

5

Statistical applications

CHAPTER OBJECTIVES:

- 4.1 The normal distribution; random variables; the parameters μ and σ ; diagrammatic representation; normal probability calculations; expected value; inverse normal calculations
- 4.2 Bivariate data: correlation; scatter diagrams; line of best fit; Pearson's product-moment correlation coefficient, r
- 4.3 The regression line for y on x
- 4.4 The χ^2 test for independence: null and alternative hypotheses; significance levels; contingency tables; expected frequencies; degrees of freedom; p -values

Before you start

You should know how to:

- 1 Find the mean and standard deviation of a set of data and comment on the relationship between them, e.g. for the data set

4, 5, 6, 8, 12, 13, 2, 5, 6, 9, 10, 9, 8, 3, 5:

Mean =

$$\frac{(4+5+6+8+12+13+2+5+6+9+10+9+8+3+5)}{15}$$

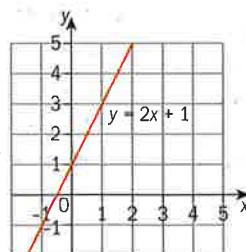
$$= \frac{105}{15} = 7$$

On a GDC, the mean is indicated by \bar{x} .

Using a GDC, standard deviation $(\sigma_x) = 3.10$ (to 3 sf).

The small standard deviation implies that the data are close to the mean.

- 2 Sketch the graph of the equation of a straight line, e.g. the straight line $y = 2x + 1$ passes through the point $(0, 1)$ and has gradient 2.



Skills check

- 1 Find the mean and standard deviation of these sets of data. Comment on your answers.

a 2, 4, 3, 6, 3, 2, 5, 3, 2, 5, 4, 4, 3, 5, 2, 3, 4, 5

x	Frequency
12	1
13	2
14	23
15	2
16	1

For help, see Chapter 2, Sections 2.4 and 2.7.

- 2 Sketch the graphs of:

a $y = -3x + 4$
b $y = 2x - 6$



The people in this photograph are a sample of a population and a source of valuable data. Like a lot of data on natural phenomena, people's heights and weights fit a 'normal distribution', which you will study in this chapter. Medical statisticians use this information to plot height and weight charts, and establish guidelines on healthy weight.

The information can also be used to chart changes in a population over time. For example, the data can be analyzed to determine whether people, on the whole, are getting taller or heavier. These results may affect or even determine government health policy. Moreover, manufacturing and other industries may use the information to decide whether to, for example, make door frames taller or aircraft seats wider.

You may think that some data might be related, for example, people's height and shoe size, or perhaps a child's height and their later adult height. This chapter shows you how to investigate correlation and the strength of relationships between data sets.

Investigation – related data?

Do you think that height and shoe size are related? Collect the height and shoe size of at least 60 students in your school.

Plot these data points on a graph. Use the x -axis for 'Height' and the y -axis for 'Shoe size'. Do not join up the points.

Does the data support your original hypothesis on height and shoe size?

The graph you will draw in this investigation is called a **scatter diagram**. You will find out more about scatter diagrams and the correlation between data sets in Section 5.2 of this chapter.

5.1 The normal distribution

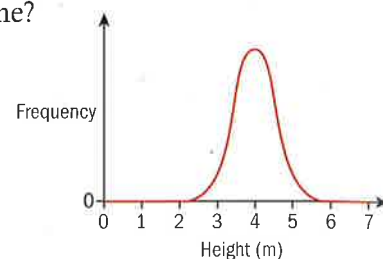
For his Mathematical Studies Project, Pedro measures the heights of all the apple trees in his father's orchard. There are 150 trees.

If Pedro drew a diagram to represent the frequency of the heights of all 150 trees, what do you think it would look like?

Pedro then measures the heights of the apple trees in his uncle's orchard. If he drew a diagram of the frequencies of these heights, do you think that this diagram would look different to the previous one?

In both orchards there would probably be a few very small trees and a few very large trees – but those would be the exception. Most of the trees would fall within a certain range of heights. They would roughly fit a bell-shaped curve that is symmetrical about the mean. We call this a **normal distribution**.

Many events fit this type of distribution: for example, the heights of 21-year-old males, the results of a national mathematics examination, the weights of newborn babies, etc.

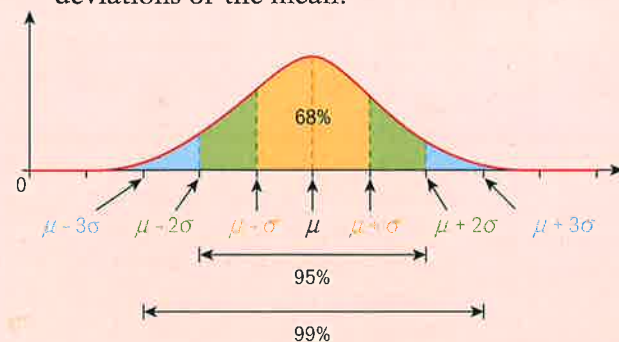


▲ Normal distribution diagram for the tree heights measured by Pedro

The properties of a normal distribution

→ The **normal distribution** is the most important continuous distribution in statistics. It has these properties:

- It is a bell-shaped curve.
- It is symmetrical about the mean, μ . (The mean, the mode and the median all have the same value.)
- The x -axis is an asymptote to the curve.
- The total area under the curve is 1 (or 100%).
- 50% of the area is to the left of the mean and 50% to the right.
- Approximately 68% of the area is within 1 standard deviation, σ , of the mean.
- Approximately 95% of the area is within 2 standard deviations of the mean.
- Approximately 99% of the area is within 3 standard deviations of the mean.

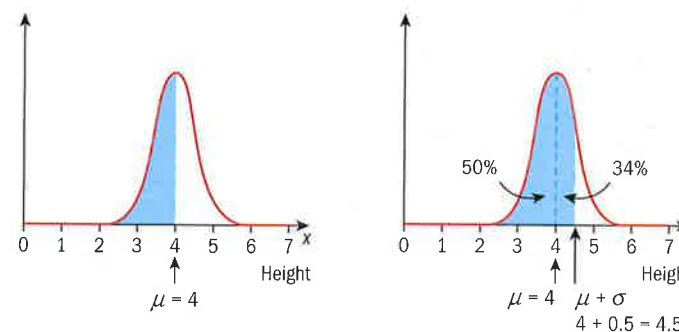


The normal curve is sometimes called the 'Gaussian curve' after the German mathematician Carl Friedrich Gauss (1777–1855). Gauss used the normal curve to analyze astronomical data in 1809. A portrait of Gauss and the normal curve appear on the old German 10 Deutschmark note.

You can calculate the probabilities of events that follow a normal distribution.

Returning to Pedro and the apple trees, imagine that the mean height of the trees is 4 m and the standard deviation is 0.5 m.

Let the height of an apple tree be x .



From the properties of the normal distribution:
Area to left of $\mu = 50\%$.
Area between μ and $\mu + \sigma = 34\%$ ($68\% \div 2$).

The probability that an apple tree is less than 4 m is $P(x < 4) = 50\%$ or 0.5. And $P(x < 4.5) = 50\% + 34\% = 84\%$ or 0.84.

→ The **expected value** is found by multiplying the number in the sample by the probability.

For example, if we chose 100 apple trees at random, the expected number of trees that would be less than 4 m = $100 \times 0.5 = 50$.

Example 1

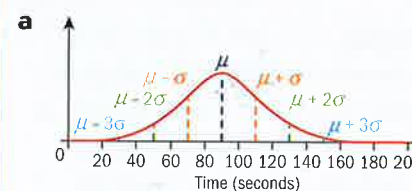
The waiting times for an elevator are normally distributed with a mean of 1.5 minutes and a standard deviation of 20 seconds.

- Sketch a normal distribution diagram to illustrate this information, indicating clearly the mean and the times within one, two and three standard deviations of the mean.
- Find the probability that a person waits longer than 2 minutes 10 seconds for the elevator.
- Find the probability that a person waits less than 1 minute 10 seconds for the elevator.

200 people are observed and the length of time they wait for an elevator is noted.

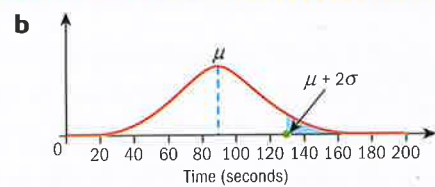
- Calculate the number of people expected to wait less than 50 seconds for the elevator.

Answers

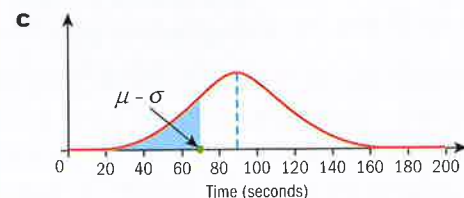


1.5 minutes = 90 seconds
 $\mu = \text{mean} = 90 \text{ seconds}$
 $\sigma = \text{standard deviation} = 20 \text{ seconds}$

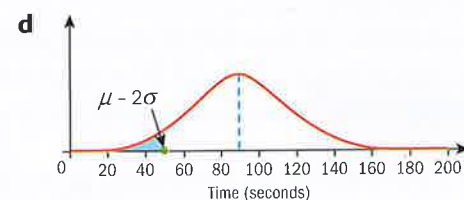
▶ Continued on next page



$P(\text{waiting longer than 2 minutes 10 seconds}) = 2.5\%$, or 0.025 .



$P(\text{waiting less than 1 minute 10 seconds}) = 16\%$, or 0.16 .



$P(\text{waiting less than 50 seconds}) = 2.5\%$, or 0.025
So, the expected number of people
 $= 200 \times 0.025 = 5$.

2 minutes 10 seconds = 130 seconds
Using symmetry about μ :
Area to right of $\mu = 50\%$
Area between μ and $\mu + 2\sigma = 47.5\%$ ($95\% \div 2$)
Area to right of $\mu + 2\sigma = 50\% - 47.5\% = 2.5\%$

1 minute 10 seconds = 70 seconds
Using symmetry about μ :
Area to left of $\mu = 50\%$
Area between μ and $\mu - \sigma = 34\%$ ($68\% \div 2$)
Area to left of $\mu - \sigma = 50\% - 34\% = 16\%$

First find the probability of waiting less than 50 seconds.
Using symmetry about μ :
Area to left of $\mu = 50\%$
Area between μ and $\mu - 2\sigma = 47.5\%$ ($95\% \div 2$)
Area to left of $\mu - 2\sigma = 50\% - 47.5\% = 2.5\%$

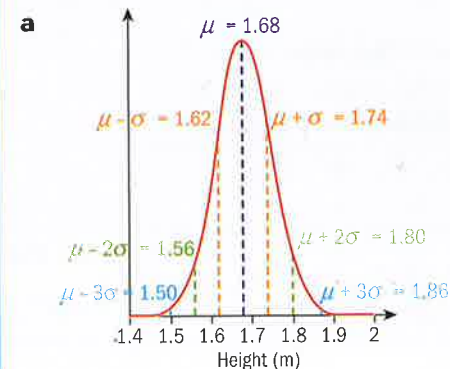
There are 200 people in the sample.

Example 2

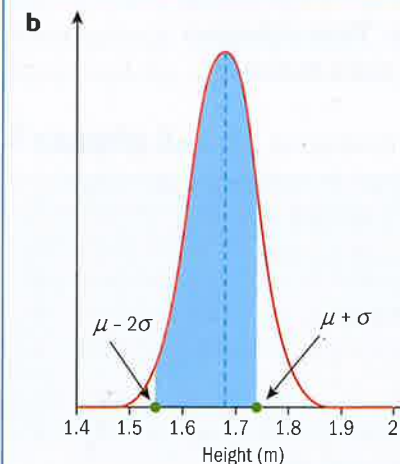
The heights of 250 twenty-year-old women are normally distributed with a mean of 1.68 m and standard deviation of 0.06 m.

- Sketch a normal distribution diagram to illustrate this information, indicating clearly the mean and the heights within one, two and three standard deviations of the mean.
- Find the probability that a woman has a height between 1.56 m and 1.74 m.
- Find the expected number of women with a height greater than 1.8 m.

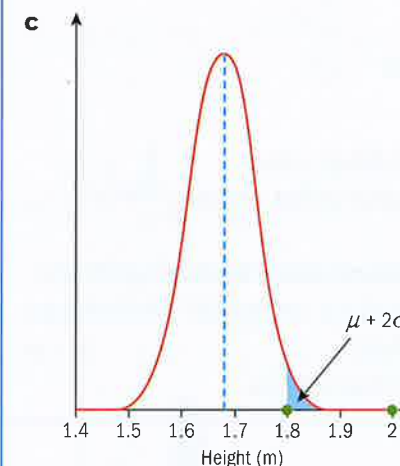
Answers



Let
 $\mu = \text{mean} = 1.68 \text{ m}$
 $\sigma = \text{standard deviation} = 0.06 \text{ m}$



$P(\text{height between 1.56 m and 1.74 m}) = 81.5\%$, or 0.815 .



