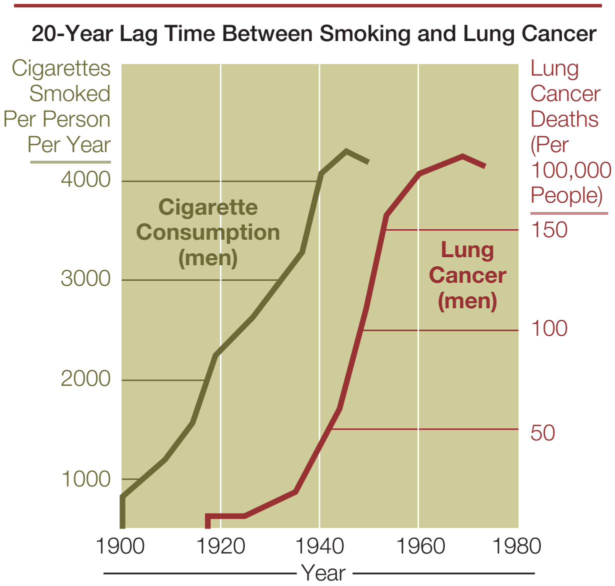
4 Statistical applications

In 1747 the government of Sweden conducted its first population census to find out how many citizens were living in the country. Today, governments all over the world use regular censuses to collect data about their citizens so that they have sufficient information to plan the services needed now and in the future.

Once information has been collected, it needs to be displayed in a meaningful way and then analysed if it is to be used constructively. Initially called 'State mathematics', the branch of mathematics that deals with the collection, organisation, presentation and analysis of data is now known as 'Statistics'. This chapter introduces three statistical techniques that have become increasingly influential.

For example, when the data on smoking habits of lung cancer patients was analysed, people came to realise that the connection or 'correlation' was too strong to have occurred by chance. Now the health ministries of most countries discourage their citizens from smoking because the risks are considered to be unacceptably high.

However, some authorities have argued that our reliance on statistical inference is now too great and is stopping us from taking events seriously that the statistics tell us are unlikely. In his book *Black Swan*, Nassim Nicholas Taleb suggests that the reliance of banks and trading firms on analysis based on the normal distribution means that they do not allow for very unlikely events (which he calls 'black swan events'), and this can lead them to ignore positive developments and to make very large mistakes when they encounter negative events.



Source: National Institutes of Health.

Prior learning topics

It will be easier to study this topic if you have already completed:

- Topic 2 (Chapters 5–7)
- Chapter 10

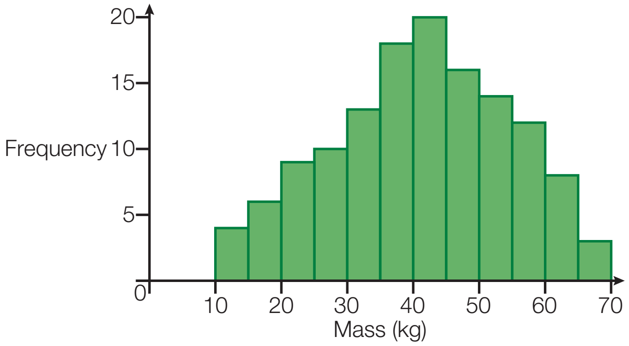
Chapter 11 The normal distribution

In this chapter you will learn:

- about the normal distribution and the concept of a random variable
- about the parameters σ and μ
- about properties of the normal distribution, such as the bell shape and the symmetry about $x = \mu$
- how to draw the normal distribution curve and areas under it on a diagram
- how to do normal probability calculations by using diagrams and your GDC
- how to do inverse normal calculations.

Every year Nando weighs his lambs before taking them to market. He can only take lambs above a certain mass, and any that are too light are left behind. He finds that most of his lambs are approximately the same mass, but some are heavier and some are lighter.

If Nando drew a histogram of the masses of his lambs, it would look like this:

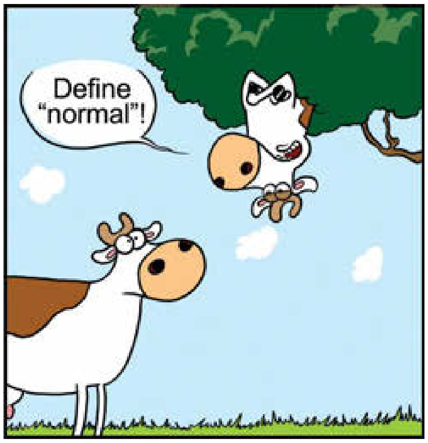


The shape of this histogram is characteristic of many sets of data that involve continuous variables such as height, length, time, blood pressure, or other quantities that are measured rather than counted.

