Statistics 2

Key words

scatter diagram **correlation** **causal relationship** **causality**
correlation coefficient **line of best fit** **normal distribution** **normal curve**
Empirical Rule **standard scores (*z*-scores)** **margin of error**
confidence interval **hypothesis testing** **null hypothesis**

Section 4.1 Scatter diagrams

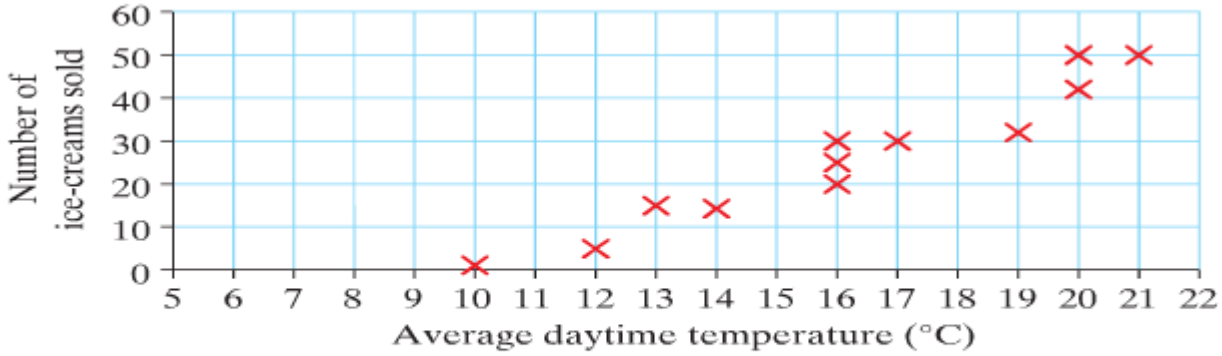
A **scatter graph** or **scatter diagram** is a graph consisting of points plotted on an x–y plane. Each point represents the values of two different variables such as the heights and weights of different individuals. Such data connecting two variables is called **bivariate data**.

After plotting the points on a scatter graph, we look for a pattern, particularly a **linear** pattern. If the points on a scatter graph lie approximately on a straight line, we say that there is a linear relationship between the two sets of data. The closer the points are to a straight line, the stronger the relationship will be.

The table below shows the number of ice-creams sold by a shop over a 12-day period.

Average temperature (°C)	10	12	16	20	13	16	14	17	19	20	21	16
No. of ice-creams sold	1	5	20	50	15	25	14	30	32	42	50	30

Using the horizontal axis for the temperature and the vertical axis for the numbers of ice-creams sold, we get the following scatter graph.



The scatter graph shows that the number of ice-creams sold increases as the temperature increases. Since the points lie close to a straight line, we say that there is a linear relationship between the two sets of data.

Correlation

Correlation is a measure of the strength of a relationship between two variables, say x and y. We say x and y are correlated if a scatter graph shows a linear pattern to the plotted points (x, y). If no pattern exists, the variables are not correlated.

The three diagrams below illustrate **positive correlation**, **negative correlation** and **no correlation**.

The variables x and y have a **positive correlation** if y increases as x increases.

Positive correlation

The variables x and y have a **negative correlation** if y decreases as x increases.

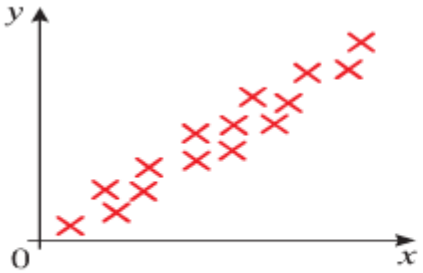
Negative correlation

The variables x and y show no linear pattern.

No correlation

The correlation is **high** if the points are close to a straight line.
The correlation is **low** if the points are more spread out.

It is possible to have strong and weak positive correlations as well as strong and weak negative correlations.
The scatter diagrams below illustrate these possibilities:



Strong positive correlation

Weak positive correlation

Strong negative correlation

Weak negative correlation

Example 1

The table shows the weights and heights of 12 people.

