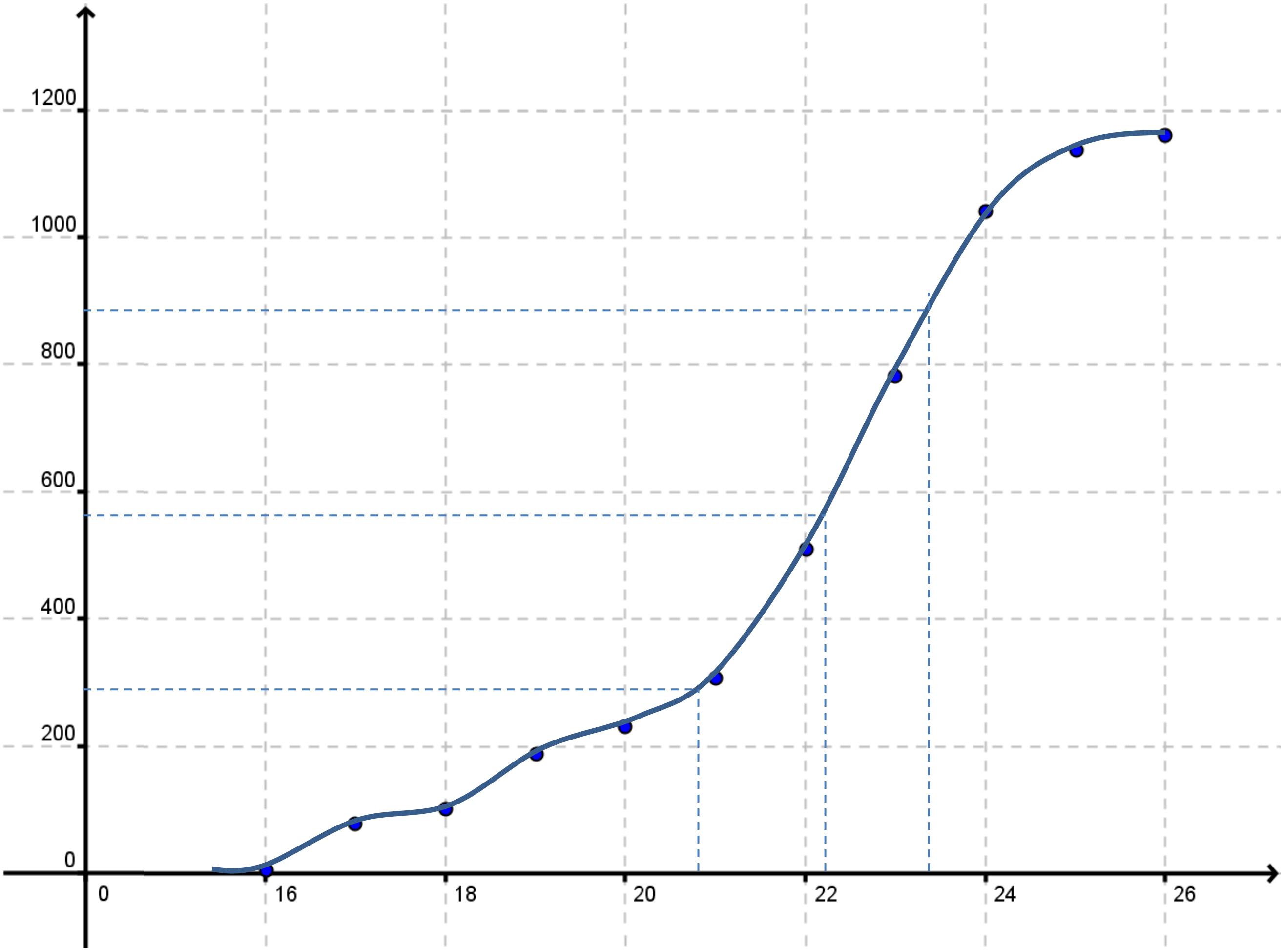
iii) Quartile 3 = $22 \leq x < 24$

4. a)



b) To draw an ogive – you need to find the cumulative frequency first:

| ***Average Age of First Job*** | ***Frequency*** | ***Cumulative Frrequency*** |
|---|---|---|
| $\leq 16$ | 6 | 6 |
| $\leq 17$ | 74 | 80 |
| $\leq 18$ | 23 | 103 |
| $\leq 19$ | 86 | 189 |
| $\leq 20$ | 43 | 232 |
| $\leq 21$ | 78 | 310 |
| $\leq 22$ | 201 | 511 |
| $\leq 23$ | 273 | 784 |
| $\leq 24$ | 258 | 1 042 |
| $\leq 25$ | 96 | 1 138 |
| $\leq 26$ | 24 | 1 162 |



c) Median = ≤ 23
Quartile 1 = ≤ 21
Quartile 3 = ≤ 24

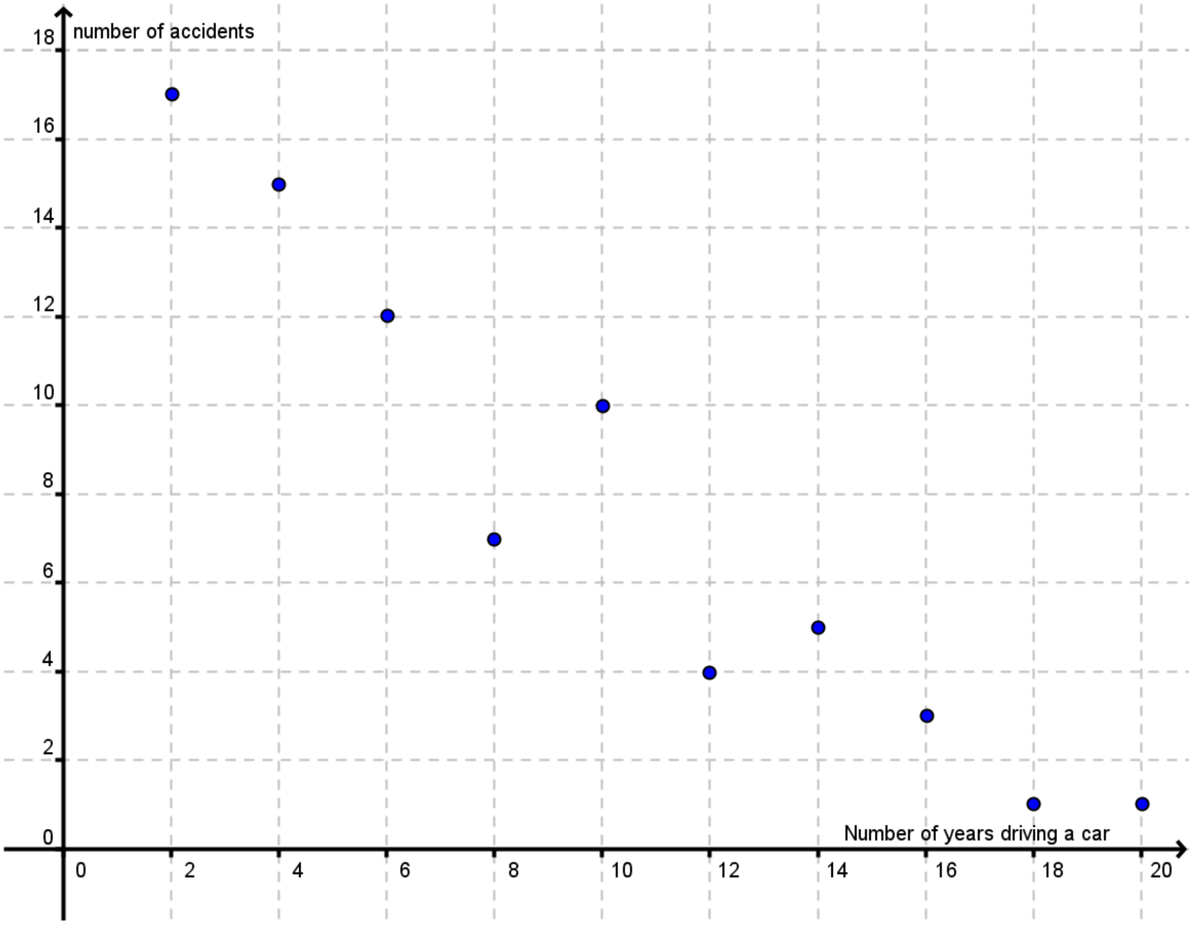
Activity 4

1. a) Skewed to the right (positively skewed) b) symmetrical
c) symmetrical d) bimodal
e) skewed to the left (negatively skewed) f) symmetrical
g) skewed to the right (positively skewed) h) bimodal