RR ***Recall from Chapter 5 that discrete data can be counted, while continuous data comes from measuring.***

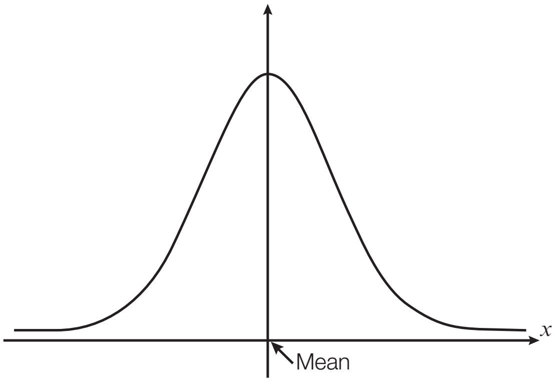
The masses of the lambs cluster around a central value; there are few lambs that are very heavy or very light, so the frequencies for masses near the middle of the range are higher than those for masses near the ends of the range. This creates a 'bell shaped' histogram, and you can see immediately that high or low values occur less frequently than those that are close to the central value.

This 'bell-shaped' distribution of frequencies is called the **normal distribution**, and it has been known and used for nearly three hundred years.



11.1 The normal distribution curve

The graph of the normal distribution looks like this:



The vertical 'axis' in the middle marks the position of the **mean**. x is the value of the continuous variable X being measured and is plotted along the horizontal axis.

You will have noticed straight away several properties of the normal distribution:

- It is symmetrical about the mean.
- The maximum height of the curve occurs at the mean.
- If you think of the height as representing 'frequency', then you can see that the mean, median and mode are all the same.

RR ***The mean, median and mode were defined in Chapter 6.***

The normal distribution has other special properties that make it useful in many different areas of statistics; it is particularly useful in the study of probabilities.

Two of these special properties are:

- The total area under the curve equals 1.
- The horizontal axis is an **asymptote** of the curve; the curve approaches the horizontal axis but never touches it.

To work with the normal distribution, you need to know the mean and standard deviation of the population from which data is being collected.

RR ***You met the standard deviation in Chapter 7. It is a measure of spread around the mean of a set of data.***

As the area under the normal distribution curve is equal to 1, you can use any partial area to calculate the probability of obtaining a particular range of data values.

RR ***All probabilities have values between 0 and 1; see Chapter 10.***

There is special notation for normal distributions. Suppose you are measuring a variable X (e.g. height, length or time). X is called a **random variable** because its value can be different each time it is observed; it is a **continuous random variable** if it can take any value within a certain interval of real numbers (which is the case when X is a measurement of some quantity). If the continuous random variable X follows a normal distribution, we write:

$$X \sim N(\mu, \sigma^2)$$

where:

- N denotes the normal distribution
- μ (the Greek letter 'mu') stands for the mean
- σ is the standard deviation.

You say this as 'the variable X is normally distributed with a mean of μ and a standard deviation of σ'.

Pierre Simon Laplace (1749–1827) discovered his mathematical talents when he entered Caen University, in France, at the age of 16. He studied the mathematical and scientific techniques current at that time, moving to Paris when he was 19.

As the scientific instruments in the eighteenth century were not very accurate, scientists worked by taking a series of measurements and then calculating their mean. Laplace proved that the mean of many observations is described by a 'bell-shaped' curve and that even when the original distribution is not normal, the mean of the repeated samples will be. He also proved that the larger the number of samples, the better the fit to the normal distribution.

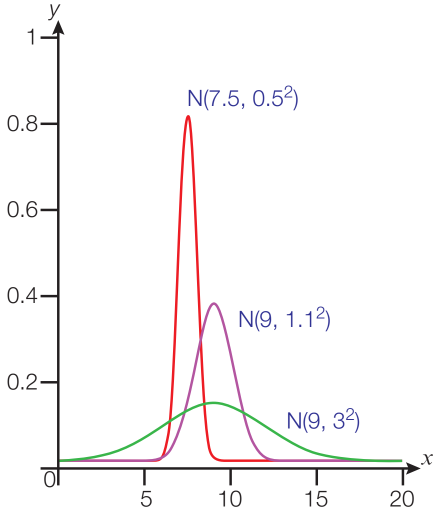
A normal distribution with a mean of 23 and a standard deviation of 3.7 is written as:

$$X \sim \mathrm{N}(23, 3.7^2)$$

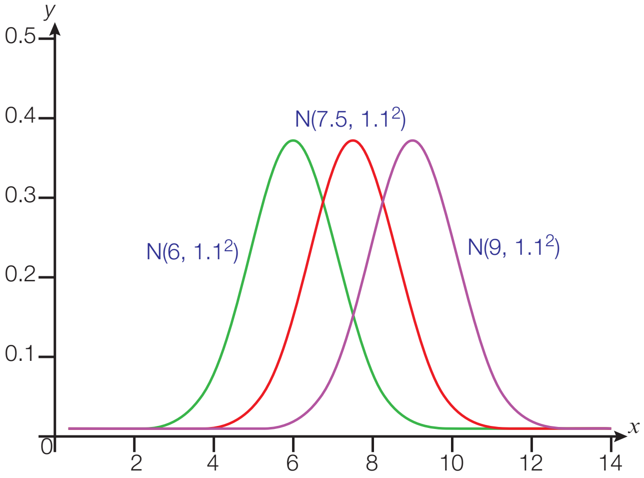
Relationship between the shape of the curve and the mean and standard deviation