Height (cm)	150	152	155	158	158	160	163	165	170	175	178	180
Weight (kg)	57	62	63	64	58	62	65	66	65	70	66	67

(i) Draw a scatter graph to show this data.
(ii) Describe the strength and type of correlation between these heights and weights.

Causal relationships and correlation

The price of a used car depends, among other things, on the age of the car. The age of the car **causes** the price of the car to decreases. We say that there is a **causal relationship** between the price of the car and the age of the car.

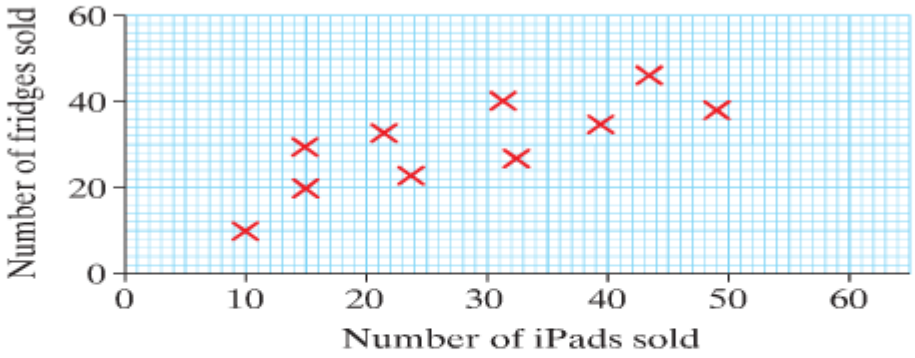
Definition

When a change in one variable causes a change in another variable, we say that there is a causal relationship between them.

The scatter graph shows the relationship between the sales of iced drinks and temperature. The correlation is strong and positive. You would expect this as a rise in temperature would tend to result in an increase in the sales of iced drinks.

It would therefore be reasonable to conclude that there is a **causal relationship** between the sales of iced drinks and an increase in temperature.

The scatter diagram below shows the number of iPads and the number of fridges sold by an electrical shop over a ten-month period.



The graph shows that there is a reasonably strong positive correlation between the number of iPads sold and the number of fridges sold. However, this does not mean that there is a causal relationship between them; buying an iPad does not cause you to buy a fridge.

Correlation does not necessarily mean that there is a causal relationship.

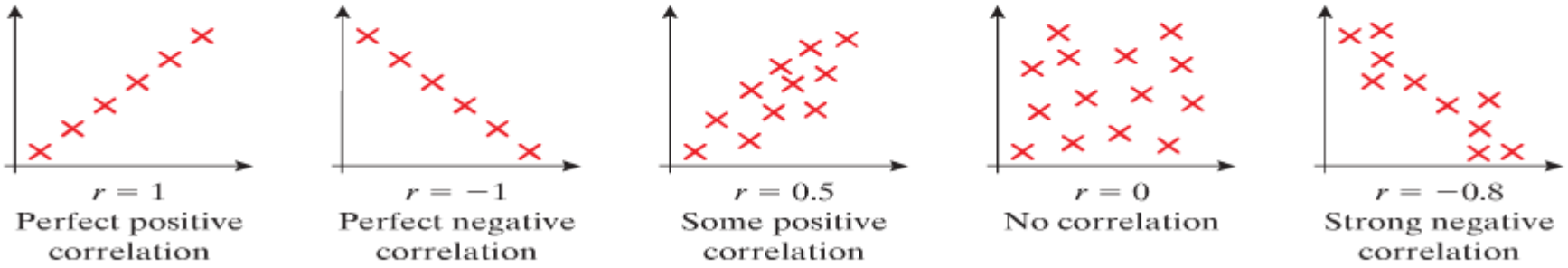
Section 4.2 Measuring correlation – Line of best fit

1. Calculating the correlation coefficient

Correlation is a measure of the strength of the relationship between two sets of data.
We use the letter r to denote the **correlation coefficient**.
The value of r will always lie between -1 and 1.