2. a) skewed to the right (positively skewed) b) symmetrical
c) skewed to the left (negatively skewed) d) symmetrical
e) skewed to the right (positively skewed)

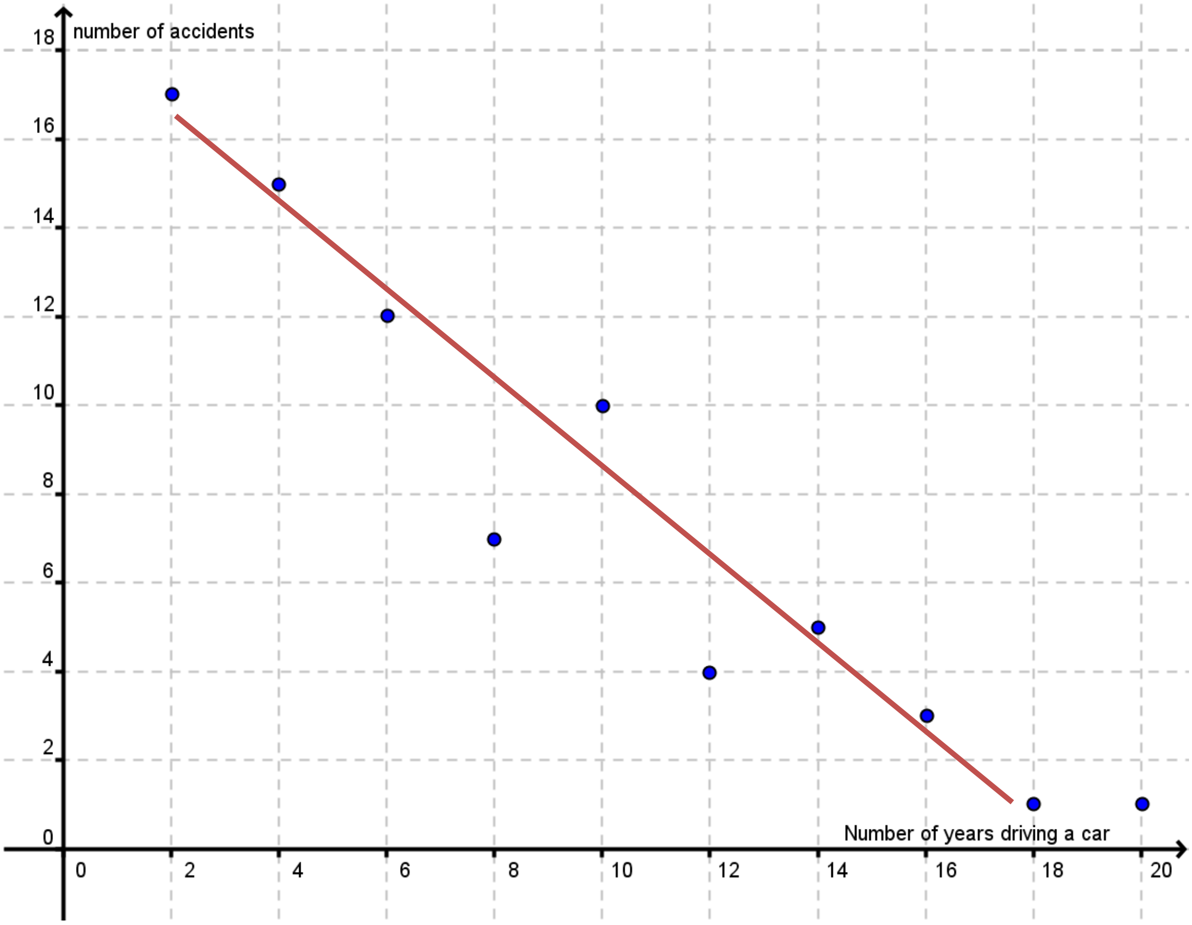
Activity 5

1. a)