The normal distribution curve is always bell-shaped, but the exact shape of the 'bell' is controlled by the values of the mean and the standard deviation.

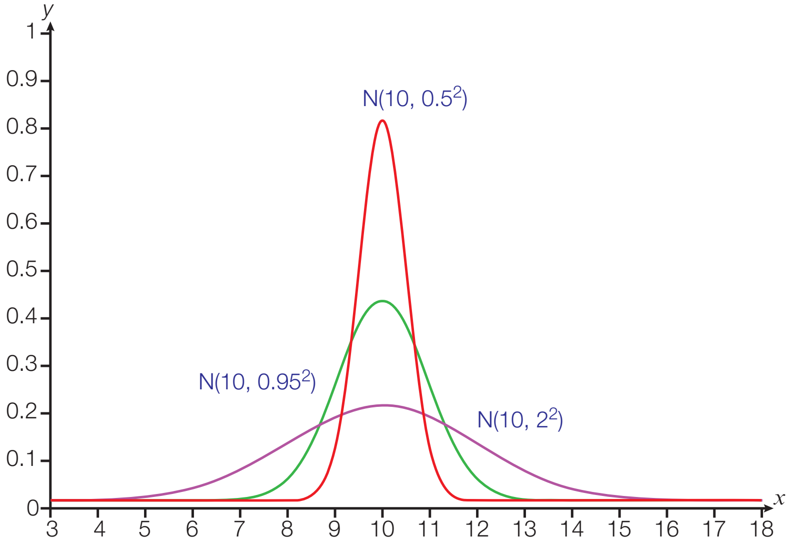
In the graph below, notice that a small standard deviation gives a curve that is tall and thin, while a large standard deviation gives a curve that is low and wide. In all cases, the total area under the curve equals 1.



The following graph shows three normal distribution curves with the same standard deviation but different mean values. The curves all have the same maximum height. Changing the value of the mean does not change the shape of the curve but moves it to a different place along the horizontal axis.



If you keep the mean at the same value but change the standard deviation, you will see that the curve remains centred at the same location on the horizontal axis, but its shape becomes more peaked or flatter depending on the value of the standard deviation: a small standard deviation gives a tall and narrow curve, while a large standard deviation gives a low, flat and wide curve.

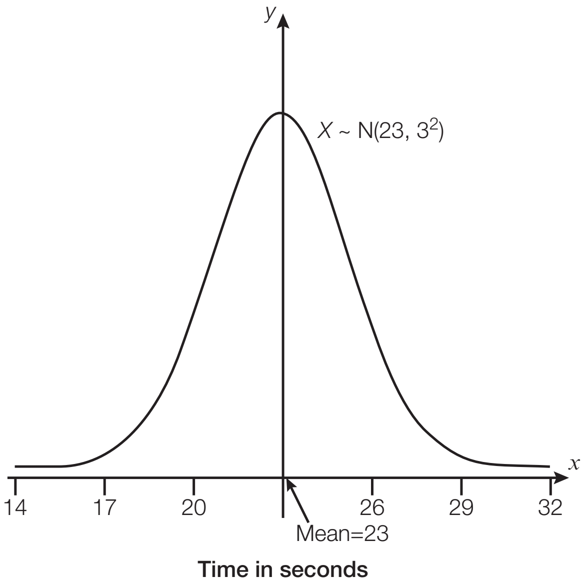


Symmetry of the normal distribution curve

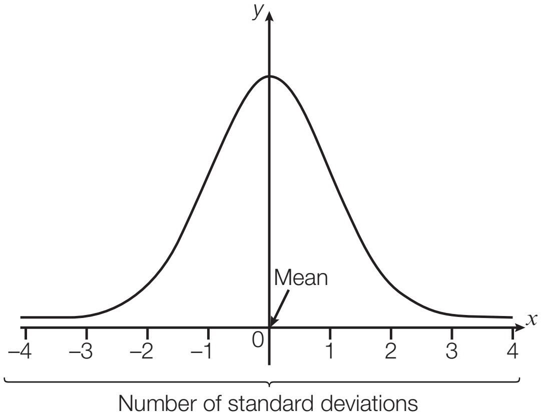
Understanding the symmetry of the normal distribution curve will be very important in learning how to use the curve for calculating probabilities.