$r = 0$ indicates no correlation.
$r = 1$ indicates **perfect positive** linear correlation.
$r = -1$ indicates **perfect negative** linear correlation.

Here are some examples of the value of r:



$r = 1$
Perfect positive correlation

$r = -1$
Perfect negative correlation

$r = 0.5$
Some positive correlation

$r = 0$
No correlation

$r = -0.8$
Strong negative correlation

The nearer the value of r is to 1 or -1, the closer the points on the scatter diagram are to a straight line.

There are several methods of calculating a correlation coefficient.
The method selected for our course is called the **product–moment correlation coefficient, r.**
The formula that is used to find this coefficient involves a lot of tedious calculations.

For our course, it is recommended that we use the electronic calculator to find the value of r.

The steps involved in the input of data and finding the value of r are given in *Appendix 1* at the end of the book.

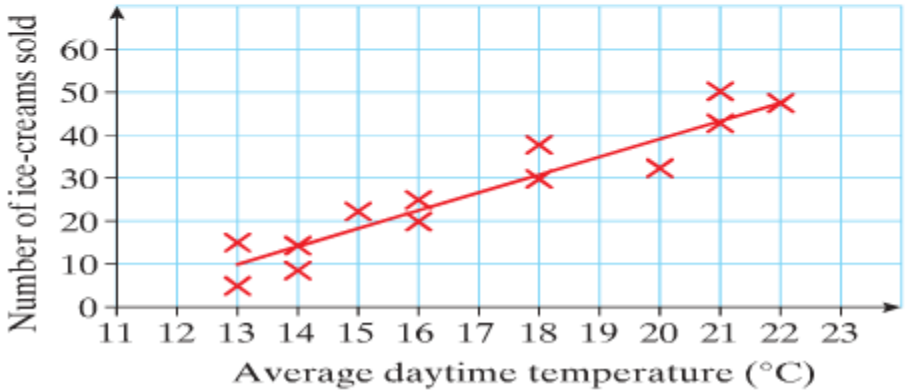
2. The line of best fit

We have already stated that when points on a scatter diagram lie on, or close to, a straight line a strong correlation exists.

When a line is drawn through the points on a scatter diagram, it is called a **line of best fit**. Try to draw the line that fits best through the points. You should aim to have roughly the same number of points on either side of your line. This method is generally referred to as drawing the line of best fit **by eye**.

This scatter diagram shows the average daytime temperature plotted against the number of ice-creams sold.

> A line that is drawn to pass as close as possible to all the plotted points on a scatter diagram is called the **line of best fit**.



The line drawn is called the line of best fit.

It can go through some, all or none of the points.

This line shows the general trend of the relationship between the two sets of data. The line can be used to estimate other values. However, estimating values where the line has been extended beyond the existing points is less reliable.

Example 1

The table below shows the number of hours of sunshine and the maximum temperature in ten Irish towns on a particular day.

Maximum temperature (°C)	12	13	14	15	16	17	18	19	20	21
Hours of sunshine	9.6	11.6	10.2	13.25	11.8	13.6		15.4	15.2	15

(i) Plot a scatter diagram and draw a line of best fit for the data.
(ii) Use your line of best fit to estimate the number of hours of sunshine when the maximum temperature was 18°C.
(iii) Describe the correlation shown in your diagram.
(iv) Use your calculator to find the correlation coefficient.

The correlation coefficient, by calculator, is 0.9176.

3. Finding the equation of the line of best fit

From our knowledge of coordinate geometry, we should be familiar with the equation of a line in the form $y = mx + c$.
In statistics, it is more usual to use it in the form $y = ax + b$.

The equation $y = ax + b$ has a gradient a and its intercept on the y-axis is $(0, b)$.