$P(\text{height greater than 1.8 m}) = 2.5\%$, or 0.025 .
So, the expected number of women
 $= 250 \times 0.025 = 6.25$, or 6 women.

Using symmetry about μ :
Area between μ and $\mu + \sigma = 34\%$ ($68\% \div 2$)
Area between μ and $\mu - 2\sigma = 47.5\%$ ($95\% \div 2$)
Area between 1.56 m and 1.74 m = $34\% + 47.5\% = 81.5\%$

First find the probability of a woman being taller than 1.8 m.
Using symmetry about μ :
Area to right of $\mu = 50\%$
Area between μ and $\mu + 2\sigma = 47.5\%$ ($95\% \div 2$)
Area to right of $\mu + 2\sigma = 50\% - 47.5\% = 2.5\%$

There are 250 women in the sample.

Exercise 5A

EXAM-STYLE QUESTION

- The heights of 200 lilies are normally distributed with a mean of 40 cm and a standard deviation of 3 cm.
 - Sketch a normal distribution diagram to illustrate this information. Indicate clearly the mean and the heights within one, two and three standard deviations of the mean.
 - Find the probability that a lily has a height less than 37 cm.
 - Find the probability that a lily has a height between 37 cm and 46 cm.
 - Find the expected number of lilies with a height greater than 43 cm.

Continued on next page


EXAM-STYLE QUESTIONS

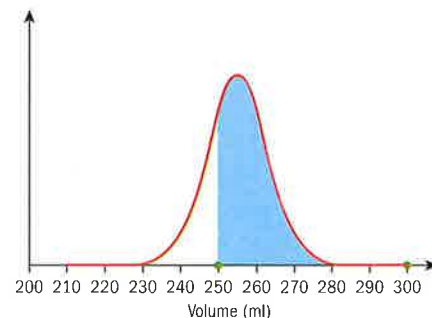
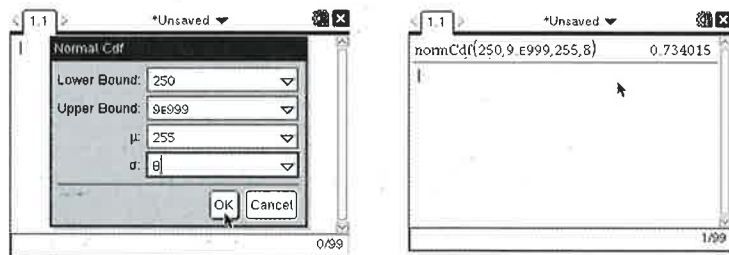
- 2 100 people were asked to estimate the length of one minute. Their estimates were normally distributed with a mean time of 60 seconds and a standard deviation of 4 seconds.
- Sketch a normal distribution diagram to illustrate this information, indicating clearly the mean and the times within one, two and three standard deviations of the mean.
 - Find the percentage of people who estimated between 52 and 64 seconds.
 - Find the expected number of people estimating less than 60 seconds.
- 3 60 students were asked how long it took them to travel to school. Their travel times are normally distributed with a mean of 20 minutes and a standard deviation of 5 minutes.
- Sketch a normal distribution diagram to illustrate this information, indicating clearly the mean and the times within one, two and three standard deviations of the mean.
 - Find the percentage of students who took longer than 25 minutes to travel to school.
 - Find the expected number of students who took between 15 and 25 minutes to travel to school.
- 4 Packets of coconut milk are advertised to contain 250 ml. Akshat tests 75 packets. He finds that the contents are normally distributed with a mean volume of 255 ml and a standard deviation of 8 ml.
- Sketch a normal distribution diagram to illustrate this information, indicating clearly the mean and the volumes within one, two and three standard deviations of the mean.
 - Find the probability that a packet contains less than 239 ml.
 - Find the expected number of packets that contain more than 247 ml.

You can use your GDC to calculate values that are not whole multiples of the standard deviation.


For example, in question 4 of Exercise 5A, suppose we wanted to find the probability that a packet contains more than 250 ml.

First sketch a normal distribution diagram.

In a Calculator page  press MENU 5:Probability | 5:Distributions | 2:Normal Cdf and enter the lower bound (250), the upper bound (9×10^{999} – a very large number), the mean (255) and the standard deviation (8) in the wizard.



To enter 9×10^{999} you need to type 9E999, but you cannot use the E key. Instead, you must use the EE key.

GDC help on CD: Alternative demonstrations for the TI-84 Plus and Casio FX-9860GII GDCs are on the CD. 

So, 73.4% of the packets contain more than 250 ml of coconut milk. Alternatively, enter normCdf, the lowest value, the highest value, the mean and the standard deviation directly into the calculator screen.

For a very small number enter -9×10^{999}



Example 3

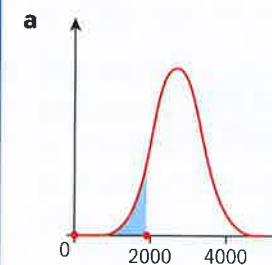
The lifetime of a light bulb is normally distributed with a mean of 2800 hours and a standard deviation of 450 hours.

- Find the percentage of light bulbs that have a lifetime of less than 1950 hours.
- Find the percentage of light bulbs that have a lifetime between 2300 and 3500 hours.
- Find the probability that a light bulb has a lifetime of more than 3800 hours.

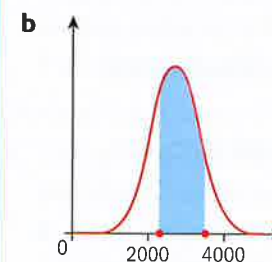
120 light bulbs are tested.

- Find the expected number of light bulbs with a lifetime of less than 2000 hours.

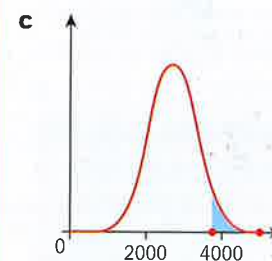
Answers



2.95% of the light bulbs have a lifetime of less than 1950 hours.



80.7% of the light bulbs have a lifetime between 2300 and 3500 hours.



Only 1.31% of the light bulbs have a lifetime of more than 3800 hours.

$\mu = \text{mean} = 2800 \text{ hours}$
 $\sigma = \text{standard deviation} = 450 \text{ hours}$


Lifetime less than 1950 hours:
 lower bound = -9×10^{999}
 upper bound = 1950

From GDC:
 $\text{normCdf}(-9\text{E}999, 1950, 2800, 450) = 0.02945 = 2.95\%$

Lifetime between 2300 and 3500 hours:
 lower bound = 2300
 upper bound = 3500

From GDC:
 $\text{normCdf}(2300, 3500, 2800, 450) = 0.8068 = 80.7\%$

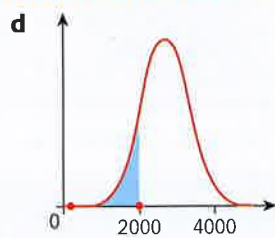
Lifetime more than 3800 hours:
 lower bound = 3800
 upper bound = 9×10^{999}

GDC help on CD: Alternative demonstrations for the TI-84 Plus and Casio FX-9860GII GDCs are on the CD. 

From GDC:
 $\text{normCdf}(3800, 9\text{E}999, 2800, 450) = 0.0131 = 1.31\%$

Remember not to use $-9\text{E}999$ notations in an exam.

▶ Continued on next page



$P(\text{lifetime less than 2000 hours}) = 3.77\%$
 Expected number = 120×0.0377
 $= 4.524$

So, you would expect 4 or 5 light bulbs to have a lifetime of less than 2000 hours.

First find $P(\text{lifetime less than 2000 hours})$:
 lower bound = -9×10^{999}
 upper bound = 2000

From GDC:
 $\text{normCdf}(-9E999, 2000, 2800, 450) = 0.0377 = 3.77\%$
 120 light bulbs are tested.

Exercise 5B

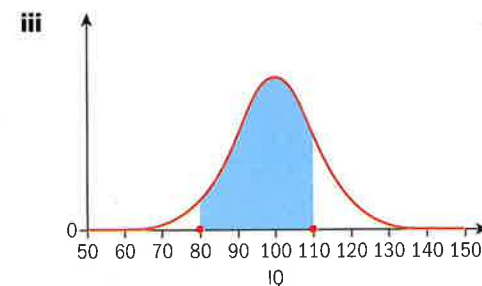
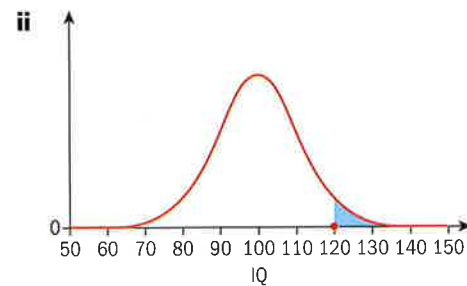
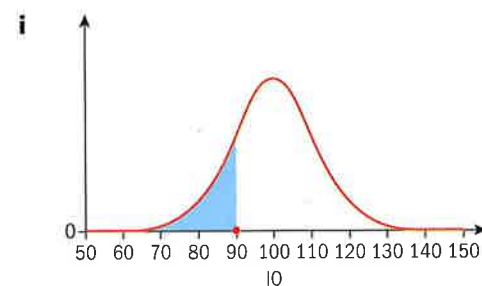
EXAM-STYLE QUESTION

- Jordi delivers daily papers to a number of homes in a village. The time taken to deliver the papers follows a normal distribution with mean 80 minutes and standard deviation 7 minutes.
 - Sketch a normal distribution diagram to illustrate this information.
 - Find the probability that Jordi takes longer than 90 minutes to deliver the papers.

Jordi delivers papers every day of the year (365 days).

 - Calculate the expected number of days on which it would take Jordi longer than 90 minutes to deliver the papers.

- A set of 2000 IQ scores is normally distributed with a mean of 100 and a standard deviation of 10.
 - Calculate the probability that is represented by each of the following diagrams.



Lambert Quételet (1796–1874), a Flemish scientist, was the first to apply the normal distribution to human characteristics. He noticed that measures such as height, weight and IQ were normally distributed.

- Find the expected number of people with an IQ of more than 115.



- A machine produces washers whose diameters are normally distributed with a mean of 40 mm and a standard deviation of 2 mm.
 - Find the probability that a washer has a diameter less than 37 mm.
 - Find the probability that a washer has a diameter greater than 45 mm.

Every week 300 washers are tested.

 - Calculate the expected number of washers that have a diameter between 35 mm and 43 mm.



EXAM-STYLE QUESTIONS

- In a certain school, the monthly incomes of members of staff are normally distributed with a mean of 2500 euros and a standard deviation of 400 euros.
 - Sketch a normal distribution diagram to illustrate this information.
 - Find the probability that a member of staff earns less than 1800 euros per month.

The school has 80 members of staff.

 - Calculate the expected number of staff who earn more than 3400 euros.
- The lengths of courgettes are normally distributed with a mean of 16 cm and a standard deviation of 0.8 cm.
 - Find the percentage of courgettes that have a length between 15 cm and 17 cm.
 - Find the probability that a courgette is longer than 18 cm.

The lengths of 100 courgettes are measured.

 - Calculate the expected number of courgettes that have a length less than 14.5 cm.



- At a market, the weights of bags of kiwi fruit are normally distributed with a mean of 500 g and a standard deviation of 8 g. A man picks up a bag of kiwi fruit at random. Find the probability that the bag weighs more than 510 g.



EXAM-STYLE QUESTIONS

- The scores in a Physics test follow a normal distribution with mean 70% and standard deviation 8%.
 - Find the percentage of students who scored between 55% and 80%.

30 students took the physics test.

 - Calculate the expected number of students who scored more than 85%.
- A machine produces pipes such that the length of each pipe is normally distributed with a mean of 1.78 m and a standard deviation of 2 cm. Any pipe whose length is greater than 1.83 m is rejected.
 - Find the probability that a pipe will be rejected.

500 pipes are tested.

 - Calculate the expected number of pipes that will be rejected.

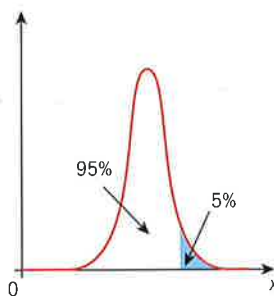
Inverse normal calculations

Sometimes you are given the percentage area under the curve, i.e. the probability or proportion, and are asked to find the value corresponding to it. This is called an inverse normal calculation.

Always make a sketch to illustrate the information given.

You must always remember to use the area to the **left** when using your GDC. If you are given the area to the **right** of the value, you must subtract this from 1 (or 100%) before using your GDC.

For example, an area of 5% above a certain value means there is an area of 95% below it.



In examinations, inverse normal questions will not involve finding the mean or standard deviation.