When drawing the curve, there are two common ways of marking the *x*-axis.

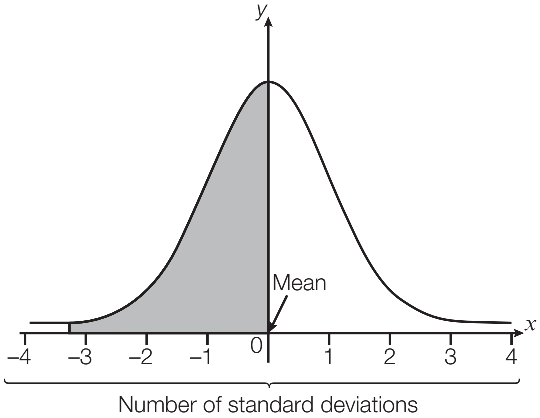
1. Mark the mean on the *x*-axis and put a sensible scale of *x* values around it. Write the distribution beside the curve.



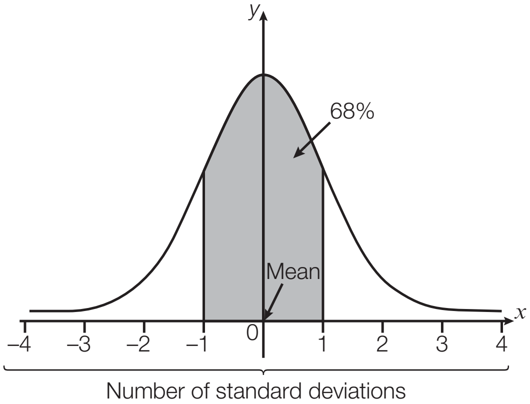
2. In some situations it is more relevant to mark the mean on the *x*-axis together with a certain number of standard deviations to the right and left of the mean. For instance, normal curves generated by spreadsheets or GDCs will usually show four standard deviations in each direction.

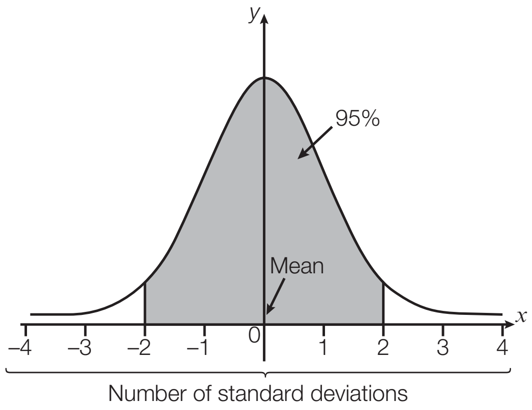
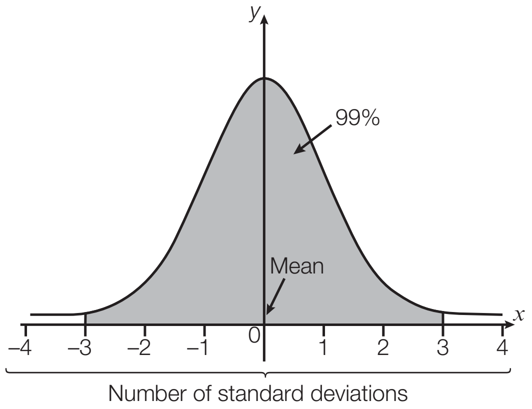


By the symmetry of the curve, 50% of the population lies to the right of the mean and 50% lies to the left; in other words, 50% of data values are expected to lie above the mean and 50% below the mean.



For applications, it is useful to remember what percentage of the population lies within a certain number of standard deviations of the mean:

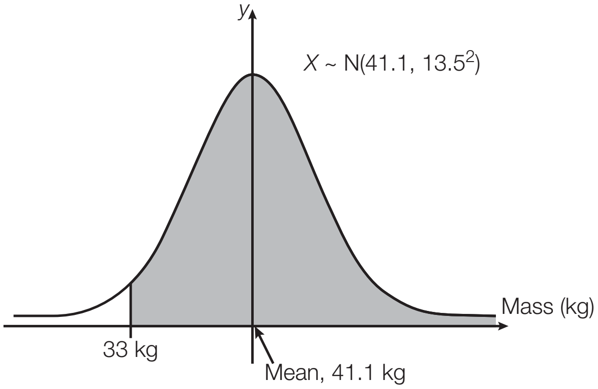
Approximately **68%** of the distribution lies between **–1** and **+1** standard deviations from the mean.	