Thus to find the equation of a line of best fit drawn by eye,

(i) find two points on the line and use these points to find the slope of the line.

(ii) Use the slope found and one of the points to find the equation of the line using

$$y - y_1 = m(x - x_1).$$

(iii) Express the equation in the form $y = ax + b$.

Example 3

The table below shows the weights and heights of 12 pupils.

Height (cm)	150	152	155	158	158	160	163	165	170	175	178	180
Weight (kg)	57	62	63	64	58	62	65	66	66.5	70	66	67

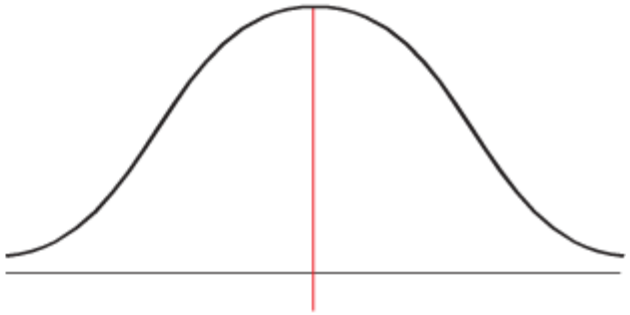
(i) Draw a scatter diagram to show this data.
(ii) Describe the strength and type of correlation between these heights and weights.
(iii) Draw a line of best fit on your scatter diagram.
(iv) Tony is 162 cm tall.
Use your line of best fit to estimate his height.
(v) Use your calculator to find the correlation coefficient, correct to two decimal places.
(vi) Find the equation of the line of best fit in the form $y = ax + b$.

By calculator, the correlation coefficient is 0.8.

Section 4.3 The normal distribution

In Section 3.6 of this book, we were introduced to the **normal distribution** which is the cornerstone of modern statistics.

The **normal curve** is a smooth bell-shaped and symmetrical curve.
The red line is the axis of symmetry.

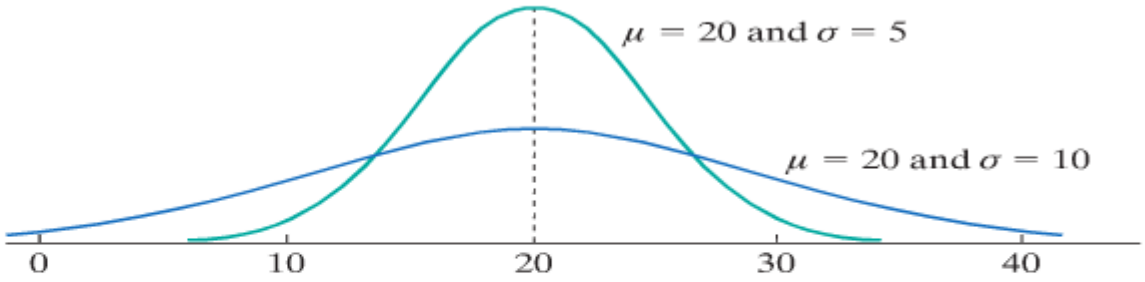


The mean, the mode and the median are all equal and they lie on the axis of symmetry.

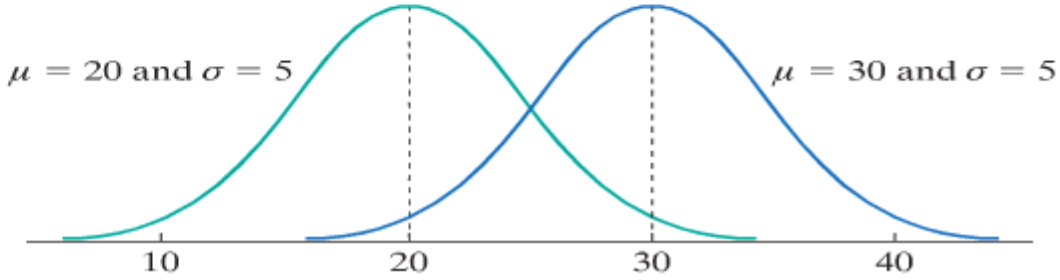
Normal distributions occur frequently in nature. For example the heights and weights of all adult males in Ireland will be **normally distributed**.
If all the heights of these adult males were plotted on a graph, the result would be a smooth bell-shaped curve, as shown above.

