**Scatter Diagrams:**

**Scatter diagrams are used to represent and compare two sets of data. By looking at a scatter diagram, we can see whether there is any connection (correlation) between the two sets of data.**

**Drawing scatter diagrams**

**Example**



Matt sells ice-creams at outdoor events. He often buys too much or too little ice-cream from the wholesalers, so does not make as much profit as he would like. He decides to record how many ice-creams he sells over a number of days, to see whether there is a link between the temperature and number of ice-creams sold. Here are his results:

|  |
| --- |
| **Temperature (°C)** | 21 | 26 | 15 | 24 | 18 | 29 | 20 | 27 | 23 | 17 | 30 | 19 |
| **Number of ice-creams sold** | 70 | 86 | 50 | 80 | 58 | 96 | 66 | 92 | 74 | 54 | 100 | 62 |

There appears to be a connection. When the temperature is low, the number of ice-creams sold is also low. When the temperature is high, the number of ice-creams sold is high.

But it is much easier to judge the results by looking at a **scatter diagram**.

Plotting a scatter diagram is easy. The results are in pairs, so it is just like plotting coordinates.

The axes have been drawn so that the temperature is on the horizontal axis and number of ice-creams sold on the vertical. We therefore plot the points (21, 70), (26, 86), (15, 50) etc.



Did you notice the jagged lines close to the origin?

For example:



These indicate a **broken scale**.

A broken scale is used when values close to 0 are not required. In this case, we only needed to start the horizontal axis at 15, and the vertical axis at 50.

Care must be taken to use the broken scale appropriately. It is OK to use it in this case, but it can sometimes be misleading.

**Types of correlation**

**Positive correlation**

If there is a **correlation** between two sets of data, it means they are connected in some way.



We have seen that as the temperature **increases**, the number of ice-creams sold **increases**. The results are approximately in a straight line, with a positive gradient. We therefore say that there is **positive correlation**.

**Negative correlation**

Look at the following scatter diagram. It shows the connection between the number of weeks a song has been in the Top 40 and sales of the single for that week.



There is a definite a connection between the two sets of data, as the results are approximately in a straight line. As the number of weeks **increases**, sales **decrease**. The line therefore has a negative gradient, and we say there is **negative correlation**.

**No correlation**

The following scatter diagram shows the connection between a person's house number and their IQ (one measure of intelligence).

**Lines of best fit**

The 'line of best fit' is a line that goes roughly through the middle of all the scatter points on a graph. The closer the points are to the line of best fit the stronger we can say the correlation is.

Look at the diagrams below:







The **line of best fit** is drawn so that the points are evenly distributed on either side of the line. There are various methods for drawing this 'precisely', but you will only be expected to draw the line 'by eye'.

You may be asked to comment on the nature of the correlation. This means you will be expected to say whether there is positive, negative or no correlation. Using terms such as 'strong', 'moderate' or 'weak' will give a clearer indication of the strength of the connection.

When drawing the line of best fit, use a transparent ruler so you can see how the line fits



It is obvious that there is no connection between these values, and this is shown by the scatter diagram. We say there is **no correlation**.

**Interpreting scatter diagrams**

Look at the following scatter diagram which shows the test results in maths and science for a class of 24 pupils.



**Question**

Is there any correlation between the results?

**Answer**

There is positive correlation.

**Question**

What is the highest mark in maths?

**Answer**

The highest mark for maths is 86. There are no marks higher than 86 on the horizontal scale.

**Question**

What is the lowest mark in science?

**Answer**

The lowest mark for science is 52.

**Question**

Jane scored 68 for her maths test, what was her mark for science?

**Answer**

There is only one result which has an 'x-coordinate' of 68. The 'y-coordinate' of this result is 70, so the science mark is 70.

**Question**

Where on the graph would you draw the line of best fit?

**Answer**



# Remember that the line of best fit goes through the middle of the distribution of all the points.

# Correlation and lines of best fit

You can also use a line of best fit to predict results.

**Question**

The heights and weights of 20 children in a class are recorded. The results are shown on the scatter diagram below.



Katie is 148 cm tall. Estimate her weight.

**Answer**



Start by drawing a line of best fit. Remember that the line of best fit is drawn so that the points are evenly distributed on either side of the line.

Katie is 148 cm tall, so we use the line of best fit to find an approximate weight. Find 148 cm on the height axis. Now follow the line up until you hit the line of best fit. Now read across the graph to the weight axis.

Katie weighs approximately 52 kg. As you are only drawing the line of best fit 'by eye', it is unlikely that your answers will be exactly the same as your friend's. The examiners will take this into account.

Looking at the graph of height against weight we need to interpret the gradient of the graph.

We choose two points on the graph and find the gradient (134, 30) and (148, 52).

As the height increases by 14 cm, the weight increases by 12 kg. So the gradient is 12/14, which tells us that for every increase of 1 cm, the weight increases by 12/14 kg (0.857 kg).

# Rogue values

A rogue value is a value that doesn’t quite fit with other findings from the same set of data.

These are various ways that we may get a rogue value. The data may have been:

* deliberately given incorrectly
* recorded incorrectly
* plotted incorrectly

Look at this graph showing the height and weight of a group of children.



Most of the points are close together but two points, at (144, 15) and (84, 62) are much further apart.

(144, 15) is likely to be a rogue value because it suggests that a person 144 cm tall weighs 15 kg. Very unlikely!

(84, 62) means a person is 84 cm tall and weighs 62 kg, which is possible and so this may be an **isolated value**.

We need to decide if the value is rogue or just isolated.

* For rogue values we should be able to suggest a reason.
* An isolated value is possible and may 'fit' if we had more data.

We use the axes to interpret a point, then decide if we have a rogue value.