Approximately **95%** of the distribution lies between **−2** and **+2** standard deviations from the mean.	
Approximately **99%** of the distribution lies between **−3** and **+3** standard deviations from the mean.	

11.2 Probability calculations using the normal distribution

The symmetry and general properties of the normal distribution make it suitable for use in many different fields. Modern statistics packages have made it easy to do calculations with the normal distribution, even if you do not understand the detailed mathematics behind it.

For example, the distribution of masses of Nando's lambs has a mean of 41.1 kg and a standard deviation of 13.5 kg. Nando can only send lambs to market if they have a mass of at least 33 kg. He can estimate the **proportion** of his lambs that can go to market by making a sketch like the following:



The proportion of lambs with a mass of 33 kg or more is represented by the area under the curve to the right of $x = 33$.

Sets of data that come from measurements can often be approximated very well by the normal distribution. These include measurements of physical, psychological, biological and economic phenomena, and so the normal distribution is used in disciplines ranging from physics and biology to finance and business.

When doing calculations with the normal distribution, it is always a good idea to draw clear diagrams showing the area that you are using to solve a problem.

As the total area under the curve is 1, you can use any partial area that you have shaded under the curve to estimate the **proportion** of your data that lies within that range of x values or, equivalently, the **probability** of your data taking on those values.

Your graphical calculator has a program that will perform the area calculations, but in order to enter all the data correctly you should sketch a diagram first. Before you do anything with your GDC, read the question carefully, and identify the area under the normal curve corresponding to the probability that you have been asked to calculate. Express the distribution in the form $X \sim \mathrm{N}(\mu, \sigma^2)$, and check ≥ and ≤ symbols in the question with particular care. Then draw a bell-shaped curve, and on your diagram:

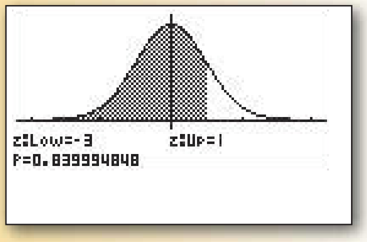
- mark the mean
- mark the lower and upper boundaries relevant to the question
- shade in the area that you need to find.

For example, if Nando wants to estimate the number of lambs whose mass is between 33 kg and 48 kg, he would shade in the following area under his normal curve: the lower boundary would be $x = 33$, and the upper boundary would be $x = 48$.

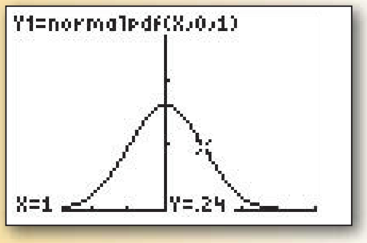
GDC

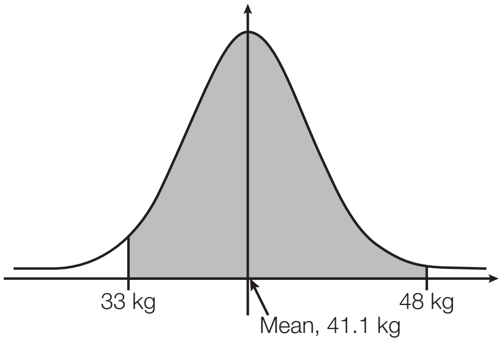
Be careful not to confuse normal **c**df with normal **p**df.

The cdf (cumulative distribution function) gives an area under the curve, i.e. the probability that you are calculating.


```
z:Low=-3          z:Up=1
P=0.839994848
```

The pdf (probability density function) gives the y-coordinate of a point on the normal curve corresponding to a given value of x.


```
Y1=normalpdf(X,0,1)
X=1          Y=.24
```

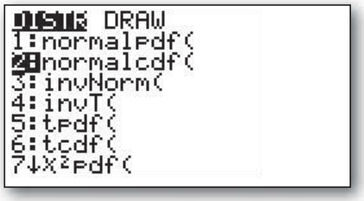


Mass (kg)

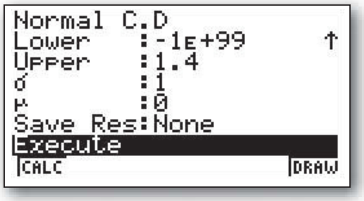
A hand-drawn diagram on which you have shaded the required area will help to ensure that you have understood the question. Now use a GDC.

Areas under the normal distribution curve are calculated by your GDC by drawing a graph (see section *11.1 (a)* on page 669 of the GDC chapter) or with the **normal cumulative distribution** (ncd) **function** (see section *11.1 (b)* on page 670 of the GDC chapter).