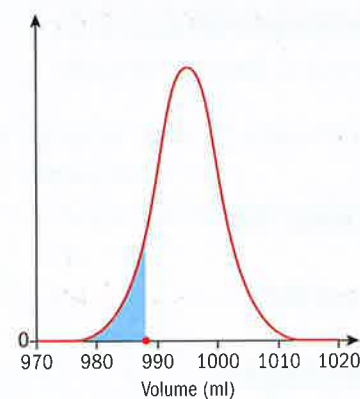
Example 4

The volume of cartons of milk is normally distributed with a mean of 995 ml and a standard deviation of 5 ml.

It is known that 10% of the cartons have a volume less than x ml.

Find the value of x .

Answer



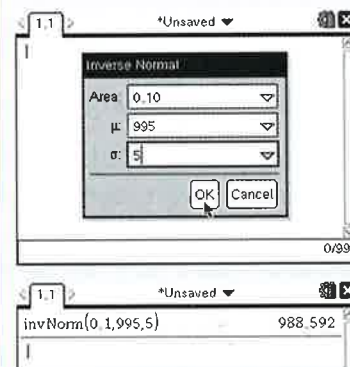
$x = 989$ ml (to 3 sf)

First sketch a diagram. The shaded area represents 10% of the cartons.

Using the GDC:

In a Calculator page press MENU 5:Probability | 5:Distributions | 3:Inverse Normal...

Enter the percentage given (as a decimal, 0.1), the mean (995) and the standard deviation (5).



$x = 989$ (3 sf)

$x = 989$ ml means that 10% of the cartons have a volume less than 989 ml.

GDC help on CD: Alternative demonstrations for the TI-84 Plus and Casio FX-9860GII GDCs are on the CD.



Example 5

The weights of pears are normally distributed with a mean of 110g and a standard deviation of 8g.

a Find the percentage of pears that weigh between 100g and 130g.

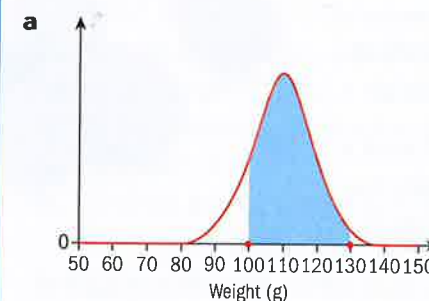
It is known that 8% of the pears weigh more than m g.

b Find the value of m .

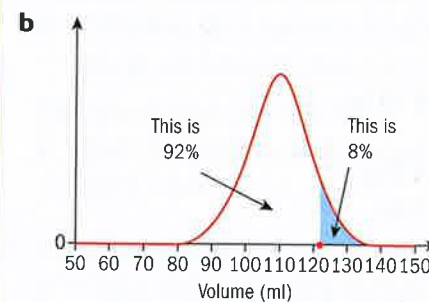
250 pears are weighed.

c Calculate the expected number of pears that weigh less than 105g.

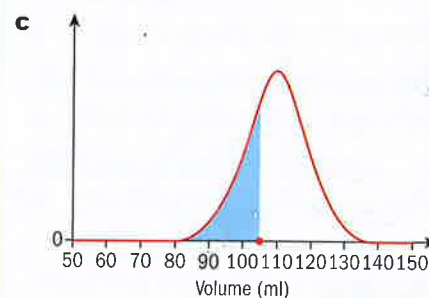
Answers



88.8% of the pears weigh between 100g and 130g.



$m = 121$ g



$P(\text{weight less than } 105\text{g}) = 0.266$
 Expected number = $250 \times 0.266 = 66.5$
 So, you would expect 66 or 67 pears to weigh less than 105g.

Sketch a diagram.

$\mu = \text{mean} = 110\text{g}$

$\sigma = \text{standard deviation} = 8\text{g}$

Weight between 100g and 130g:

lower bound = 100

upper bound = 130

From GDC:

$\text{normCdf}(100, 130, 110, 8) = 0.888 = 88.8\%$

8% weighing more than m g is the same as saying that 92% weigh less than m g.

From GDC:

$\text{invNorm}(0.92, 110, 8) = 121$

$m = 121$ g means that 8% of the pears weigh more than 121g.

Weight less than 105g:

lower bound = -9×10^{999}

upper bound = 105

From GDC:

$\text{normCdf}(-9E999, 105, 110, 8) = 0.266$

250 pears are weighed.



Exercise 5C

- The mass of coffee grounds in Super-strength coffee bags is normally distributed with a mean of 5 g and a standard deviation of 0.1 g. It is known that 25% of the coffee bags weigh less than p grams. Find the value of p .
- The heights of Dutch men are normally distributed with a mean of 181 cm and a standard deviation of 5 cm. It is known that 35% of Dutch men have a height less than h cm. Find the value of h .
- The weight of kumquats is normally distributed with a mean of 20 g and a standard deviation of 0.8 g. It is known that 15% of the kumquats weigh more than k grams. Find the value of k .
- The weight of cans of sweetcorn is normally distributed with a mean of 220 g and a standard deviation of 4 g. It is known that 30% of the cans weigh more than w grams. Find the value of w .



EXAM-STYLE QUESTIONS

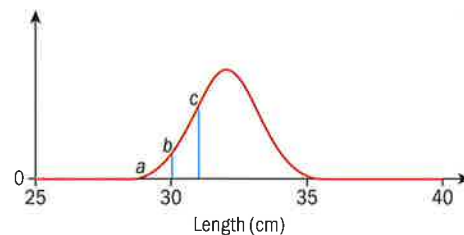
- The weights of cats are normally distributed with a mean of 4.23 kg and a standard deviation of 0.76 kg.
 - Write down the weights of the cats that are within one standard deviation of the mean.

A vet weighs 180 cats.

 - Find the number of these cats that would be expected to be within one standard deviation of the mean.
 - Calculate the probability that a cat weighs less than 3.1 kg.
 - Calculate the percentage of cats that weigh between 3 kg and 5.35 kg.

It is known that 5% of the cats weigh more than w kg.

- Find the value of w .
- A manufacturer makes drumsticks with a mean length of 32 cm. The lengths are normally distributed with a standard deviation of 1 cm.
 - Calculate the values of a , b and c shown on the graph.
 - Find the probability that a drumstick has a length greater than 30.6 cm.



It is known that 80% of the drumsticks have a length less than d cm.

- Find the value of d .
- One week 5000 drumsticks are tested.
- Calculate the expected number of drumsticks that have a length between 30.5 cm and 32.5 cm.



- The average lifespan of a television set is normally distributed with a mean of 8000 hours and a standard deviation of 1800 hours.
 - Find the probability that a television set will break down before 2000 hours.
 - Find the probability that a television set lasts between 6000 and 12 000 hours.
 - It is known that 12% of the television sets break down before t hours. Find the value of t .



EXAM-STYLE QUESTIONS

- The speed of cars on a motorway is normally distributed with a mean of 120 km h^{-1} and a standard deviation of 10 km h^{-1} .
 - Draw a normal distribution diagram to illustrate this information.
 - Find the percentage of cars that are traveling at speeds of between 105 km h^{-1} and 125 km h^{-1} .

It is known that 8% of the cars are traveling at a speed of less than $p \text{ km h}^{-1}$.

- Find the value of p .

One day 800 cars are checked for their speed.

- Calculate the expected number of cars that will be traveling at speeds of between 96 km h^{-1} and 134 km h^{-1} .

The speed limit is 130 km h^{-1} .

- Find the number of cars that are expected to be exceeding the speed limit.

- The weights of bags of rice are normally distributed with a mean of 1003 g and a standard deviation of 2 g.

- Draw a normal distribution diagram to illustrate this information.
- Find the probability that a bag of rice weighs less than 999 g.

The manufacturer states that the bags of rice weigh 1 kg.

- Find the probability that a bag of rice is underweight.

400 bags of rice are weighed.

- Calculate the expected number of bags of rice that are underweight.

5% of the bags of rice weigh more than p g.

- Find the value of p .





EXAM-STYLE QUESTION

10 The weights of babies are normally distributed with a mean of 3.8 kg and a standard deviation of 0.5 kg.

a Find the percentage of babies who weigh less than 2.5 kg.

In a space of 15 minutes two babies are born. One weighs 2.34 kg and the other weighs 5.5 kg.

b Calculate which event is more likely to happen.

One month 300 babies are weighed.

c Calculate the number of babies expected to weigh more than 4.5 kg.

It was found that 10% of the babies weighed less than w kg.

d Find the value of w .

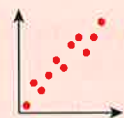
5.2 Correlation

When two sets of data appear to be connected, that is, one set of data is dependent on the other, then there are various methods that can be used to check whether or not there is any **correlation**. One of these methods is the scatter diagram.

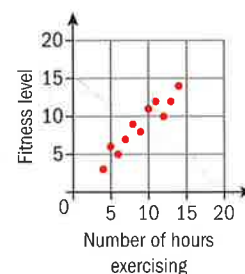
Data can be plotted on a scatter diagram with the **independent variable** on the horizontal axis and the **dependent variable** on the vertical axis. The pattern of dots will give a visual picture of how closely, if at all, the variables are related.

Types of correlation

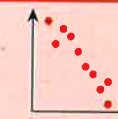
→ In a **positive** correlation the dependent variable increases as the independent variable increases.



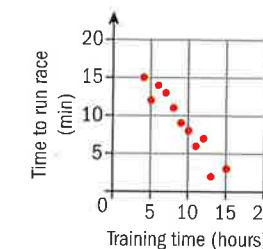
For example, fitness levels (dependent variable) increase as the number of hours spent exercising (independent variable) increase:



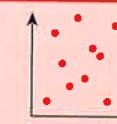
→ In a **negative** correlation the dependent variable decreases as the independent variable increases.



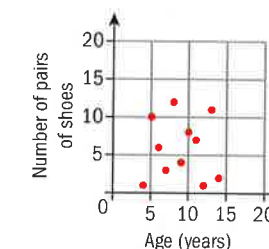
For example, the time taken to run a race (dependent variable) decreases as the training time (independent variable) increases:



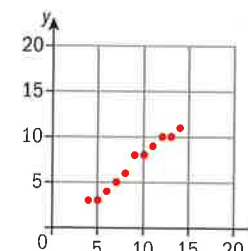
→ When the points are scattered randomly across the diagram there is **no** correlation.



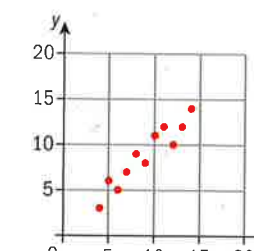
For example, the number of pairs of shoes that a person owns is not related to their age:



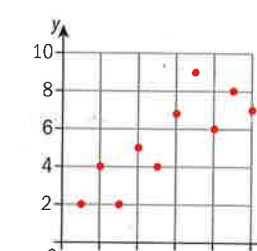
→ Correlations can also be described as strong, moderate or weak.



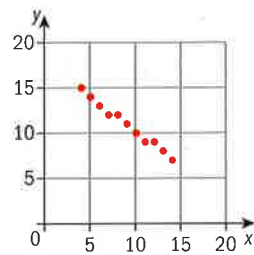
This is an example of a **strong positive** correlation.



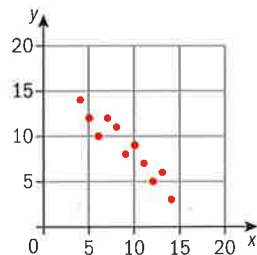
This is an example of a **moderate positive** correlation.



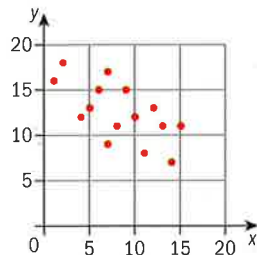
This is an example of a **weak positive** correlation.



This is an example of a **strong negative** correlation.

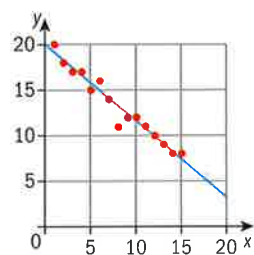


This is an example of a **moderate negative** correlation.

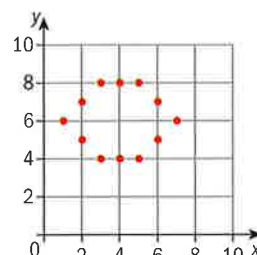


This is an example of a **weak negative** correlation.

Correlations can also be classed as linear or non-linear.



This is an example of a **linear** correlation.



This is an example of a **non-linear** correlation.

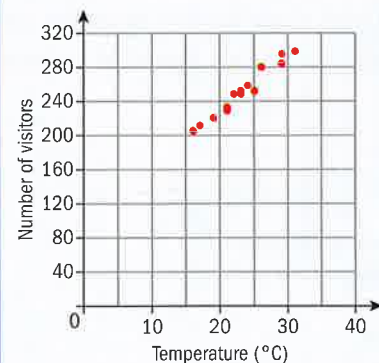
For Mathematical Studies, you will only need to learn about **linear** correlations. However you may use other types of correlation in your project.

