**Collecting Data**

Data is another term for information.

**Types of Data**

Data which contains individual values is known as **discrete data.**

Some examples:-

* Number of goals scored
* Number of chairs in a classroom
* Marks achieved

Data which is measured on a scale is known as **continuous data**.

Some examples:-

* Heights of students in a class
* Speed of cars passing
* Time taken to complete an exam

A **variable** is something that is measured or observed.

**Categorical Data** is data that fits into a group and can be written in words, e.g.

* male/female,
* Colour of cars.

**Ordinal Data** is a type of categorical data, where the data can be written in an obvious order, e.g.

* Type of house – 1 bedroom, 2 bedrooms, 3 bedrooms…..
* Standard of hotel – 1 star, 2 stars, 3 stars ……….
* Opinion of scale – strongly disagree, disagree, neutral, agree, strongly agree

**Univariate** is when one type of data is collected from a group, e.g.

* Colour of hair
* Distance from school
* Weight in kilos

**Bivariate** is when two types of data are collected, e.g.

* Height and weight
* Gender and colour of hair
* Age of a car and price of a car

**Collecting Data**

**Primary Data** is collected by an organisation or an individual who is going to use the information.

Some ways in which Primary Data can be collected are as follows; - Questionnaire, experiment, investigation, observing.

**Primary Data**

|  |  |
| --- | --- |
| Advantages | Disadvantages |
| * You collect exactly what you require. * You know how it is collected. * You know from whom it is collected. | * Expensive. * Time consuming |

**Secondary data** is already available or has been collected by somebody else for a different purpose.

Examples of secondary data include Internet, Published Statistics, tables and charts in newspaper and magazines.

**Secondary Data**

|  |  |
| --- | --- |
| Advantages | Disadvantages |
| * Cheap to get. * Easy to get. | * Unknown method of collection. * May be out of date. * May not be exactly what you require. |

Surveys are useful for collecting data that is likely to be personal.

* Methods of surveys
* Postal surveys
* Personal interviews (used in market research)
* Telephone surveys
* Observation – which involves monitoring behaviour or information

Observation

|  |  |
| --- | --- |
| Advantages | Disadvantages |
| * Systematic and mechanical | * Results are prone to chance |

Questionnaire – A questionnaire is a set of questions designed to obtain data from individuals. The people who answer the questions are called respondents. Questions in questionnaire should be

* Clear to find out what data you need
* Short and simple questions
* Simple questions to start with to encourage the respondents to participate
* Provide response boxes where possible
* Avoid leading questions
* Avoid personal questions

**Bias –** when collecting data you need to make sure that your survey is fair and avoids biased. If bias exists, the data collected might be distorted.

**Measuring Data**

**Mode** – The most common value in a set of data.

**Mean (** – It considers every piece of data. It can be affected by extremes. The sum of all the numbers in a set of data divided by the number of numbers in the set is called the mean.

There is a different formula for grouped data.

**Median** – Put list of numbers in order. The median is the middle number ½(n+1) = median

**Range**

This is the highest value minus the smallest value.

**Interquartile range**

Upper Quartile - Lower Quartile.

**Standard Deviation - σ**

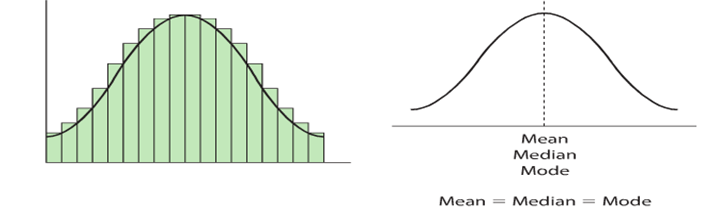
This measures the spread of the data around the mean. It can be done on the calculator as follows,

**Shapes of distribution**

If a histogram is symmetrical /balanced, it is not skewed. However when it does not have an axis of symmetry it is said to be skewed or it has less balance.