TEXAS


```
DISTR DRAW
1:normalpdf(
2:normalcdf(
3:invNorm(
4:invT(
5:tpdf(
6:tcdf(
7↓X²pdf(
```

CASIO


```
Normal C.D
Lower    :-1E+99       ↑
Upper    :1.4
σ        :1
μ        :0
Save Res:None
Execute
CALC                DRAW
```

Standard normal distribution N(0, 1^2)

A normal distribution with a mean of 0 and a standard deviation of 1 is called the **standard normal distribution**.

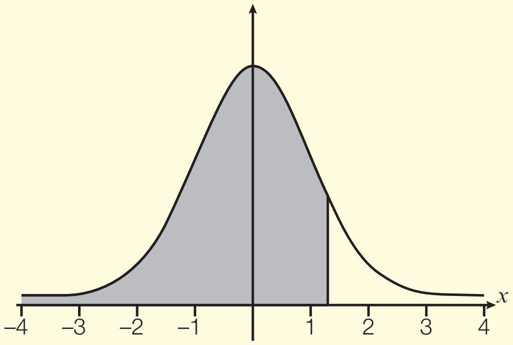
By default, the program on your GDC calculates areas for the standard normal distribution. This means that if the question you want to answer involves the standard normal distribution, you will only need to enter the upper and lower boundaries.

Worked example 11.1

Q. (a) If $X \sim \mathrm{N}(0, 1^2)$, find the probability that X is 1.15 or less; this is written as $\mathrm{P}(X \leq 1.15)$.

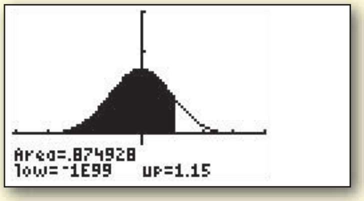
A.

Sketch a diagram first, either by hand or on a GDC or computer. Shade the area under the curve to the **left** of 1.15 because we are looking for $X \leq 1.15$.



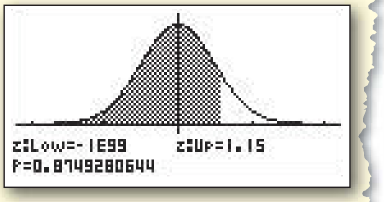
Calculate the probability using your GDC; you can use either method *11.1 (a)* on page 669 or method *11.1(b)* on page 670 of the GDC chapter. There is no lower boundary defined in the question, so enter a very large **negative** number such as –1E99 on your GDC. Enter 1.15 as the upper boundary.

TEXAS



```
Area=.874928
low=⁻1E99    up=1.15
```

CASIO



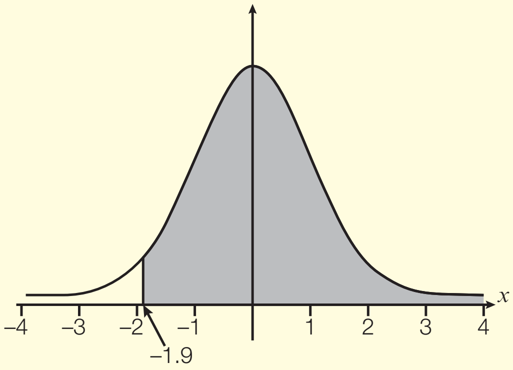
```
z:Low=-1E99      z:Up=1.15
P=0.8749280644
```

From GDC: $\mathrm{P}(X \leq 1.15) = 0.875$

Q. (b) If $X \sim \mathrm{N}(0, 1^2)$, find $\mathrm{P}(X \geq -1.9)$.

A.

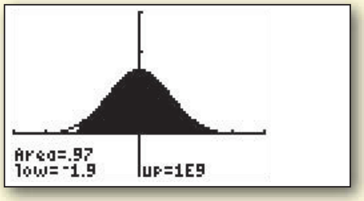
Sketch a diagram and shade the area under the curve to the **right** of –1.9.



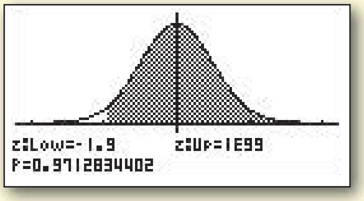
continued . . .

Use your GDC to calculate the probability by drawing a graph or using the ncd function. Enter −1.9 as the lower boundary. This time there is no upper boundary defined in the question, so use a very **large** positive number such as 1E9 or 1E99.