All normal distributions will have a mean (μ) and standard deviation (σ).
Different values for μ and σ will give different normal distributions.

The diagram on the right shows two normal distributions with the same mean but different standard deviations.



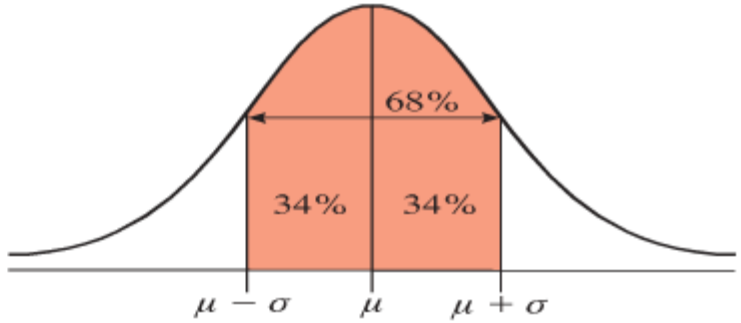
This diagram shows two normal distributions with the same standard deviation but different means.



All normal distributions share some very important characteristics.

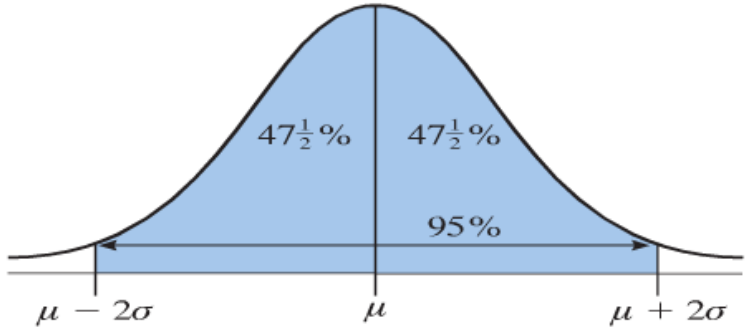
1. About 68% of all the values of any normal distribution lie within one standard deviation of the mean, i.e., in the range $[\mu - \sigma \text{ and } \mu + \sigma]$.

34% lie to the right of the mean.
34% lie to the left of the mean.

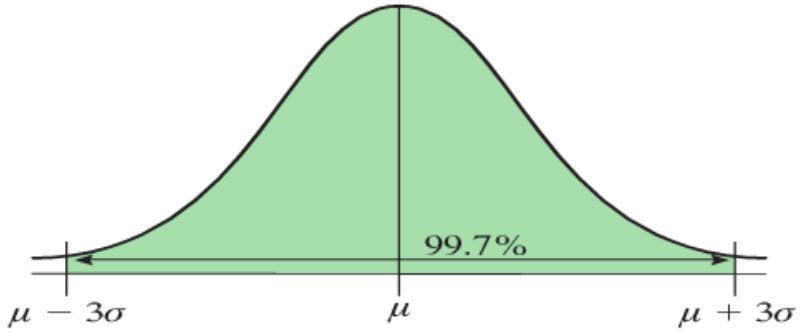


2. About 95% of all values lie within two standard deviations of the mean, i.e., in the range $[\mu - 2\sigma \text{ and } \mu + 2\sigma]$.

$47\frac{1}{2}\%$ lie to the right of the mean.
$47\frac{1}{2}\%$ lie to the left of the mean.



3. Almost all (99.7%) of the values lie within three standard deviations of the mean.



The three characteristics of the normal distribution listed above are generally known as **The Empirical Rule**.