Example 6

The manager of a recreation park thought that the number of visitors to the park was dependent on the temperature. He kept a record of the temperature and the numbers of visitors over a two-week period. Plot these points on a scatter diagram and comment on the type of correlation.

Temperature (°C)	16	22	31	19	23	26	21	17	24	29	21	25	23	29
Number of visitors	205	248	298	223	252	280	233	211	258	295	229	252	248	284

Answer



There is a **strong positive** correlation between temperature and the number of visitors to the park.

Draw the *x*-axis 'Temperature (°C)' from 0 to 40 and the *y*-axis 'Number of visitors' from 0 to 320.

Plot the points.

Describe the correlation.

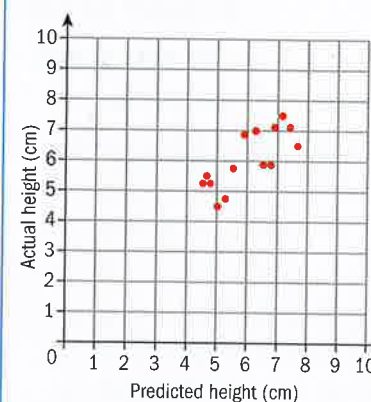
Example 7

A Mathematical Studies student wanted to check if there was a correlation between the predicted heights of daisies and their actual heights.

Draw a scatter diagram to illustrate the data and comment on the correlation.

Predicted height (cm)	5.3	6.2	4.9	5.0	4.8	6.6	7.3	7.5	6.8	5.5	4.7	6.8	5.9	7.1
Actual height (cm)	4.7	7.0	5.3	4.5	5.6	5.9	7.2	6.5	7.2	5.8	5.3	5.9	6.8	7.6

Answer



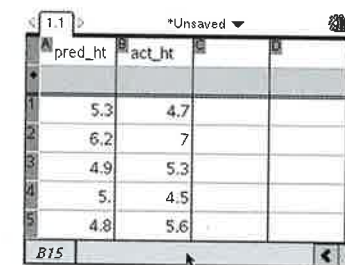
There is a **moderate positive** correlation between predicted height and actual height.

Draw *x*- and *y*-axes from 0 to 10.

Plot 'Predicted height' on the horizontal axis and 'Actual height' on the vertical axis.

Describe the correlation.

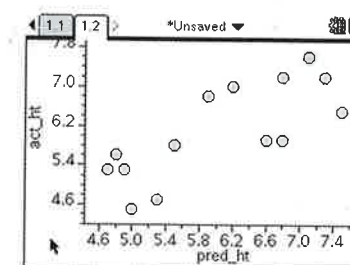
You can also use a GDC to plot a scatter diagram. For Example 7:



GDC help on CD: Alternative demonstrations for the TI-84 Plus and Casio FX-9860GII GDCs are on the CD.

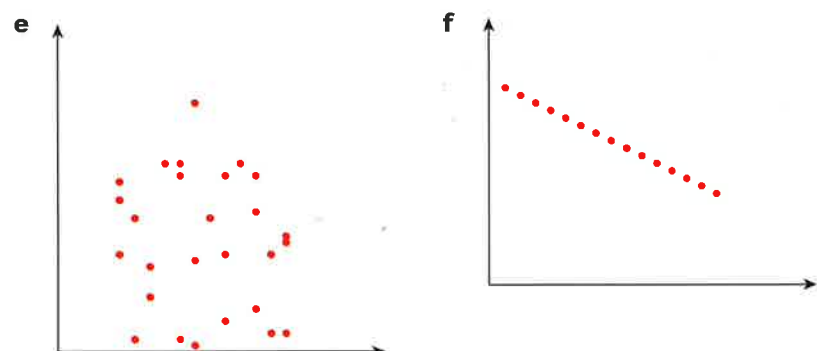
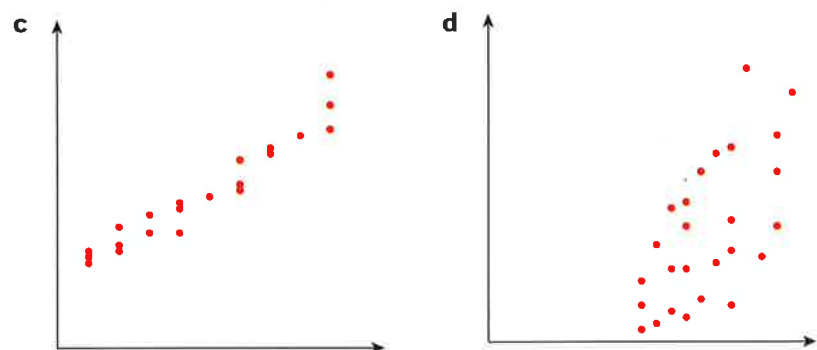
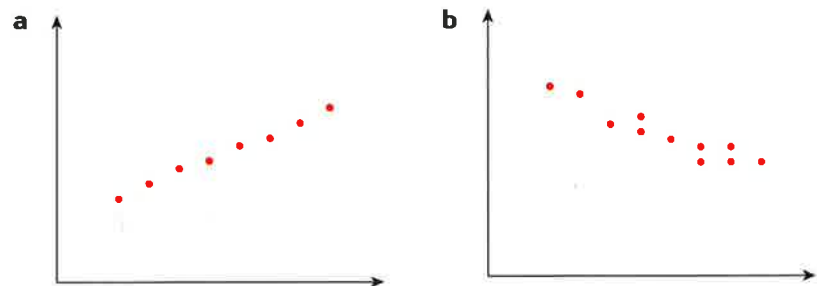
First enter the data in two lists on a List & Spreadsheet page

Then enter the variables onto the axes on a Data and Statistics page to draw the scatter diagram.

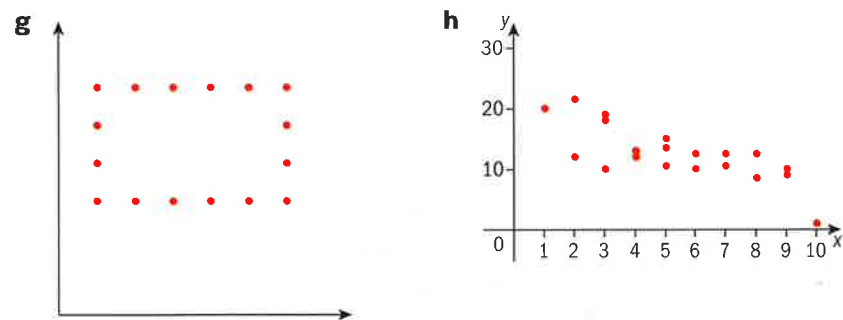


Exercise 5D

1 For each diagram, state the type of correlation (positive/negative and linear/non-linear) and the strength of the relationship (perfect/strong/moderate/weak/none).



A **perfect correlation** is one where all the plotted points lie on a straight line.



2 For each set of data, plot the points on a scatter diagram and describe the type of correlation.

x	28	30	25	35	19	38	25	33	41	22	35	44
y	24	36	30	40	15	34	28	34	44	23	37	45

x	3	7	7	11	16	15	17	17	18	20
y	16	11	12	9	6	7	3	9	5	6

Line of best fit

A **line of best fit** is a line that is drawn on a graph of two sets of data, so that approximately as many points lie above the line as below it.

- To draw the **line of best fit** by eye:
 - Find the mean of each set of data and plot this point on your scatter diagram.
 - Draw a line that passes through the mean point and is close to all the other points – with approximately an equal number of points above and below the line.

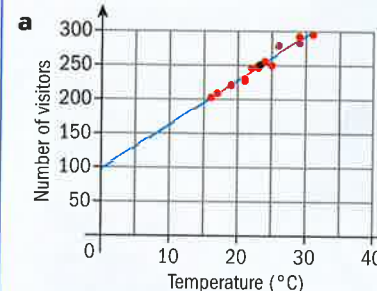


The line of best fit does not need to go through the origin and, in fact, in most cases it will not go through the origin.

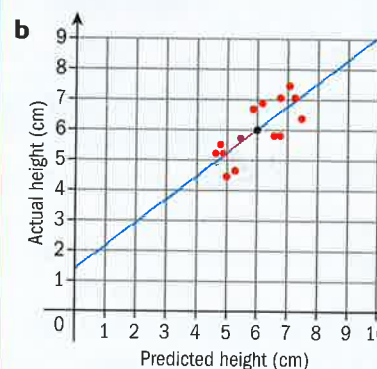
Example 8

- a For Example 6 draw the line of best fit on the diagram.
- b For Example 7 draw the line of best fit on the diagram.

Answers



Calculate the means using your GDC. The mean temperature is 23.3, and the mean number of visitors is 251. Plot the mean point (23.3, 251) on the scatter diagram. Draw a line of best fit through the mean point so that there are roughly an equal number of points above and below the line.



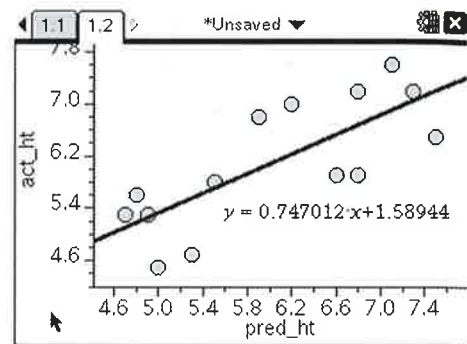
The mean predicted height is 6.03, and the mean actual height is 6.09. Plot the mean point (6.03, 6.09) on the scatter diagram and draw a straight line through it so that there are roughly an equal number of points above and below the line.

Geosciences use a line of best fit in

- flood frequency curves
- earthquake forecasting
- meteorite impact prediction
- climate change.

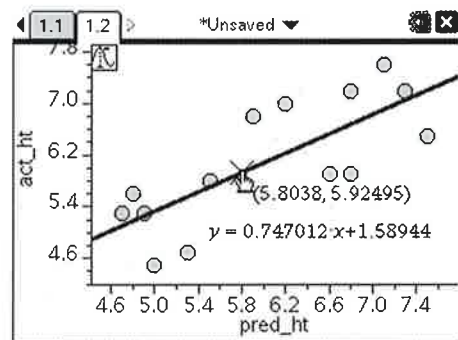
You can also use a GDC to draw a line of best fit.
For Example 7:
Select MENU 4:Analyze | 6:Regression | 2:Show Linear ($ax + b$).

Given a value of predicted height, use trace (MENU 4:Analyze | A:Graph Trace) to find the value of the actual height from the graph.



In the Data & Statistics mode it is not possible to find exact values when using trace

GDC help on CD: Alternative demonstrations for the TI-84 Plus and Casio FX-9860GII GDCs are on the CD.



There is often a lot of confusion between the concepts of *causation* and *correlation*. However, they should be easy enough to distinguish. One action can *cause* another (such as smoking can cause lung cancer), or it can *correlate* with another (for example, blue eyes are correlated with blonde hair).
If one action *causes* another, then they are also *correlated*. But if two things are *correlated* it does *not* mean that one *causes* the other. For example, there could be a strong *correlation* between the predicted grades that teachers give and the actual grades that the students achieve. However, the achieved grades are not *caused* by the predicted grades.
Can you think of other examples?
Can you find articles in newspapers, magazines or online where *cause* is used incorrectly?

Exercise 5E

- 1 For each set of data:
 - i Plot the points on a scatter diagram and describe the type of correlation.
 - ii Find the mean of x and the mean of y .
 - iii Plot the mean point on your diagram and draw a line of best fit by eye.

a

x	2	4	6	8	10	12	14	16	18	20	22	24
y	14	15	18	21	24	25	27	29	30	32	35	39

b

x	12	13	14	15	16	17	18	19	20	21
y	32	29	30	25	22	22	15	10	10	7

EXAM-STYLE QUESTIONS

- 2 The following table gives the heights and weights of 12 giraffes.

Height (xm)	4.8	4.1	4.2	4.7	5.0	5.0	4.8	5.2	5.3	4.3	5.5	4.5
Weight (ykg)	900	600	650	750	1100	950	850	1150	1100	650	1250	800

- a Plot the points on a scatter diagram and describe the correlation.
- b Find the mean height and the mean weight.
- c Plot the mean point on your diagram and draw a line of best fit by eye.
- d Use your diagram to estimate the weight of a giraffe of height 4.6 m.



- 3 Fourteen students took a test in Chemistry and ITGS (Information Technology in a Global Society). The results are shown in the following table.

Chemistry (%)	45	67	72	34	88	91	56	39	77	59	66	82	96	42
ITGS (%)	42	76	59	44	76	88	55	45	69	62	58	94	85	58

- a Plot the points on a scatter diagram and describe the correlation.
 - b Find the mean score for each test.
 - c Plot the mean point on your diagram and draw a line of best fit by eye.
 - d Use your diagram to estimate the result for an ITGS test when the chemistry score was 50%.
- 4 Twelve mothers were asked how many hours per day, on average, they held their babies and how many hours per day, on average, the baby cried. The results are given in the following table.