TEXAS


Area=.97
low=-1.9 up=1E9

CASIO

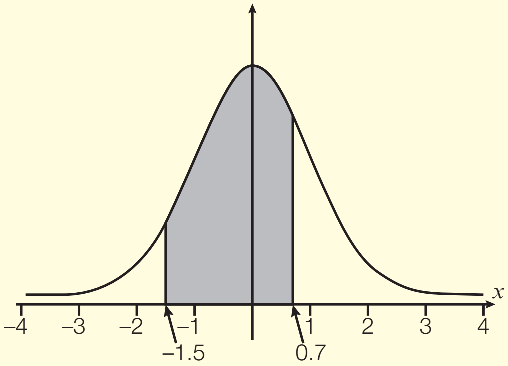

z:Low=-1.9 z:Up=1E99
P=0.9712834402

From GDC: $P(X \le 1.15) = 0.971$

Q. (c) If $X \sim N(0, 1^2)$, find the probability that X lies between −1.5 and 0.7.

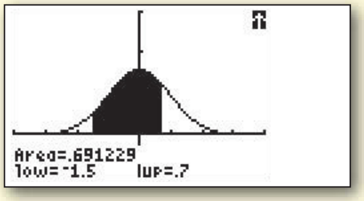
A.

Sketch a diagram. Shade the area under the curve between $x = -1.5$ and $x = 0.7$.

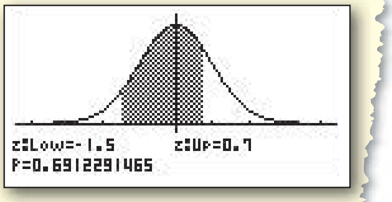


Use your GDC to calculate the probability; enter the given lower and upper boundaries into your GDC.

TEXAS


Area=.691229
low=-1.5 up=.7

CASIO


z:Low=-1.5 z:Up=0.7
P=0.6912291465

$P(-1.5 \le X \le 0.7) = 0.691$

Exercise 11.1

1. Assume that $X \sim N(0, 1^2)$. For each of the given probabilities, sketch the normal distribution curve and indicate, by shading and labelling, the required area.

(a) $P(X \ge 1.7)$

(b) $P(X \le 0.9)$

(c) $P(0 \le X \le 0.9)$

(d) $P(0.8 \le X \le 1.9)$

(e) $P(X \ge -0.8)$

(f) $P(X \le -1.9)$

(g) $P(-2.5 \le X \le -0.5)$

(h) $P(-0.5 \le X \le 2.5)$

(i) $P(0.5 \le X \le 2.5)$

(j) $P(-1.0 \le X \le 0)$

2. Calculate each of the probabilities in question 1.

3. (a) Sketch, shade and label the following areas, given that $X \sim \text{N}(0, 1^2)$.

(i) $\text{P}(X \geq 1.5)$ (ii) $\text{P}(X \leq -1.5)$

(b) Comment on the relationship between the two areas.

(c) Calculate both probabilities.

4. Given that $X \sim \text{N}(0, 1^2)$, what is the connection between the three probabilities $\text{P}(X \leq -2)$, $\text{P}(-2 \leq X \leq 2)$ and $\text{P}(X \geq 2)$?

General normal distribution N(μ, σ^2)

In most problems the normal distribution will not be the standard one with mean 0 and standard deviation 1.

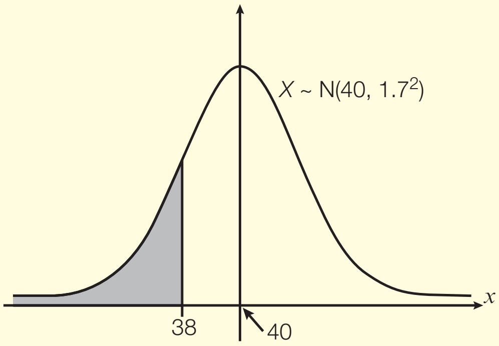
You can still calculate probabilities using your GDC, but now you will need to enter the mean and standard deviation as well as the lower and upper boundary values. On most GDCs, you enter the values of μ and σ **after** the lower and upper limits. It is still essential to draw a diagram to visualise the area you are calculating.

Worked example 11.2

Q. (a) If $X \sim \text{N}(40, 1.7^2)$, find the probability that $X \leq 38$.

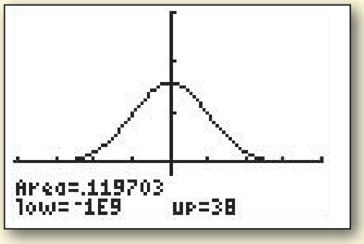
A.

Sketch a diagram.

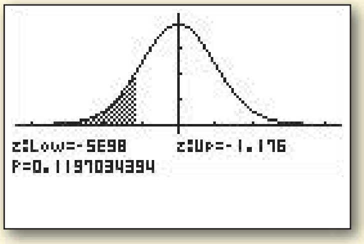


TEXAS

CASIO



```
Area=.119703
low=-1E9      up=38
```



```
z:Low=-5E98        z:Up=-1.176
P=0.1197034394
```

Use your GDC to find the probability. Since no lower limit is given, enter a large negative number. See '*11.1 Finding the area under a normal distribution curve*' on page 669 of the GDC chapter if you need a reminder.