Histograms are useful to see where data lies

**Types of distribution**

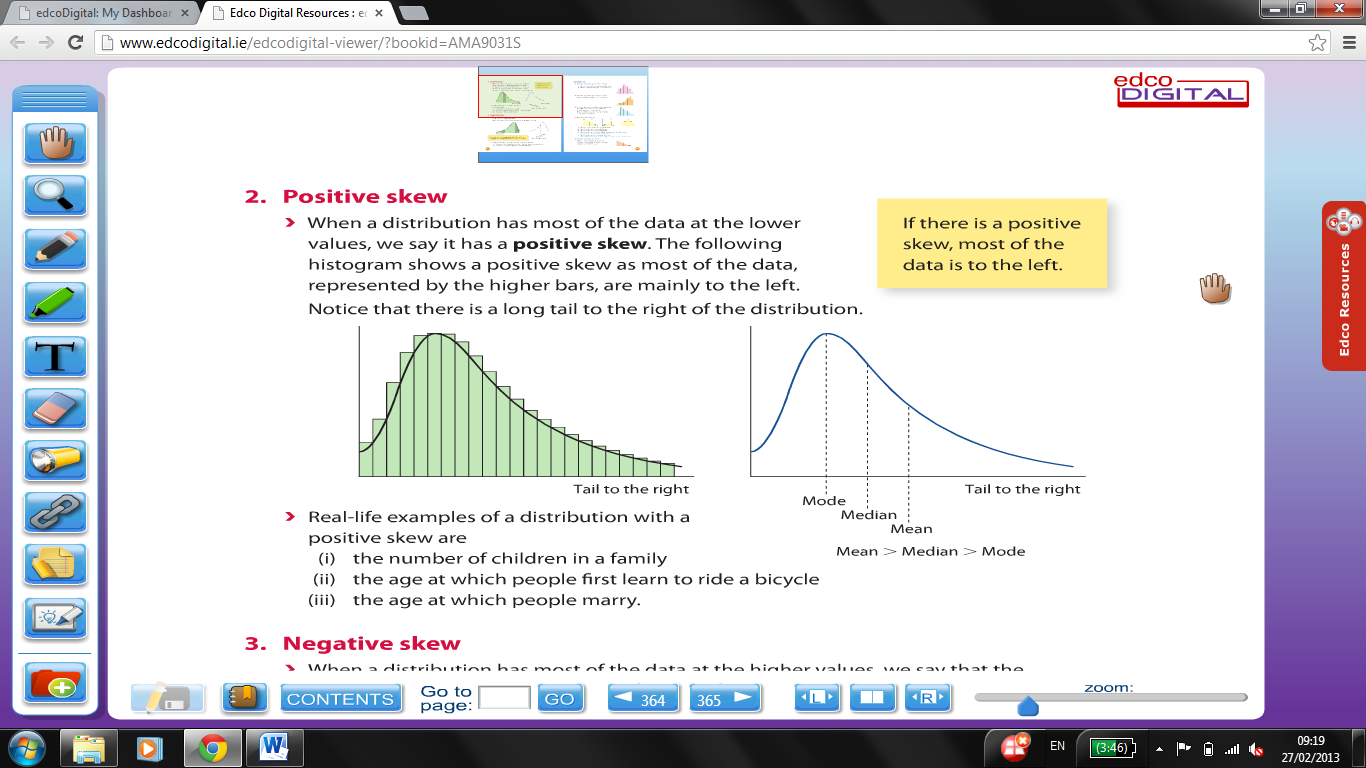
Symmetrical Distributions

This has an axis of symmetry right down the middle

* It is the most common and most important of distributions in statistics.
* It is known as the normal distribution
* Real life examples, the heights of random sample of people, the I.Q. of a population

**Positive Skew**

When a distribution has most of its data at the lower values, it is said to have a positive skew. If is a positive skew, most of the data is on the left.