The Empirical Rule

- Approximately 68% of a normal distribution lie within one standard deviation of the mean
- 95% lie within two standard deviations of the mean
- 99.7% lie within three standard deviations of the mean.

Example 1

The marks, out of 100, in an examination are normally distributed.
The mean mark is 60 and the standard deviation is 6 marks.

(i) Work out the mark that is two standard deviations above the mean.

(ii) What percentage of the marks lie between 48 and 72 marks?

(iii) What percentage of the marks lie between 60 and 72 marks?

(iv) If 1000 students took the examination, how many students scored less than 54 marks?

Example 2

Bottles of 300 ml shampoo are filled with the amounts in the bottles normally distributed with a mean of 300 ml and a standard deviation of 3 ml.
If 10 000 bottles are filled, how many bottles contain amounts that are

(i) within one standard deviation of the mean?

(ii) more than two standard deviations above the mean?

(iii) If the manufacturer rejects bottles that contain amounts more than 3 standard deviations from the mean, what is the largest amount of shampoo in a bottle you would find for sale?

Example 3

The weights of a group of 1000 school children were normally distributed with a mean of 42 kg and a standard deviation of σ.
If 950 of the children were in the range 30 kg to 54 kg,

(i) find the value of σ

(ii) find the probability that a child selected at random was in the range 36 kg to 48 kg.

Standard scores (*z*-scores)

In a state examination, Karen got 72% in her English examination and 68% in her Maths examination. In which examination did she achieve the better result? To determine this, we would need to know the average mark and standard deviation for each subject. We would then need to find the number of standard deviations Karen's mark was above or below the mean in each subject. If her mark was 1 standard deviation above the mean in English and 0.75 standard deviations above the mean in Maths, then Karen would have done relatively better in her English examination.

The number of standard deviations that a value lies above or below the mean is called a **standard score** or ***z*-score**.

In general, if x is a measurement belonging to a set of data with mean μ and standard deviation σ, then its value in z-units is given below:

$$z = \frac{x - \mu}{\sigma}, \text{ where}$$

x is the score or value
μ is the mean
σ is the standard deviation

Standard scores are very useful when comparing values from different normal distributions.

Example 4

Simon and Susan did a test in French and a test in Science.
Both tests had a maximum mark of 100. The results are given in the table below:

	Susan's mark	Simon's mark	Mean mark	Standard deviation
French	75	50	60	10
Science	65	40	50	5

Work out the z-scores for each subject and comment on the performance of Simon and Susan.

Section 4.4 Margin of error – Confidence intervals – Hypothesis testing

When dealing with sampling in *Statistics 1*, it was stated that the purpose of sampling is to gain information about the whole population by surveying a small part of the population. If data from a sample is collected in a proper way, then the sample survey can give an accurate indication of the population characteristic that is being studied.

Before a general election, a national newspaper generally requests a market research company to survey a sample of the electorate regarding their voting intentions in the election. The number surveyed is generally about 1000.

The result of the survey might appear in the daily newspaper as follows:

40% support for **The Democratic Right**.

The **40%** support is called the **sample proportion**, that is, the part or portion of the sample who indicated that they would vote for **The Democratic Right**.
A sample proportion is used to give an estimate of the **population proportion** who intend to vote for **The Democratic Right**.

The notation $\hat{p}$ is used to denote **sample proportion**.
The notation p is used to represent **population proportion**.
Since p is generally not known, $\hat{p}$ is used as an **estimor** for the true population proportion, p.

Of course everybody knows that sample surveys are not always 100% accurate. There is generally some 'element of chance' or **error** involved.

The newspaper might add to their headline the following sentence:

The margin of error is 3%.

The **margin of error** of 3% is a way of saying that the result of the survey is 40% ± **3%**. That means that the research company is quite 'confident' that the proportion of the whole electorate who intend to vote for **The Democratic Right** could be anywhere between 37% and 43%.

How does the research company calculate 'the margin of error'?