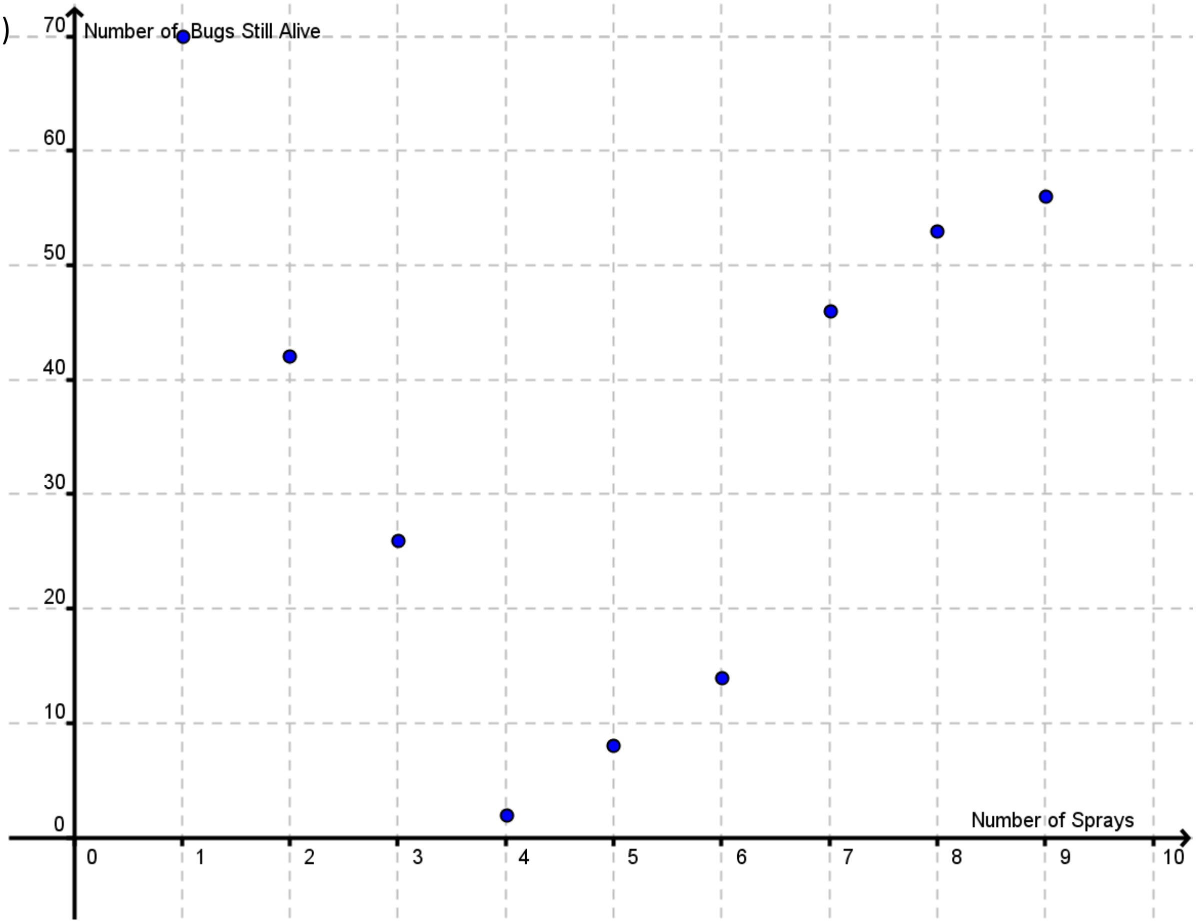
b) A straight line.

c)



d) A strong negative relationship.