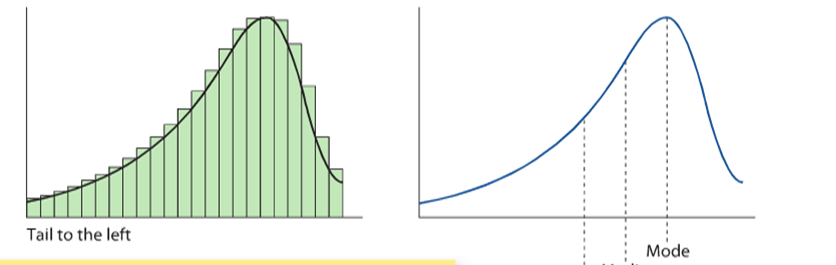
(Tail to the right) 

Mean > Median >Mode

Real life example:-the number of children in a family.

**Negative Skew**

When a distribution has most of the data at the higher values, we say it has negative skew. (Tail to the left)



Mean < Median < Mode

Real life examples: the ages at which people get their first reading glasses.

**Stem/Leaf Graph**

A stem/leaf is another useful way of representing data. It shows the original data and gives an overall view picture of the shapes of distribution.

It is similar to a horizontal bar chart, with numbers representing the amounts of data. They are suitable for small amounts of data.

You must always include a key to show how the stem/leaf combines and you must put the data in order.

Median – Total plus 1 divided by 2.

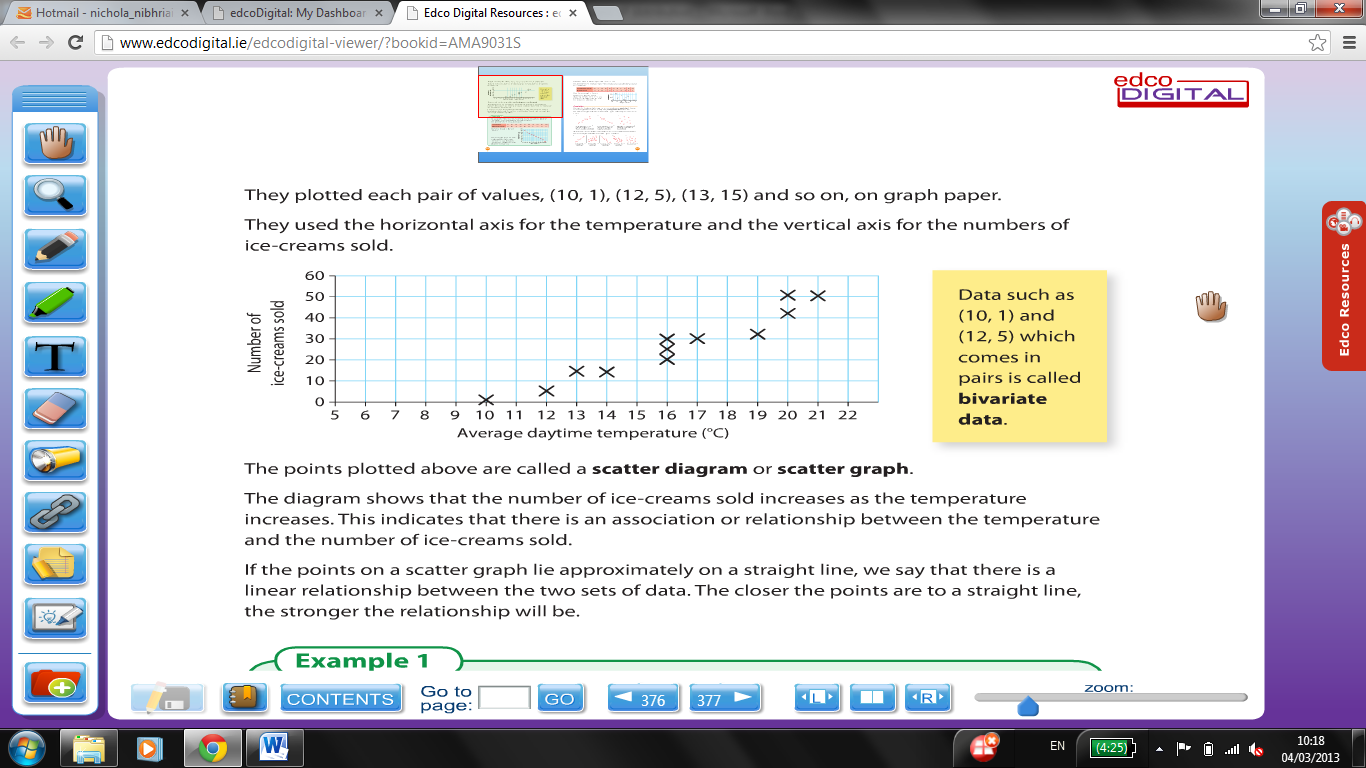
Interquartile Range – Upper Quartile – Lower Quartile