The margin of error in opinion polls is generally calculated using the formula,

$E = \frac{1}{\sqrt{n}}$, where n is the sample size.

If the sample size is 1000, then $E = \frac{1}{\sqrt{1000}} \approx 3\%$.

If the sample size is increased, the margin of error will be reduced.

Margin of error

$$E = \frac{1}{\sqrt{n}}$$

Confidence interval

The result of the opinion poll above was given as $40\% \pm 3\%$.
That could be written as $37\% < p < 43\%$, where p is the population proportion.
$37\% < p < 43\%$ is called the **confidence interval**.

The 'confidence' level is pitched at 95%.
The 95% confidence implies that the interval was obtained by a method which 'works 95% of the time'.

The confidence interval, $37\% < p < 43\%$, is a way of stating that if you surveyed many samples of 1000 people on the same day, the results would be in the interval 37% to 43% in 95% of the samples.

In our course, the confidence level is always at 95%.

At the 95% confidence level, the confidence interval for a population proportion is given on the right.

Confidence interval is

$$\hat{p} - \frac{1}{\sqrt{n}} < p < \hat{p} + \frac{1}{\sqrt{n}}$$

The confidence interval above may be also expressed as $\hat{p} \pm \frac{1}{\sqrt{n}}$.

Example 1

What sample size would be required to have a margin of error of

(i) 0.05 (ii) $2\frac{1}{2}\%$?

Example 2

A random sample of 400 persons are given a flu vaccine and 136 of them experienced some discomfort.
Construct a 95% confidence interval, p, for the population proportion who might experience discomfort.

Example 3

A survey of 100 residents of a Dublin suburb were asked if they remembered seeing an advertisement for McCain's chips on television. 60 respondents said that they had.

(i) Calculate the sample proportion, $\hat{p}$.

(ii) Find the margin of error, E.

(iii) Construct a 95% confidence interval for p.

Hypothesis testing

A **hypothesis** is a statement or conjecture made about some statistic or characteristic of a population.
Here is an example of a hypothesis:

'A football team is most likely to concede a goal just after it has scored a goal'.

A **hypothesis test** is a statistical method of proving the truth or otherwise of the statement or claim.

A local council reduced the speed limit on a dangerous 8 km stretch of country road from 80 km/hr to 60 km/hr. The number of accidents on the stretch was reduced from 5 per mon to 3 per month. The council claimed that the speed reduction was effective. Is the council correct in its claim?

In cases like this, a hypothesis test is set up to prove or disprove the claim.

Procedure for carrying out a hypothesis test

The procedure for carrying out a hypothesis test will involve the following steps:

1. Write down $\mathbf{H_0}$, the **null hypothesis**, and $\mathbf{H_1}$, the **alternative hypothesis**
   For example, to test if a coin is biased if we get 7 heads in 10 tosses, we could formulate the following hypothesis:
   H_0 : The coin is not biased.
   H_1 : The coin is biased.

2. Write down or calculate the sample proportion, $\hat{p}$.

3. Find the margin of error.

4. Write down the confidence interval for p, using

$$\hat{p} - \frac{1}{\sqrt{n}} < p < \hat{p} + \frac{1}{\sqrt{n}}$$

5. (i) If the value of the population proportion stated is within the confidence interval, accept the null hypothesis H_0 and reject H_1.
   (ii) If the value of the population proportion is outside the confidence interval, reject the null hypothesis H_0 and accept H_1.

Example 4

A drugs company produced a new pain-relieving drug for migraine sufferers and claimed that the drug had a 90% success rate. A group of doctors doubted the company's claim. They prescribed the drug for a group of 150 patients. After six months, 120 of these patients said that their migraine symptoms had been relieved by the drug.

At the 95% level of confidence, can the company's claim be upheld?

Example 5

A coin is tossed 1000 times and heads occur 550 times.
At the 95% confidence level, does the result indicate that the coin is biased?