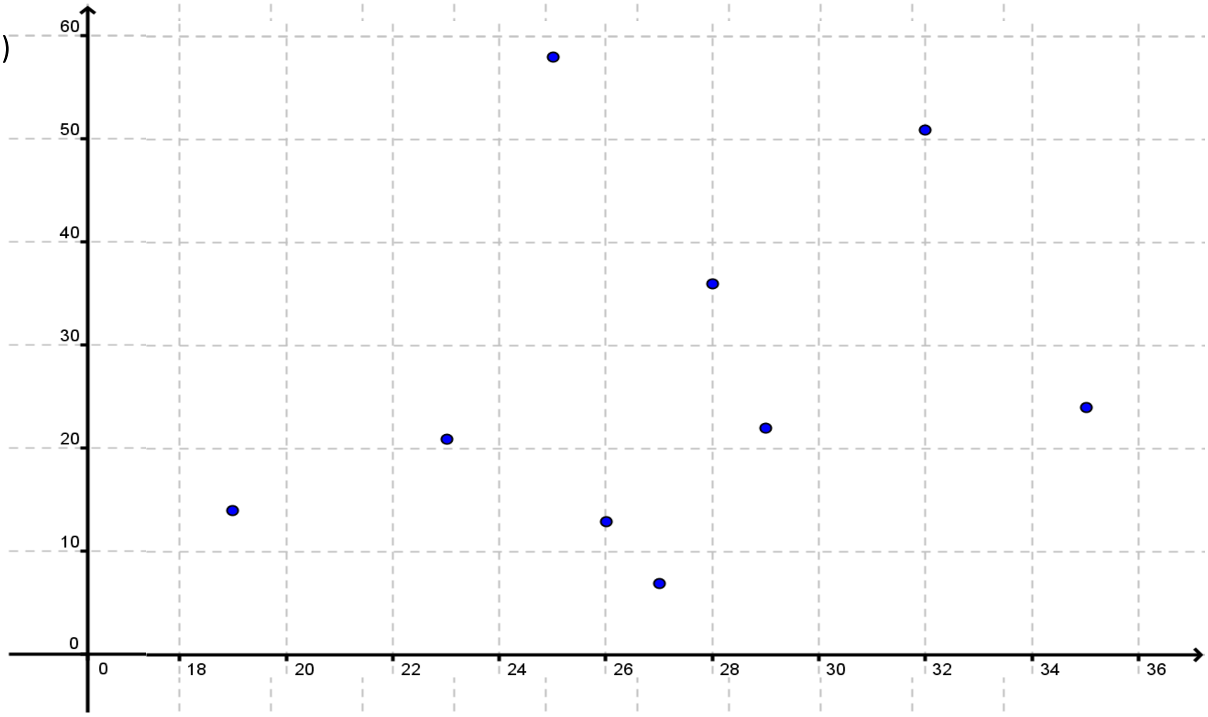
2. a)



b) A parabola (quadratic function)

c) The optimum number of sprays would be when the least number of bugs remain alive – thus the optimum number of sprays would be 4.

3. a)



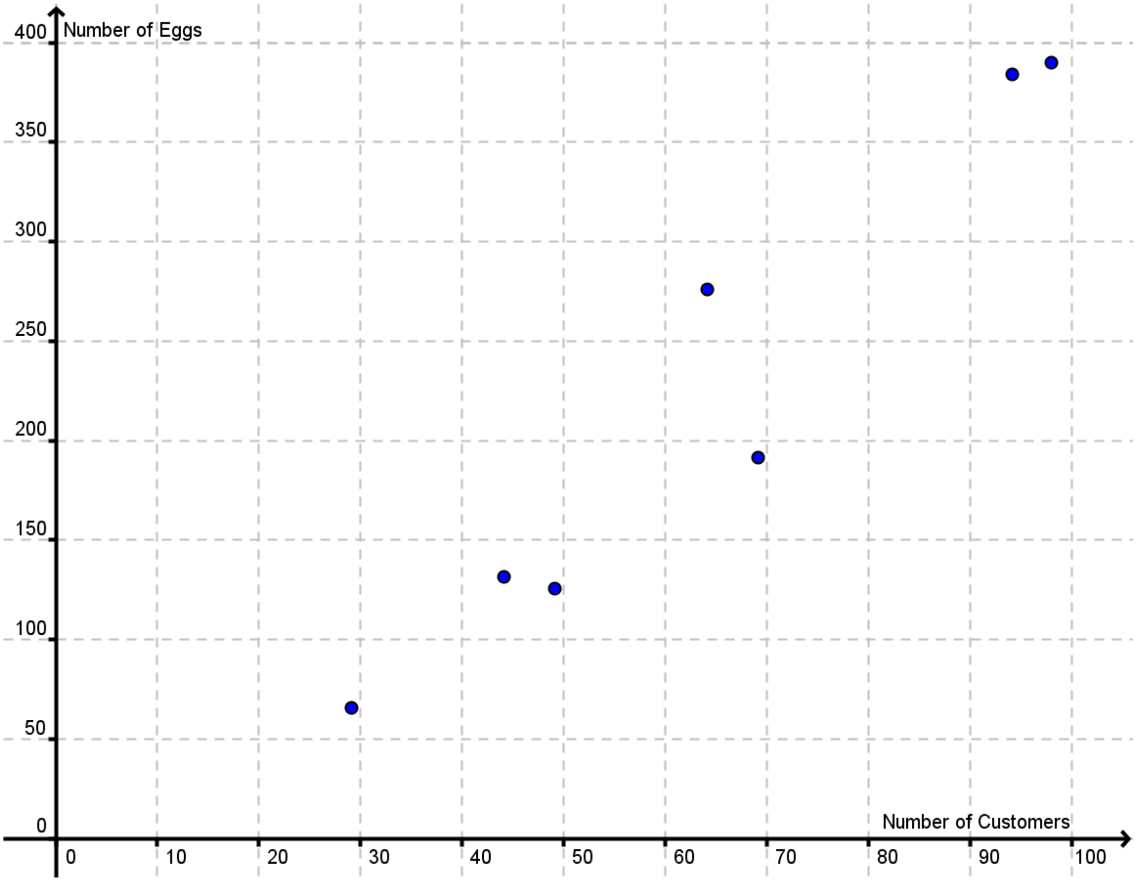
b) No there does not appear to be a relationship between the temperature and the number of murders committed on the day.