Total plus multiplied by ¼ =Lower

Total plus one multiplied by ¾ = Upper

**Scatter plot**

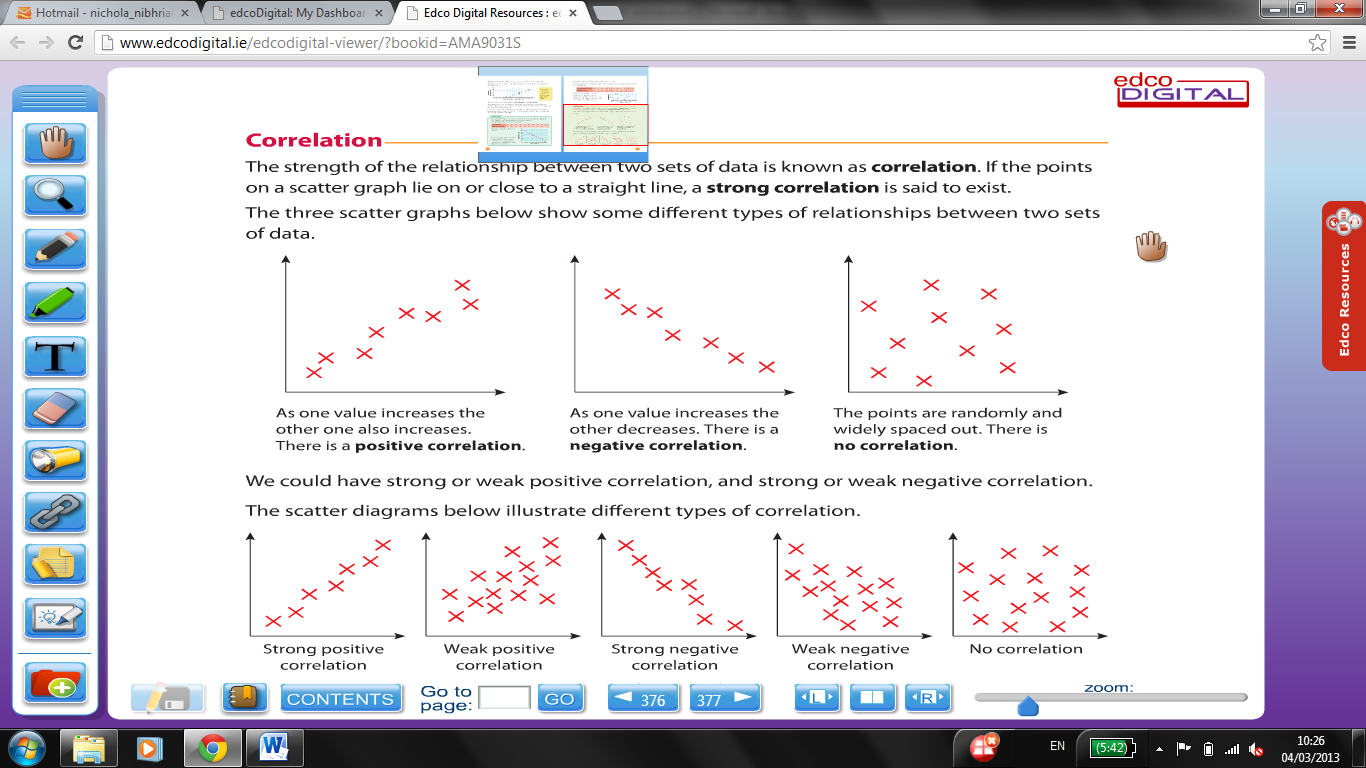
Scatter plots are used to compare two sets of data and to see if there is a relationship between the two sets of data