Baby held (hours)	1	2	3	3	4	4	5	6	6	7	8	9
Baby cried (hours)	6	6	5	5.5	4	3	3.5	2	2.5	2	1.5	1

- a Plot the points on a scatter diagram and describe the correlation.
- b Find the mean number of hours held and the mean number of hours spent crying.
- c Plot the mean point on your diagram and draw a line of best fit by eye.
- d Use your diagram to estimate the number of hours a baby cries if it is held for 3.5 hours.

EXAM-STYLE QUESTION

- 5 The table shows the size of a television screen and the cost of the television.

Size (inches)	32	37	40	46	50	55	59
Cost (\$)	450	550	700	1000	1200	1800	2000

- Plot the points on a scatter diagram and describe the correlation.
- Find the mean screen size and the mean cost.
- Plot the mean point on your diagram and draw a line of best fit by eye.
- Use your diagram to estimate the cost of a 52-inch TV.

Pearson's product-moment correlation coefficient

Karl Pearson (1857–1936) was an English lawyer, mathematician and statistician.

His contributions to the field of statistics include the product-moment correlation coefficient and the chi-squared test.

Pearson's career was spent largely on applying statistics to the field of biology.

He founded the world's first University statistics department at University College London in 1911.

► Karl Pearson



It is useful to know the **strength** of the relationship between any two sets of data that are thought to be related.

Pearson's product-moment correlation coefficient, r , is one way of finding a numerical value that can be used to determine the strength of a linear correlation between the two sets of data.

- **Pearson's product-moment correlation coefficient, r** , can take all values between -1 and $+1$ inclusive.
- When $r = -1$, there is a **perfect negative** correlation between the data sets.
 - When $r = 0$, there is **no** correlation.
 - When $r = +1$, there is a **perfect positive** correlation between the data sets.
 - A **perfect correlation** is one where **all** the plotted points lie on a straight line.

When r is between

- 0 and 0.25, the correlation is very weak
- 0.25 and 0.5, the correlation is weak
- 0.5 and 0.75, there is a moderate correlation
- 0.75 and 1, the correlation is strong.

In examinations you will only be expected to use your GDC to find the value of r .

The formula for Pearson's product-moment correlation coefficient for two sets of data, x and y , is: $r = \frac{s_{xy}}{s_x s_y}$

where s_{xy} is the covariance (beyond the scope of this course) and s_x and s_y are the standard deviations of x and y respectively.

You will be expected to use this formula to enhance your project.

Other formulae that you will need are:

$$s_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n} \text{ or } \frac{\sum xy}{n} - \frac{\sum x}{n} \frac{\sum y}{n}$$

$$s_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} \text{ or } \sqrt{\left(\frac{\sum x^2}{n} - \bar{x}^2\right)} \quad s_y = \sqrt{\frac{\sum (y - \bar{y})^2}{n}} \text{ or } \sqrt{\left(\frac{\sum y^2}{n} - \bar{y}^2\right)}$$



Example 9

The data given below for a first-division football league show the position of the team and the number of goals scored.

Find the correlation coefficient, r , and comment on this value.

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Goals	75	68	60	49	59	50	55	46	57	49	48	39	44	56	54	37	42	37	40	27

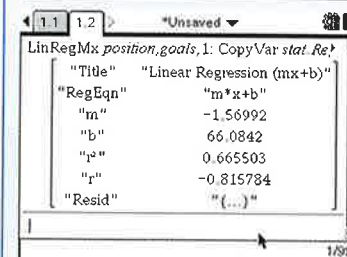
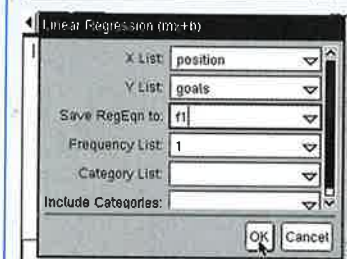
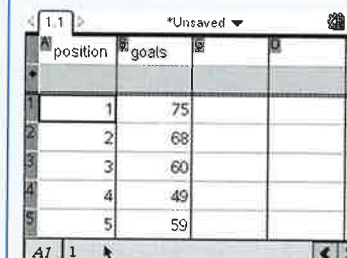
Answer

$r = -0.816$ (to 3 sf)

So, there is a **strong negative** correlation between the position of the team and the number of goals scored.

Using a GDC:

First enter 'Position' numbers and 'Goals' into two lists (X and Y respectively).



GDC help on CD: *Alternative demonstrations for the TI-84 Plus and Casio FX-9860GII GDCs are on the CD.*



Your GDC also gives r^2 , **the coefficient of determination**. This is an indication of how much of the variation in one set of data, y , can be explained by the variation in the other set of data, x . For example, if $r^2 = 0.821$, this means that 82.1% of the variation in set y is caused by the variation in set x . Here, either $r = 0.906$ which is a strong positive linear correlation, or $r = -0.906$ which is a strong negative linear correlation.



Example 10

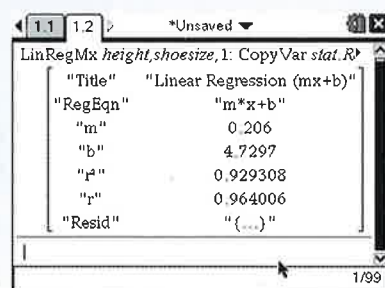
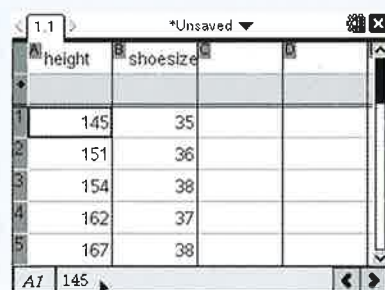
The heights and shoe sizes of the students at Learnwell Academy are given in the table below. Find the correlation coefficient, r , and comment on your result.

Height (x cm)	145	151	154	162	167	173	178	181	183	189	193	198
Shoe size	35	36	38	37	38	39	41	43	42	45	44	46

Answer

$r = 0.964$ (to 3 sf)

This means that there is a **strong positive** correlation between height and shoe size.



GDC help on CD: Alternative demonstrations for the TI-84 Plus and Casio FX-9850GII GDCs are on the CD.



Exercise 5F

- 1 The table gives the temperature ($^{\circ}\text{C}$) at midday and the number of ice creams sold over a period of 21 days.

Temperature ($^{\circ}\text{C}$)	22	23	22	19	20	25	23	20	17	18	23	24	22	26	19	19	20	22	23	22	20
Number of ice creams sold	59	61	55	40	51	72	55	45	39	35	59	72	63	77	37	41	44	50	59	48	38

Find the correlation coefficient, r , and comment on this value.

- 2 A chicken farmer selected a sample of 12 hens. During a two-week period, he recorded the number of eggs each hen produced and the amount of feed each hen ate. The results are given in the table.

Number of eggs	Units of feed eaten
11	6.2
10	4.9
13	7.1
10	6.2
11	5.0
15	7.9
9	4.8
12	6.9
11	5.3
12	5.9
13	6.5
9	4.5



- a Find the correlation coefficient, r .
b Comment on the value of the correlation coefficient.

- 3 The table gives the average temperature for each week in December, January and February and the corresponding number of hours that an average family used their central heating.

Average temperature ($^{\circ}\text{C}$)	4	1	3	-2	-9	-12	-8	-9	-2	1	3	5
Hours of heating	43	45	51	52	58	64	57	60	55	43	40	30

Find the correlation coefficient, r , and comment on this value.

- 4 Eight students complete examination papers in Economics and Biology. The results are shown in the table.

Student	A	B	C	D	E	F	G	H
Economics	64	55	43	84	67	49	92	31
Biology	53	42	44	79	75	52	84	29

Find the correlation coefficient, r , and comment on your result.

- 5 The table shows the age of a baby, measured in days, and the weight, in kilograms, at 08:00 on the corresponding day.

Age (days)	0	7	14	21	28	35	42
Weight (kg)	3.50	3.75	3.89	4.15	4.42	4.55	5.02

Find the correlation coefficient, r , and comment on your result.

- 6 The heights and weights of 10 students selected at random are shown in the table.

Height (xcm)	155	161	173	150	182	165	170	185	175	145
Weight (ykg)	50	75	80	46	81	79	64	92	74	108

Find the correlation coefficient, r , and comment on your answer.

- 7 The table shows the mock examination results and the actual results of 15 students at Top High College.

Mock	32	35	28	24	19	39	44	41	23	29	28	35	38	43	21
Actual	33	34	30	25	18	36	43	42	24	27	29	36	39	44	22

Find the correlation coefficient, r , and comment on your result.

- 8 The ages of 14 people and the times it took them to run 1 km are shown in the table.

Age (years)	9	12	13	15	16	19	21	29	32	43	48	55	61	66
Time (minutes)	7.5	6.8	7.2	5.3	5.1	4.9	5.2	4.6	4.9	6.8	6.2	7.5	8.9	9.2

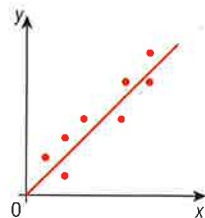
Find the correlation coefficient, r , and comment on your result.

5.3 The regression line

→ The **regression line for y on x** is a more accurate version of a line of best fit, compared to best fit by eye.

The regression line for y on x , where y is the dependent variable, is also known as the least squares regression line. It is the line drawn through a set of points such that the sum of the squares of the distance of each point from the line is a minimum.

→ If there is a strong or moderate correlation, you can use the regression line for y on x to predict values of y for values of x within the range of the data.



You should only calculate the equation of the regression line if there is a moderate or strong correlation coefficient.

In your project you can work out the equation of the regression line for y on x using the formula:

$$(y - \bar{y}) = \frac{s_{xy}}{(s_x)^2} (x - \bar{x})$$

where \bar{x} and \bar{y} are the means of the x and y data values respectively, s_x is the standard deviation for the x data values, and s_{xy} is the covariance.

In examinations you will only be expected to use your GDC to find the equation of the regression line.



Example 11

Ten students train for a charity walk.

The table shows the average number of hours per week that each member trains and the time taken to complete the walk.

Training time (hours)	9	8	12	3	25	6	10	5	6	21
Time to complete walk (minutes)	15.9	14.8	15.3	18.4	13.8	16.2	14.1	16.1	16	14.2

- Find the correlation coefficient, r .
- Find the equation of the regression line.
- Using your equation, estimate how many minutes it will take a student who trains 18 hours per week to complete the walk.

British scientist and mathematician Francis Galton (1822–1911) coined the term 'regression'.

Answers

- a $r = -0.767$ (to 3 sf)

First enter the data into two lists and then compute the results.

- b The equation of the regression line is:
 $y = -0.147x + 17.0$

The general form of the equation is:

$$y = mx + c$$

From the GDC:

$$m = -0.147 \text{ (to 3 sf)}$$

$$c = 17.0 \text{ (to 3 sf)}$$

- c $y = -0.147(18) + 17.0 = 14.4$
(to 3 sf)

Substitute 18 (hours) for x in the equation from part b.

Therefore, the time taken is approximately 14.4 minutes.

GDC help on CD: Alternative demonstrations for the TI-84 Plus and Casio FX-9860GII GDCs are on the CD.

In this book we use $y = mx + c$ as the general form of a linear equation. The GDC uses $y = mx + b$ as the general form. Some people use $y = ax + b$.



Example 12

The table shows the number of mice for sale in a pet shop at the end of certain weeks.

Time (x weeks)	3	5	6	9	11	13
Number of mice (y)	41	57	61	73	80	91

- Find the correlation coefficient, r .
- Find the equation of the regression line for y on x .
- Use your regression line to predict the number of mice for sale after 10 weeks.
- Can you accurately predict the number of mice after 20 weeks?

Answers

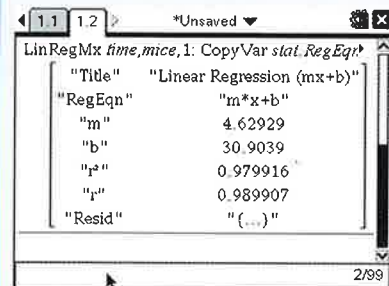
a $r = 0.990$ (to 3 sf)

b The equation of the regression line is:
 $y = 4.63x + 30.9$

c $y = 4.63(10) + 30.9 = 77.2 = 77$
After 10 weeks, the number of mice is 77.

d No, because it is too far away from the data in the table.

First enter the data into two lists.



The general form of the equation is:
 $y = mx + c$
From the GDC:
 $m = 4.63$ (to 3 sf)
 $c = 30.9$ (to 3 sf)

Substitute 10 (weeks) for x in the equation from part b.

How do we know what we know? How sure can we be of our predictions? What predictions are made about population, or the climate?