4. a) no relationship
b) strong, negative relationship
c) relatively strong, positive relationship
d) relatively strong, negative relationship
e) weak positive relationship
f) relatively strong positive relationship.

Activity 6

1. Regression line:
$y = bx + a$
$\therefore y = -0{,}906x + 17{,}467$

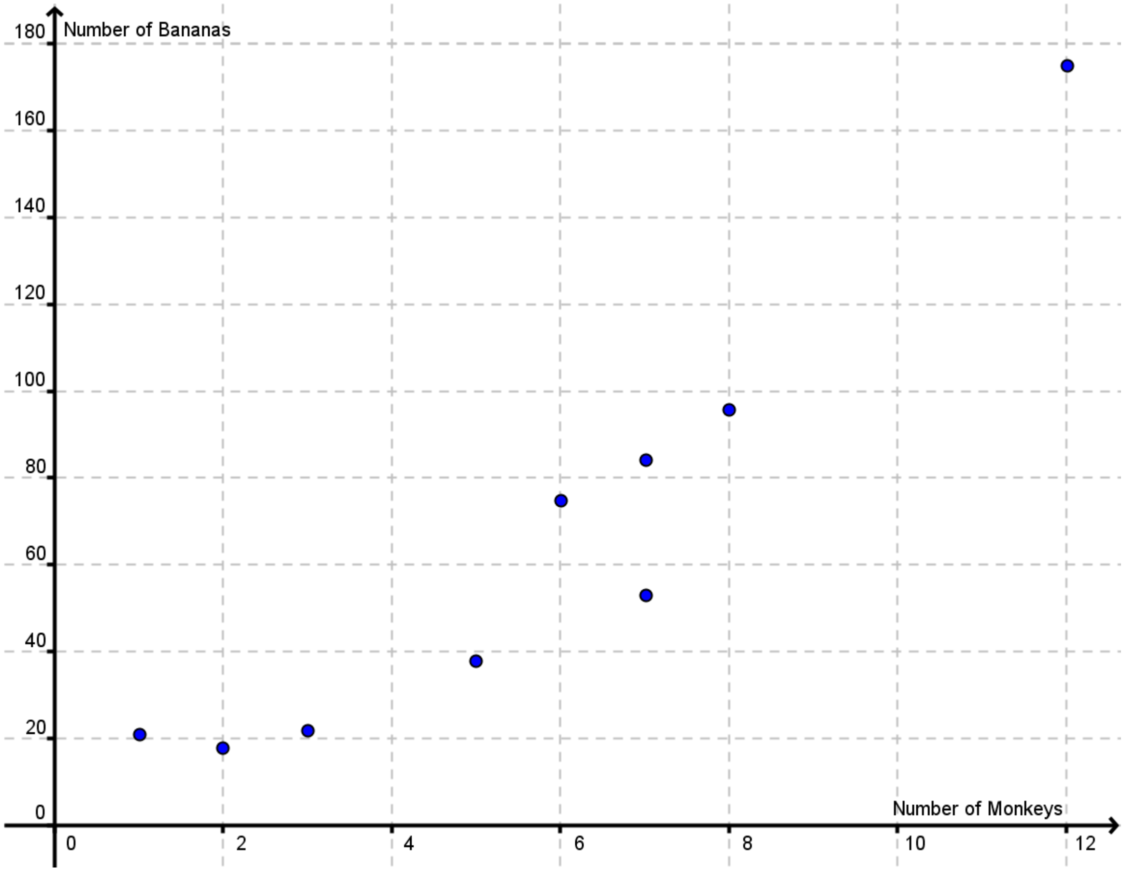
2. a)



b) a positive relationship

c) $y = bx + a$
$\therefore y = 4{,}865x - 86{,}952$

3. a)