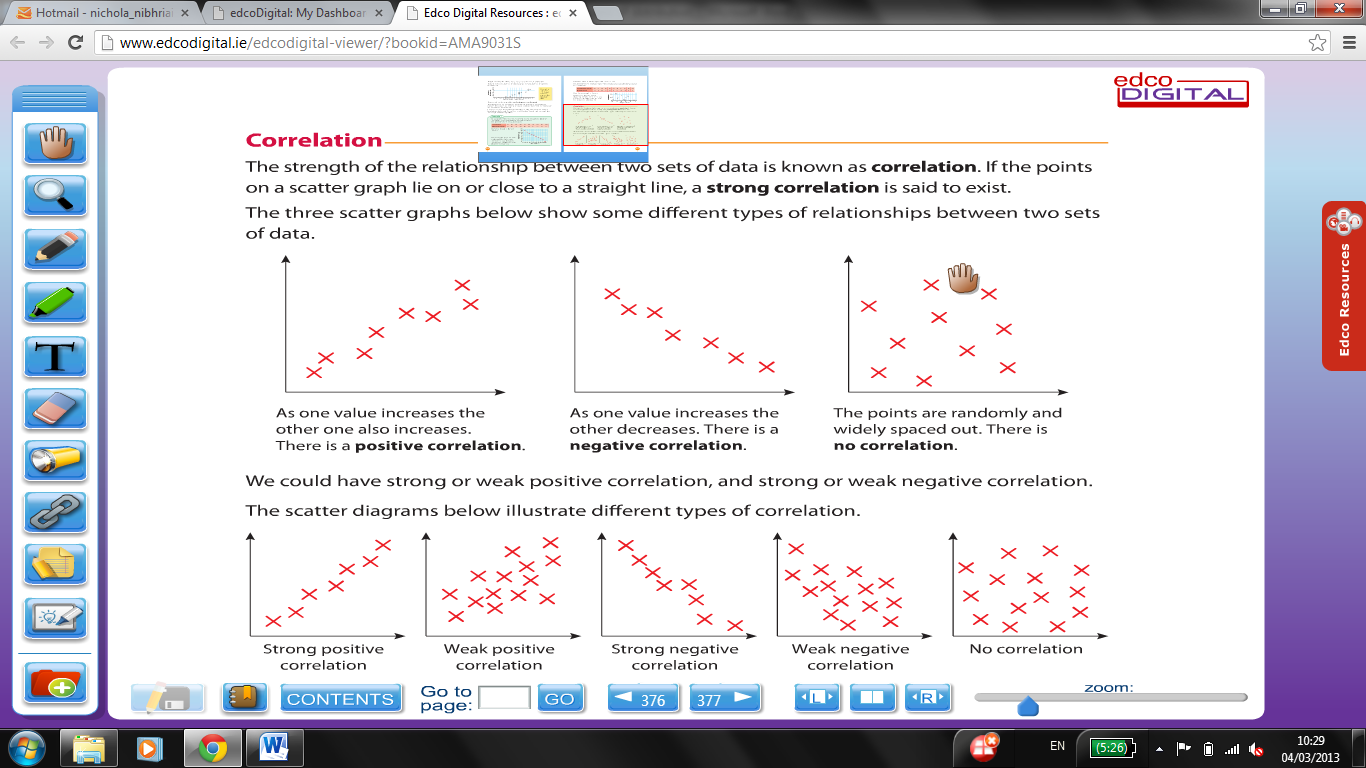
If the points on a scatter diagram are almost on a straight line, we say there is a linear relationship between the 2 sets of data. The closer the points are to a straight line the stronger the relationship between the 2 sets of data.