GDC help on CD: Alternative demonstrations for the TI-84 Plus and Casio FX-9860GII GDCs are on the CD.

Remember that you cannot use the regression line to predict values beyond the region of the given data.



Exercise 5G

EXAM-STYLE QUESTION

- 1 The table shows the distance travelled by train between various places in India and the cost of the journey.

Distance (km)	204	1407	1461	793	1542	343	663	780
Cost (rupees)	390	2200	2270	1390	2280	490	1200	1272

- Find the correlation coefficient, r , and comment on your result.
- Find the equation of the regression line.
- Use your equation to estimate the cost of a 1000 km train journey.



EXAM-STYLE QUESTIONS

- 2 Different weights were attached to a vertical spring and the length of the spring measured. The results are shown in the table.

Load (x kg)	0	2	3	5	6	7	9	11
Length (y cm)	15	16.5	17.5	18.5	18.8	19.2	20	20.4

- Find the correlation coefficient, r .
 - Find the equation of the regression line.
 - Use your equation to estimate the length of the spring when a weight of 8 kg is added.
- 3 Lijn is a keen swimmer. For his Mathematical Studies Project he wants to investigate whether or not there is a correlation between the length of the arm of a swimmer and the time it takes them to swim 200 m. He selects 15 members of a swimming club to swim 200 m. Their times (y seconds) and arm lengths (x cm) are shown in the table below.

Length of arm (x cm)	78	72	74	67	79	58	62	67	71	69	75	65	73	59	60
Time (y seconds)	130	135	132	143	133	148	140	139	135	145	129	140	130	145	142

- Calculate the mean and standard deviation of x and y .
 - Calculate the correlation coefficient, r .
 - Comment on your value for r .
 - Calculate the equation of the regression line for y on x .
 - Using your equation, estimate how many seconds it will take a swimmer with an arm length of 70 cm to swim 200 m.
- 4 Saif asked his classmates how many minutes it took them to travel to school and their stress level, out of 10, for this journey. The results are shown in the table.

Travel time (x minutes)	14	28	19	22	24	8	16	5	18	20	25	10
Stress level (y)	3	7	5	6	6	2	3	2	4	5	6	6

- Find the correlation coefficient, r .
- Find the equation of the regression line.
- Use your equation to estimate the stress level of a student who takes 15 minutes to travel to school.



EXAM-STYLE QUESTION

- 5 The table shows the weight (g) and the cost (Australian dollars) of various candy bars.

Weight (xg)	62	84	79	65	96	58	99	48	73	66
Cost (yAUD)	1.45	1.83	1.78	1.65	1.87	1.42	1.82	1.15	1.64	1.55

- a Calculate the equation of the regression line for y on x .
 b Use your equation to estimate the cost of a candy bar weighing 70 g.



- 6 Ten students in Mr Craven's PE class did pushups and situps. Their results are shown in the following table.

Number of pushups (x)	23	19	31	53	34	46	45	22	39	27
Number of situps (y)	31	26	35	51	36	48	45	28	41	30

- a Find the equation of the regression line.

A student can do 50 pushups.

- b Use your equation to estimate the number of situps the student can do.

- 7 Fifteen students were asked for their average grade at the end of their last year of high school and their average grade at the end of their first year at university. The results are shown in the table below.

High school grade (x)	44	49	53	47	52	58	67	73	75	79	82	86	88	91	97
University grade (y)	33	52	55	48	51	60	71	72	69	83	84	89	96	92	89

- a Find the equation of the regression line.

A student scores 60 at the end of their last year of high school.

- b Use your equation to estimate the average university grade for the student.

- 8 A secretarial agency has a new computer software package. The agency records the number of hours it takes people of different ages to master the package. The results are shown in the table.

Age (x)	32	40	21	45	24	19	17	21	27	54	33	37	23	45
Time (y hours)	10	12	8	15	7	8	6	9	11	16	12	13	9	17

- a Find the equation of the regression line.

- b Using your equation, estimate the time it would take a 40-year-old person to master the package.

5.4 The chi-squared test

You may be interested in finding out whether or not certain sets of data are independent. Suppose you collect data on the favorite color of T-shirt for men and women. You may want to find out whether color and gender are independent or not. One way to do this is to perform a **chi-squared test** (χ^2) for independence.

To perform a chi-squared test (χ^2) there are four main steps.

Step 1: Write the **null (H_0)** and **alternative (H_1) hypotheses**.

H_0 states that the data sets are independent.

H_1 states that the data sets are not independent.

For example, the hypotheses for color of T-shirt and gender could be:

H_0 : Color of T-shirt is independent of gender.

H_1 : Color of T-shirt is not independent of gender.

Step 2: Calculate the chi-squared test statistic.

Firstly, you may need to put the data into a **contingency table**, which shows the frequencies of two variables. The elements in the table are the **observed** data. The elements should be frequencies (not percentages).

For the example above, the contingency table could be:

	Black	White	Red	Blue	Totals
Male	48	12	33	57	150
Female	35	46	42	27	150
Totals	83	58	75	84	300

If you are given the contingency table, you may need to extend it to include an extra row and column for the 'Totals'.

From the observed data, you can calculate the **expected frequencies**. Since you are testing for independence, you can use the formula for the probability of independent events to calculate the expected values. So:

The expected number of men who like black T-shirts is $\frac{150}{300} \times \frac{83}{300} \times 300 = 41.5$.

The expected number of men who like white T-shirts is $\frac{150}{300} \times \frac{58}{300} \times 300 = 29$ and so on.

The expected table of values would then look like this:

	Black	White	Red	Blue	Totals
Male	41.5	29	37.5	42	150
Female	41.5	29	37.5	42	150
Totals	83	58	75	84	300

When two variables are independent, one does not affect the other. Here, you are finding out whether a person's gender influences their colour choice. You will learn more about mathematical independence in Chapter 8.

The main entries in this table form a 2×4 **matrix** (array of numbers) - do not include the row and column for the totals.

In examinations, the largest contingency table will be a 4×4 .

Note:

- The expected values can **never** be less than 1.
- The expected values must be 5 or higher.
- If there are entries between 1 and 5, you can combine table rows or columns.

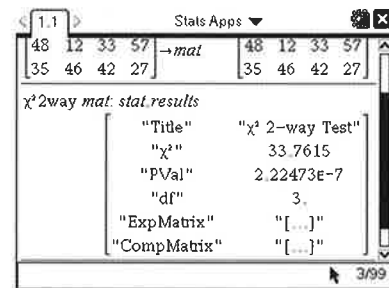
For calculations by hand, you need the expected frequencies to find the χ^2 value.

→ To calculate the χ^2 value use the formula $\chi^2_{\text{calc}} = \sum \frac{(f_o - f_e)^2}{f_e}$, where f_o are the observed frequencies and f_e are the expected frequencies.

For our example,

$$\begin{aligned} \chi^2_{\text{calc}} &= \frac{(48-41.5)^2}{41.5} + \frac{(12-29)^2}{29} + \frac{(33-37.5)^2}{37.5} + \frac{(57-42)^2}{42} + \frac{(35-41.5)^2}{41.5} \\ &\quad + \frac{(46-29)^2}{29} + \frac{(42-37.5)^2}{37.5} + \frac{(27-42)^2}{42} \\ &= 33.8 \end{aligned}$$

Using your GDC to find the χ^2 value, enter the contingency table as a matrix (array) and then use the matrix with the χ^2 2-way test.



From the screenshot, you can see that $\chi^2_{\text{calc}} = 33.8$ (to 3 sf). This confirms our earlier hand calculation.

Step 3: Calculate the critical value.

First note the **level of significance**. This is given in examination questions but you have to decide which level to use in your project. The most common levels are 1%, 5% and 10%.

Now you need to calculate the number of **degrees of freedom**.

→ To find the degrees of freedom for the chi-squared test for independence, use this formula based on the contingency table:
Degrees of freedom = (number of rows - 1) × (number of columns - 1)

If the number of degrees of freedom is 1, you will be expected to use **Yates' continuity correction** to work out the chi-squared value. (In examinations the degrees of freedom will always be greater than 1.)

So, in our ongoing example, the number of degrees of freedom is $(2 - 1) \times (4 - 1) = 3$

In examinations, you will only be expected to use your GDC to find the χ^2 value.

Your GDC calculates the expected values for you but you must know how to find them by hand in case you are asked to show one or two calculations in an exam question. To see the matrix for the expected values, type 'stat.' and then select 'expmatrix' from the menu that pops up.

GDC help on CD: Alternative demonstrations for the TI-84 Plus and Casio FX-9850GII GDCs are on the CD.

The level of significance and degrees of freedom can be used to find the critical value. However, in examinations, the **critical value** will always be given.

For our example, at the 1% level, the critical value is 11.345. At the 5% level, the critical value is 7.815. At the 10% level, the critical value is 6.251.

Step 4: Compare χ^2_{calc} against the critical value.

→ If χ^2_{calc} is **less than** the critical value then **do not reject** the null hypothesis.
If χ^2_{calc} is **more than** the critical value then **reject** the null hypothesis.

In our example, at the 5% level, $33.8 > 7.815$. Therefore, we reject the null hypothesis that T-shirt color is independent of gender.

Using a GDC, you can compare the p -value against the significance level.

→ If the p -value is **less** than the significance level then **reject** the null hypothesis.
If the p -value is **more** than the significance level then **do not reject** the null hypothesis.

Use the significance level as a decimal, so 1% = 0.01, 5% = 0.05 and 10% = 0.1.

So, for our example, p -value = 0.0000002 (see the GDC screenshot on page 234).

$0.0000002 < 0.05$, so we reject the null hypothesis.

→ **To perform a χ^2 test:**

- 1 Write the null (H_0) and alternative (H_1) hypotheses.
- 2 Calculate χ^2_{calc} :
 - a using your GDC (examinations)
 - b using the χ^2_{calc} formula (project work)
- 3 Determine:
 - a the p -value by using your GDC
 - b the critical value (given in examinations)
- 4 Compare:
 - a the p -value against the significance level
 - b χ^2_{calc} against the critical value

The p -value is the probability value. It is the probability of evidence against the null hypothesis.

Investigation – shoe size and gender

Use the information that you collected at the beginning of this chapter to test if shoe size is independent of gender.



Example 13

One hundred people were interviewed outside a chocolate shop to find out which flavor of chocolate cream they preferred. The results are given in the table, classified by gender.

	Strawberry	Coffee	Orange	Vanilla	Totals
Male	23	18	8	8	57
Female	15	6	12	10	43
Totals	38	24	20	18	100

Perform a χ^2 test, at the 5% significance level, to determine whether the flavor of chocolate cream is independent of gender.

- State the null hypothesis and the alternative hypothesis.
- Show that the expected frequency for female and strawberry flavor is approximately 16.3.
- Write down the number of degrees of freedom.
- Write down the χ^2_{calc} value for this data.

The critical value is 7.815.

- Using the critical value or the p -value, comment on your result.

Answers

- H_0 : Flavor of chocolate cream is independent of gender.
 H_1 : Flavor of chocolate cream is not independent of gender.

$$\text{b } \frac{43}{100} \times \frac{38}{100} \times 100 = 16.34$$

So, the expected frequency for female and strawberry flavor is approximately 16.3.

$$\text{c } \text{Degrees of freedom} = (2 - 1)(4 - 1) = 3$$

$$\text{d } \chi^2_{\text{calc}} = 6.88$$

- $6.88 < 7.815$; therefore, we do not reject the null hypothesis. There is enough evidence to conclude that flavor of chocolate cream is independent of gender.

Write H_0 using 'independent of'.
Write H_1 using 'not independent of'.

From the contingency table:
Total for 'female' row = 43
Total for 'strawberry' column = 38
Total surveyed = 100

Degrees of freedom = (number of rows - 1) (number of columns - 1)

Here, there are 2 rows and 4 columns in the observed matrix of the contingency table.

Using your GDC:
Enter the contingency table as a matrix. Use the matrix with χ^2 2-way test. Read off χ^2 value.
The p -value = 0.0758.

Using the given critical value, check:
 $\chi^2_{\text{calc}} < \text{critical value} \rightarrow \text{do not reject}$, or
 $\chi^2_{\text{calc}} > \text{critical value} \rightarrow \text{reject}$.
Or, using the p -value, check:
 $p\text{-value} < \text{significance level} \rightarrow \text{reject}$, or
 $p\text{-value} > \text{significance level} \rightarrow \text{do not reject}$.
Significance level = 5% = 0.05. So, $0.0758 > 0.05$ and we do not reject the null hypothesis.



Example 14

Members of a club are required to register for one of three games: billiards, snooker or darts.

The number of club members of each gender choosing each game in a particular year is shown in the table.

	Billiards	Snooker	Darts
Male	39	16	8
Female	21	14	17

Perform a χ^2 test, at the 10% significance level, to determine if the chosen game is independent of gender.