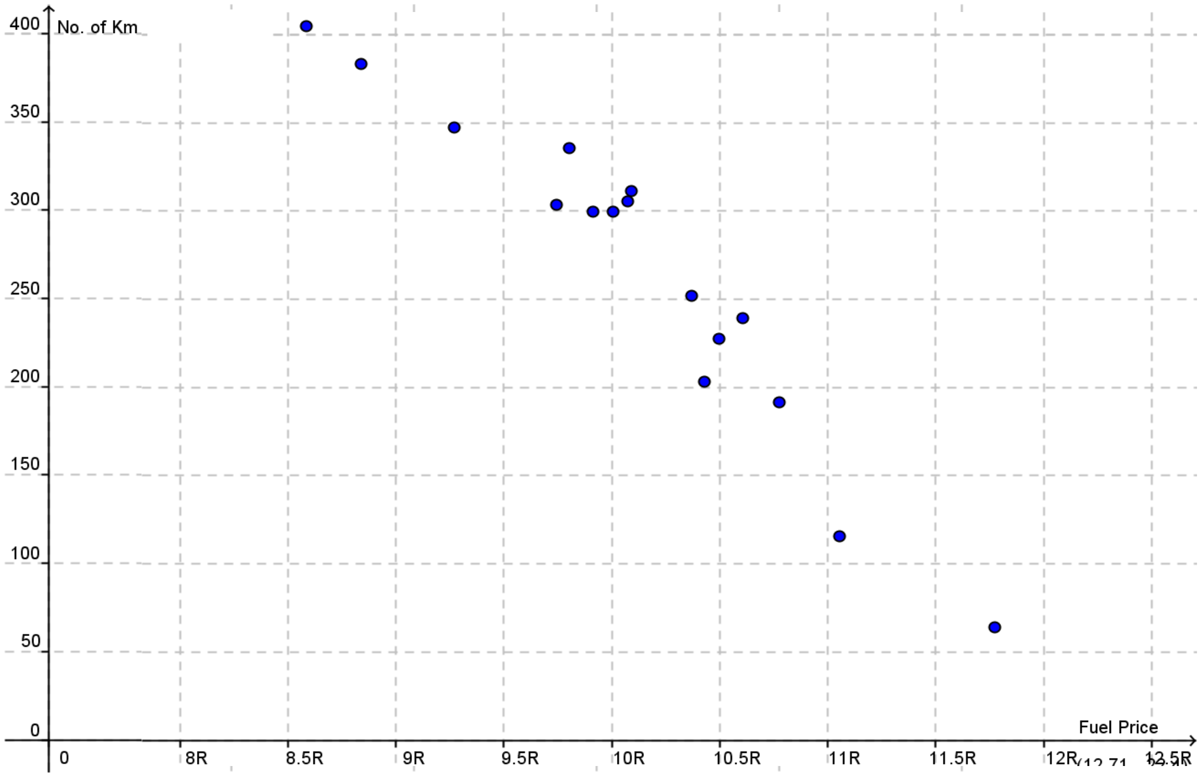


b) a positive relationship

c) $y = bx + a$

$\therefore y = 14{,}043x - 14{,}913$

4. a)



b) a negative relationship

c) $y = bx + a$

$\therefore y = -110{,}288x + 1383, 338$

d) $r = -0{,}957$

Activity 7

1. a) Graph B – there is a heading, a key, and correct labels.

b) Graph A

c) Graph A – the gap between the two lines looks bigger – thus it would be more effective in giving the impression that JHB is less safe than CT.

d) Graph B – the gap between the two lines looks smaller so it would like approximately the same number of accidents occurred and that there was therefore not much difference in the safety.

e) The difference is in the size of the graphs and the scale of the graphs. Because graph A's scale is more spread out than Graph B's scale the lines are more spread out – it also makes the data easier to read.

2. a) Graph A – the pictures between the graph are very graphic and thus effective in highlighting the plight of the rhinos.

b) the number of deaths of rhino have been increasing drastically over the last year.

c) Graph A.