Correlation

The strength of the relationship between 2 sets of data is known as correlation.



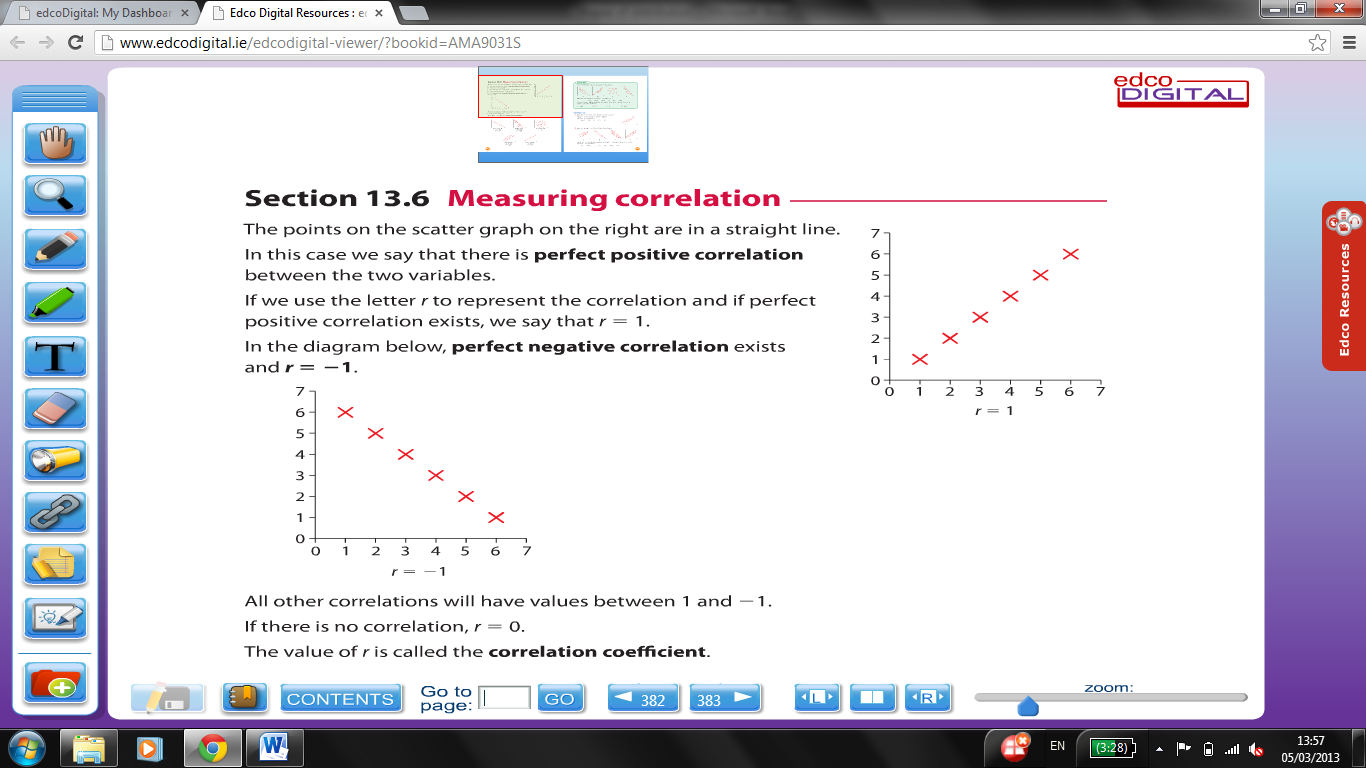
There can also be a strong/weak positive correlation and strong/weak negative correlation.



Measuring Correlation

r is called the correlation coefficient

Perfect positive correlation, r=1



Perfect Negative Correlation, r=-1,