- State the null hypothesis and the alternative hypothesis.
- Show that the expected frequency for female and billiards is approximately 27.1.
- Write down the number of degrees of freedom.
- Write down the χ^2_{calc} value for this data.

The critical value is 4.605.

- Using the critical value or the p -value, comment on your result.

Answers

- H_0 : The choice of game is independent of gender.
 H_1 : The choice of game is not independent of gender.

$$\text{b } \left(\frac{52}{115} \right) \left(\frac{60}{115} \right) (115) = 27.130 \approx 27.1$$

So, the expected frequency for female and billiards is approximately 27.1.

$$\text{c } \text{Degrees of freedom} = (2 - 1)(3 - 1) = 2$$

$$\text{d } \chi^2_{\text{calc}} = 7.79$$

- $7.79 > 4.605$; therefore, we reject the null hypothesis. There is enough evidence against H_0 to conclude that the choice of game is not independent of gender.

Expected value table from the GDC:

	Billiards	Snooker	Darts
Male	32.9	16.4	13.7
Female	27.1	13.6	11.3

The p -value = 0.0203
Or, using the p -value,
 $0.0203 < 0.10$. Therefore, we reject the null hypothesis.



Exercise 5H

EXAM-STYLE QUESTIONS

- 1 300 people were interviewed and asked which genre of books they mostly read. The results are given below in a table of observed frequencies, classified by age.

		Genre			Totals
		Fiction	Non-fiction	Science fiction	
Age	0–25 years	23	16	41	80
	26–50 years	54	38	38	130
	51+ years	29	43	18	90
Totals		106	97	97	300

Perform a χ^2 test, at the 5% significance level, to determine whether genre of book is independent of age.

- State the null hypothesis and the alternative hypothesis.
- Show that the expected frequency for science fiction and the 26–50 age group is 42.
- Write down the number of degrees of freedom.
- Write down the χ^2_{calc} value for this data.

The critical value is 9.488.

- Using the critical value or the p -value, comment on your result.
- 2 Tyne was interested in finding out whether natural hair color was related to eye color. He surveyed all the students at his school. His observed data is given in the table below.

		Hair color			Totals
		Black	Brown	Blonde	
Eye color	Brown/Black	35	43	12	90
	Blue	8	27	48	83
	Green	9	20	25	54
	Totals	52	90	85	227

Perform a chi-squared test, at the 10% significance level, to determine if hair color and eye color are independent.

- State the null hypothesis and the alternative hypothesis.
- Find the expected frequency of a person having blonde hair and brown eyes.
- Write down the number of degrees of freedom.
- Write down the chi-squared value for this data.

The critical value is 7.779.

- Using the critical value or the p -value, comment on your result.



EXAM-STYLE QUESTIONS

- 3 Three different flavors of dog food were tested on different breeds of dog to find out if there was any connection between favorite flavor and breed. The results are given in the table.

	Beef	Chicken	Fish	Totals
Poodle	13	11	8	32
Boxer	15	10	10	35
Terrier	16	12	9	37
Great Dane	17	11	8	36
Totals	61	44	35	140



A χ^2 test, at the 5% significance level, is performed to investigate the results.

- State the null hypothesis and the alternative hypothesis.
- Show that the expected frequency of a Boxer's favorite food being chicken is 11.
- Show that the number of degrees of freedom is 6.
- Write down the χ^2_{calc} value for this data.

The critical value is 12.59.

- Using the critical value or the p -value, comment on your result.



- 4 Eighty people were asked to identify their favorite film genre. The results are given in the table below, classified by gender.

	Adventure	Crime	Romantic	Sci-fi	Totals
Male	15	12	2	12	41
Female	7	9	18	5	39
Totals	22	21	20	17	80

A χ^2 test, at the 1% significance level, is performed to decide whether film genre is independent of gender.

- State the null hypothesis and the alternative hypothesis.
- Show that the expected frequency of a female's favorite film genre being crime is 10.2.
- Write down the number of degrees of freedom.
- Write down the chi-squared value for this data.

The critical value is 11.345.

- Using the critical value or the p -value, comment on your result.



EXAM-STYLE QUESTIONS

- 5 Kyu Jin was interested in finding out whether or not the number of hours spent playing computer games per week had an influence on school grades. He collected the following information.

	Low grades	Average grades	High grades	Totals
0–9 hours	6	33	57	96
10–19 hours	11	35	22	68
> 20 hours	23	22	11	56
Totals	40	90	90	220

Perform a chi-squared test, at the 5% significance level, to decide whether the grade is independent of the number of hours spent playing computer games.

- State the null hypothesis and the alternative hypothesis.
- Show that the expected frequency of a high grade and 0–9 hours of playing computer games is 39.3.
- Show that the number of degrees of freedom is 4.
- Write down the χ^2_{calc} value for this data.

The critical value is 9.488.

- Using the critical value or the p -value, comment on your result.
- 6 The local authority conducted a survey in schools in Rotterdam to determine whether the employment grade in the school was independent of gender. The results of the survey are given in the table.

	Directors	Management	Teachers	Totals
Male	26	148	448	622
Female	6	51	1051	1108
Totals	32	199	1499	1730

Perform a χ^2 test, at the 10% significance level, to determine whether the employment grade is independent of gender.

- State the null hypothesis and the alternative hypothesis.
- Write down the table of expected frequencies.
- Write down the number of degrees of freedom.
- Write down the chi-squared value for this data.

The critical value is 4.605.

- Using the critical value or the p -value, comment on your result.



EXAM-STYLE QUESTIONS

- 7 Ayako had a part-time job working at a sushi restaurant. She calculated the average amount of sushi sold per week to be 2000. She decided to find out if there was a relationship between the day of the week and the amount of sushi sold. Her observations are given in the table.

	< 1700	1700–2300	> 2300	Totals
Monday–Wednesday	38	55	52	145
Thursday–Friday	39	65	55	159
Saturday–Sunday	43	60	63	166
Totals	120	180	170	470

Perform a χ^2 test, at the 5% significance level, to determine whether the amount of sushi sold is independent of the day of the week.

- State the null hypothesis and the alternative hypothesis.
- Show that the expected frequency of selling over 2300 sushi on Monday–Wednesday is 52.4.
- Write down the number of degrees of freedom.
- Write down the χ^2_{calc} value for this data.

The critical value is 9.488.

- Using the critical value or the p -value, comment on your result.

- 8 Haruna wanted to investigate the connection between the weight of dogs and the weight of their puppies. Her observed results are given in the table.

		Puppy			Totals
		Heavy	Medium	Light	
Dog	Heavy	23	16	11	50
	Medium	10	20	16	46
	Light	8	15	22	45
Totals		41	51	49	141

Perform a χ^2 test, at the 1% significance level, to determine whether a puppy's weight is independent of its parent's weight.

- State the null hypothesis and the alternative hypothesis.
- Show that the expected frequency of a medium dog having a heavy puppy is 13.4.
- Write down the number of degrees of freedom.
- Write down the χ^2_{calc} value for this data.

The critical value is 13.277.

- Using the critical value or the p -value, comment on your result.



Extension material on CD:
Worksheet 5 - Useful
statistical techniques for
the project

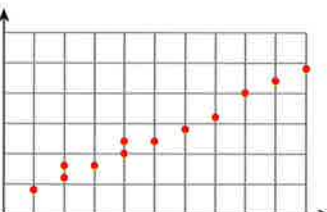
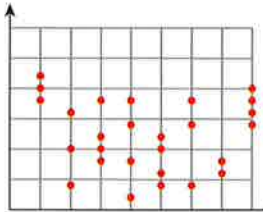
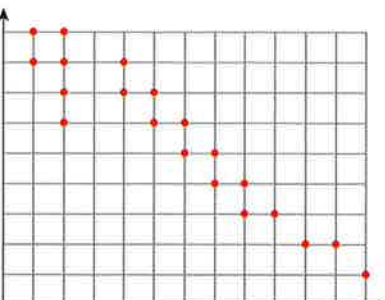


Review exercise

Paper 1 style questions



EXAM-STYLE QUESTIONS

- It is stated that the content of a can of drink is 350 ml. The content of thousands of cans is tested and found to be normally distributed with a mean of 354 ml and a standard deviation of 2.5 ml.
 - Sketch a normal distribution diagram to illustrate this information.
 - Find the probability that a can contains less than 350 ml. 100 cans are chosen at random.
 - Find the expected number of cans that contain less than 350 ml.
- 6000 people were asked how far they lived from their work. The distances were normally distributed with a mean of 4.5 km and a standard deviation of 1.5 km.
 - Find the percentage of people who live between 2 km and 4 km from their work.
 - Find the expected number of people who live less than 1 km from their work.
- The weights of bags of tomatoes are normally distributed with a mean of 1.03 kg and a standard deviation of 0.02 kg.
 - Find the percentage of bags that weigh more than 1 kg. It is known that 15% of the bags weigh less than p kg.
 - Find the value of p .
- For each diagram, state the type of correlation.
 - 
 - 
 - 

- Plot these points on a diagram.

x	6	8	10	12	14	16
y	20	21	24	27	28	30

- State the nature of the correlation.
 - Find the mean of the x -values and the mean of the y -values. Plot this mean point on your diagram.
 - Draw the line of best fit by eye.
 - Find the expected value for y when $x = 9$.
- The heights and arm lengths of 10 people are shown in the table.

Height (cm)	145	152	155	158	160	166	172	179	183	185
Arm length (cm)	38	42	45	53	50	59	61	64	70	69

- Find the correlation coefficient, r , and comment on your result.
 - Write down the equation of the regression line.
 - Use your equation to estimate the arm length of a person of height 170 cm.
- The time taken to eat three doughnuts and the person's age is recorded in the table.

Age (years)	8	12	15	18	21	30	33	35	44	52	63	78
Time (seconds)	23	21	17	14	15	18	20	21	23	25	27	35

- Find the correlation coefficient, r , and comment on your result.
 - Write down the equation of the regression line.
 - Use your equation to estimate the time taken by a 40-year-old to eat three doughnuts.
- 100 people are asked to identify their favorite flavor of ice cream. The results are given in the contingency table, classified by age (x).

	$x < 25$	$25 \leq x < 45$	$x \geq 45$	Totals
Vanilla	14	13	10	37
Strawberry	11	9	8	28
Chocolate	13	10	12	35
Totals	38	32	30	100

Perform a chi-squared test, at the 5% significance level, to determine whether flavor of ice cream is independent of age. State clearly the null and alternative hypotheses, the expected values and the number of degrees of freedom.

- 9 60 students go ten-pin bowling. They each have one throw with their right hand and one throw with their left. The number of pins knocked down each time is noted. The results are collated in the table.

	0-3	4-7	8-10	Totals
Right hand	8	28	24	60
Left hand	12	30	18	60
Totals	20	58	42	120

A χ^2 test is performed at the 10% significance level.

- State the null hypothesis.
- Write down the number of degrees of freedom.
- Show that the expected number of students who knock down 0-3 pins with their right hand is 10.

The p -value is 0.422.

- Write down the conclusion reached at the 10% significance level.

Give a clear reason for your answer.

- 10 Erland performs a chi-squared test to see if there is any association between the preparation time for a test (short time, medium time, long time) and the outcome (pass, does not pass). Erland performs this test at the 5% significance level.

- Write down the null hypothesis.
- Write down the number of degrees of freedom.

The p -value for this test is 0.069.

- What conclusion can Erland make? Justify your answer.

Paper 2 style questions

EXAM-STYLE QUESTIONS

- 1 The heights of Dutch men are normally distributed with a mean of 181 cm and a standard deviation of 9 cm.
- Sketch a normal distribution diagram to illustrate this information.
 - Find the probability that a man chosen at random has a height less than 175 cm.
 - Find the probability that a man chosen at random has a height between 172 cm and 192 cm.

Sixty men are measured.

- Find the expected number of men with a height greater than 195 cm.

It is known that 5% of the men have a height less than k cm.

- Find the value of k .

- 2 The weights of bags of sweets are normally distributed with a mean of 253 g and a standard deviation of 3 g.

- Sketch a diagram to illustrate this information clearly.
- Find the percentage of bags expected to weigh less than 250 g.

Three hundred bags are weighed.

- Find the expected number of bags weighing more than 255 g.

- 3 The heights and weights of 10 students selected at random are shown in the table.

Height (x cm)	158	167	178	160	152	160	173	181	185	155
Weight (y kg)	50	75	80	46	61	69	64	86	74	68

- Plot this information on a scatter graph. Use a scale of 1 cm to represent 25 cm on the x -axis and 1 cm to represent 10 kg on the y -axis.
- Calculate the mean height.
- Calculate the mean weight.
- Find the equation of the regression line.
 - Draw the regression line on your graph.
- Use your line to estimate the weight of a student of height 170 cm.



- 4 An employment agency has a new computer software package. The agency investigates the number of hours it takes people of different ages to reach a satisfactory level using this package. Fifteen people are tested and the results are given in the table.

Age (x)	33	41	22	46	25	18	16	23	26	55	37	34	25	48	17
Time (y hours)	8	10	7	16	8	9	7	10	12	15	11	14	10	16	7

- Find the product-moment correlation coefficient, r , for these data.
- What does the value of the correlation coefficient suggest about the relationship between the two variables?
- Write down the equation of the regression line for y on x in the form $y = mx + c$.
- Use your equation for the regression line to predict the time that it would take a 35-year-old person to reach a satisfactory level. Give your answer correct to the nearest hour.

EXAM-STYLE QUESTIONS

- 5 Ten students were asked for their average grade at the end of their last year of high school and their average grade at the end of their first year at university. The results were put into a table as follows.

Student	High school grade, x	University grade, y
1	92	3.8
2	76	2.9
3	83	3.4
4	71	1.8
5	93	3.9
6	84	3.2
7	96	3.5
8	77	2.9
9	91	3.7
10	86	3.8

- a Find the correlation coefficient, r , giving your answer to one decimal place.
 b Describe the correlation between the high school grades and the university grades.
 c Find the equation of the regression line for y on x in the form $y = mx + c$.

- 6 Several bars of chocolate were purchased and the following table shows the weight and the cost of each bar.

	Yum	Choc	Marl	Twil	Chuns	Lyte	BigM	Bit
Weight (x grams)	58	75	70	68	85	52	94	43
Cost (y euros)	1.18	1.45	1.32	1.05	1.70	0.90	1.53	0.95

- a Find the correlation coefficient, r , giving your answer correct to two decimal places.
 b Describe the correlation between the weight of a chocolate bar and its cost.
 c Calculate the equation of the regression line for y on x .
 d Use your equation to estimate the cost of a chocolate bar weighing 80 g.

- 7 The heights and dress sizes of 10 female students selected at random are shown in the table.

Height (x cm)	175	160	180	155	178	159	166	185	189	173
Dress size (y)	12	14	14	8	12	10	14	16	16	14

- a Write down the equation of the regression line for dress size (y) on height (x), giving your answer in the form $y = ax + b$.
 b Use your equation to estimate the dress size of a student of height 170 cm.
 c Write down the correlation coefficient.
 d Describe the correlation between height and dress size.

EXAM-STYLE QUESTIONS

- 8 Members of a certain club are required to register for one of three games: badminton, table tennis or darts. The number of club members of each gender choosing each game in a particular year is shown in the table.

	Badminton	Table tennis	Darts
Male	37	16	28
Female	32	10	19

Use a chi-squared test, at the 5% significance level, to test whether choice of game is independent of gender. State clearly the null and alternative hypotheses, the expected values and the number of degrees of freedom.



- 9 For his Mathematical Studies Project a student gave his classmates a questionnaire to find out which extra-curricular activity was the most popular. The results are given in the table below, classified by gender.

	Reading	Surfing	Skating	
Female	22	16	22	(60)
Male	14	18	8	(40)
	(36)	(34)	(30)	

The table below shows the expected values.

	Reading	Surfing	Skating
Female	p	20.4	18
Male	q	r	12

- a Calculate the values of p , q and r .

The chi-squared test, at the 10% level of significance, is used to determine whether the extra-curricular activity is independent of gender.

- b i State a suitable null hypothesis.
 ii Show that the number of degrees of freedom is 2.

The critical value is 4.605.

- c Write down the chi-squared statistic.
 d Do you accept the null hypothesis? Explain your answer.

EXAM-STYLE QUESTIONS

- 10 A company conducted a survey to determine whether position in upper management was independent of gender. The results of this survey are tabulated below.

	Managers	Junior executives	Senior executives	Totals
Male	135	90	75	300
Female	45	130	25	200
Totals	180	220	100	500

The table below shows the expected number of males and females at each level, if they were represented proportionally to the total number of males and females employed.

	Managers	Junior executives	Senior executives	Totals
Male	a	c	60	300
Female	b	d	40	200
Totals	180	220	100	500

- a i Show that the expected number of male managers (a) is 108.
 ii Hence, write down the values of b , c and d .
 b Write suitable null and alternative hypotheses for these data.
 c i Find the chi-squared value.
 ii Write down the number of degrees of freedom.
 iii Given that the critical value is 5.991, what conclusions can be drawn regarding gender and position in upper management?

- 11 In the small town of Schiedam, population 8000, an election was held. The results were as follows.

	Urban voters	Rural voters
Candidate A	1950	1730
Candidate B	1830	1360
Candidate C	500	630

In a–d below, use a chi-squared test, at the 1% significance level, to decide whether the choice of candidate depends on where the voter lives.

H_0 : The choice of candidate is independent of where the voter lives.

- a Write down the alternative hypothesis.
 b Show that the expected number of rural voters for candidate A is 1711.
 c i Calculate the chi-squared value.
 ii Write down the number of degrees of freedom.

The critical value is 9.21.

- d i State your conclusion.
 ii Explain why you reached your conclusion.

EXAM-STYLE QUESTION

- 12 This table of observed results gives the number of candidates taking a Mathematics examination classified by gender and grade obtained.

	6 or 7	4 or 5	1, 2 or 3	Totals
Males	34	50	6	90
Females	40	60	10	110
Totals	74	110	16	200

The question posed is whether gender and grade obtained are independent.

- a Show that the expected number of males achieving a grade of 4 or 5 is 49.5.

A chi-squared test is set up at the 5% significance level.

- b i State the null hypothesis.
 ii State the number of degrees of freedom.
 iii Write down the chi-squared value.

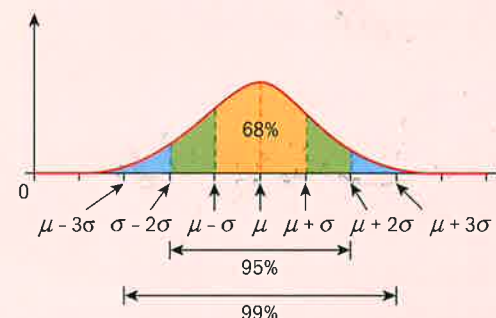
The critical value is 5.991.

- c What can you say about gender and grade obtained?

CHAPTER 5 SUMMARY

The normal distribution

- The **normal distribution** is the most important continuous distribution in statistics. It has these properties:
 - It is a bell-shaped curve.
 - It is symmetrical about the mean, μ . (The mean, the mode and the median all have the same value.)
 - The x -axis is an asymptote to the curve.
 - The total area under the curve is 1 (or 100%).
 - 50% of the area is to the left of the mean and 50% to the right.
 - Approximately 68% of the area is within 1 standard deviation, σ , of the mean.
 - Approximately 95% of the area is within 2 standard deviations of the mean.
 - Approximately 99% of the area is within 3 standard deviations of the mean.

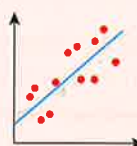
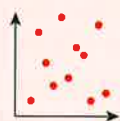
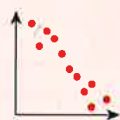
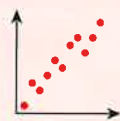


- The **expected value** is found by multiplying the number in the sample by the probability.

Continued on next page

Correlation

- In a **positive** correlation the dependent variable increases as the independent variable increases.
- In a **negative** correlation the dependent variable decreases as the independent variable increases.
- When the points are scattered randomly across the diagram there is **no** correlation.
- Correlations can also be described as strong, moderate or weak.
- To draw the **line of best fit** by eye:
 - Find the mean of each set of data and plot this point on your scatter diagram.
 - Draw a line that passes through the mean point and is close to all the other points – with approximately an equal number of points above and below the line.
- **Pearson's product-moment correlation coefficient**, r , can take all values between -1 and $+1$ inclusive.
 - When $r = -1$, there is a **perfect negative** correlation between the data sets.
 - When $r = 0$, there is **no** correlation.
 - When $r = +1$, there is a **perfect positive** correlation between the data sets.
 - A **perfect correlation** is one where **all** the plotted points lie on a straight line.



- If χ^2_{calc} is **less than** critical value, **do not reject** the null hypothesis.
- If χ^2_{calc} is **more than** critical value, **reject** the null hypothesis.
- If the p -value is **less than** significance level, **reject** the null hypothesis.
- If the p -value is **more than** significance level, **do not reject** the null hypothesis.
- To perform a χ^2 test:
 - 1 Write the null (H_0) and alternative (H_1) hypotheses.
 - 2 Calculate χ^2_{calc} : **a** using your GDC (examinations), or **b** using the χ^2_{calc} formula (project work).
 - 3 Determine: **a** the p -value using your GDC, or **b** the critical value (given in examinations).
 - 4 Compare: **a** the p -value against the significance level, or **b** χ^2_{calc} against the critical value.

The regression line

- The **regression line for y on x** is a more accurate version of a line of best fit, compared to best fit by eye.
- If there is a strong or moderate correlation, you can use the regression line for y on x to predict values of y for values of x within the range of the data.

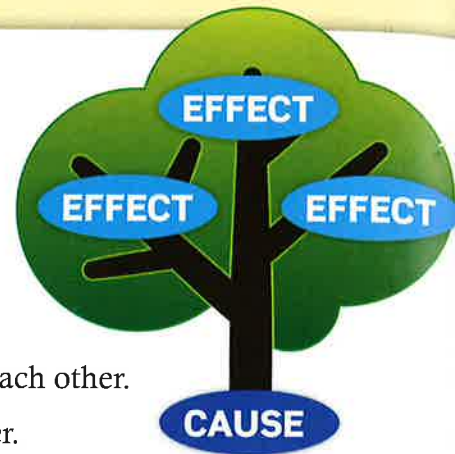
The chi-squared test

- To calculate the χ^2 value use the formula $\chi^2_{\text{calc}} = \sum \frac{(f_o - f_e)^2}{f_e}$, where f_o are the observed frequencies and f_e are the expected frequencies.
- To find the degrees of freedom for the chi-squared test for independence, use this formula based on the contingency table:
Degrees of freedom = (number of rows - 1)(number of columns - 1)

Continued on next page

Correlation or causation?

Correlation shows how closely two variables vary with each other.
Causation is when two variables directly affect each other.



Shaving less than once a day increases risk of stroke by 70%!

In 2003 British researchers found that there was a correlation between men's shaving habits and their risk of a stroke. This link emerged from a 20-year study of over 2,000 men aged 45–59 in Caerphilly, South Wales.



A strong **correlation** between two variables does not mean that one **causes** the other. There may be a cause and effect relation between the two variables, but you cannot claim this if they are only correlated. This is the **fallacy of correlation** – one of the most common logical fallacies.

Do you think a man could decrease his chance of having a stroke by shaving more? This seems silly, and suggests there might be a hidden intermediary variable at work.

In this case, the researchers think that shaving and stroke risk are linked by another variable – hormone levels. For example, testosterone has already been used to explain the link between baldness and a higher risk of heart disease.



If there is a correlation between two variables, be careful about assuming that there is a relationship between them. There may be no logical or scientific connection at all.

Analyse these examples of assumed correlation or causation. Which illustrate the fallacy of correlation?

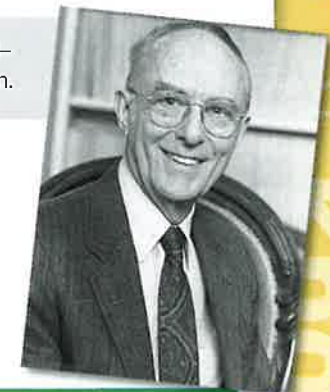


- Joining the military made me a disciplined and strong person
- I wore a hat today on my way to school and I was involved in a car accident; I will not be wearing that red hat again
- People who own washing machines are more likely to die in a car accident.

Anscombe's Quartet

Anscombe's Quartet is a group of four data sets that provide a useful caution against applying individual statistical methods to data without first graphing them. They have identical simple statistical properties (mean, variance, etc.) but look totally different when graphed.

► Francis Anscombe (1918–2001), British statistician.



- Find the mean of x , the mean of y , the variance of x and the variance of y and the r -value for each data set.

Set 1		Set 2		Set 3		Set 4	
x	y	x	y	x	y	x	y
4	4.26	4	3.1	4	5.39	8	6.58
5	5.68	5	4.74	5	5.73	8	5.76
6	7.24	6	6.13	6	6.08	8	7.71
7	4.82	7	7.26	7	6.42	8	8.84
8	6.95	8	8.14	8	6.77	8	8.47
9	8.81	9	8.77	9	7.11	8	7.04
10	8.04	10	9.14	10	7.46	8	5.25
11	8.33	11	9.26	11	7.81	8	5.56
12	10.84	12	9.13	12	8.15	8	7.91
13	7.58	13	8.74	13	12.74	8	6.89
14	9.96	14	8.1	14	8.84	19	12.5

- 1 Write down what you think the graphs and their regression lines will look like.
- 2 Using your GDC, sketch the graph of each set of points on a different graph.
- 3 Draw the regression line on each graph.
- 4 Explain what you notice.