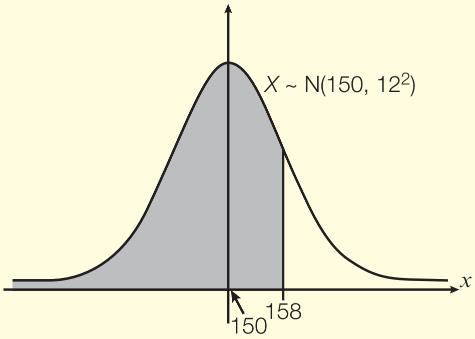
$\text{P}(X \leq 38) = 0.120$

continued . . .

Q. (b) If $X \sim \mathrm{N}(150, 12^2)$, find the probability that $X \leq 158$.

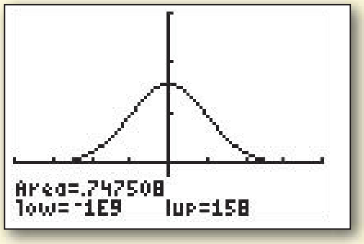
A.

From the diagram, you can predict that the answer should be greater than 0.5.

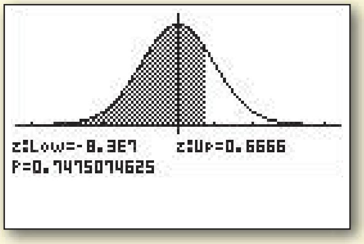


Use your GDC to calculate the probability.

TEXAS


Area=.747508
low=-1E9 up=158

CASIO

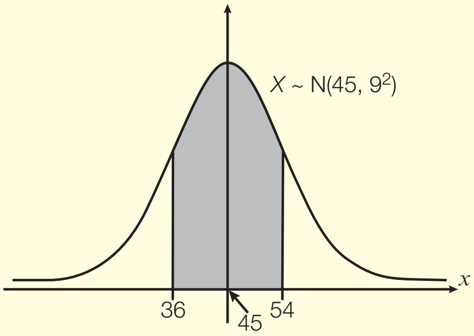

z:Low=-8.3E7 z:Up=0.6666
P=0.7475074625

$P(X \leq 158) = 0.748$

Q. (c) If $X \sim \mathrm{N}(45, 9^2)$, find $\mathrm{P}(36 \leq X \leq 54)$.

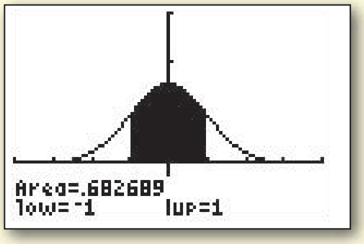
A.

Shade the area between $x = 36$ and $x = 54$.

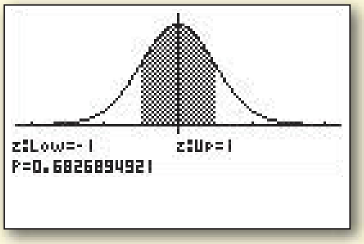


Use your GDC to calculate the probability.

TEXAS


Area=.682689
low=-1 up=1

CASIO


z:Low=-1 z:Up=1
P=0.6826894921

$P(36 \leq X \leq 54) = 0.683$

Exercise 11.2

1. For each of the following probabilities, sketch the normal distribution curve and indicate, by shading and labelling, the required areas.

(a) $X \sim \mathrm{N}(7, 2^2)$

(i) $\mathrm{P}(X \geq 9)$ (ii) $\mathrm{P}(X \leq 11.5)$

(iii) $\mathrm{P}(X \leq 4.8)$ (iv) $\mathrm{P}(5.4 \leq X \leq 8.2)$

(b) $X \sim \mathrm{N}(48, 9^2)$

(i) $\mathrm{P}(36 \leq X \leq 52)$ (ii) $\mathrm{P}(55 \leq X \leq 75)$

(iii) $\mathrm{P}(25 \leq X \leq 48)$ (iv) $\mathrm{P}(22 \leq X \leq 32)$

2. Calculate each of the probabilities in question 1.

3. Given that $X \sim \mathrm{N}\,(2.4, 3^2)$, calculate the following probabilities. Remember to shade the required area before using your GDC.

(a) $\mathrm{P}(X \leq -2)$ (b) $\mathrm{P}(-2 \leq X \leq 0)$ (c) $\mathrm{P}(-2 \leq X \leq 2.4)$

(d) $\mathrm{P}(-1 \leq X \leq 6.6)$ (e) $\mathrm{P}(X \geq 5.7)$

4. Let X be a normally distributed variable with a mean of 3 and a standard deviation of 2. What is $\mathrm{P}(X \geq 6)$?

11.3 Solving practical problems with the normal distribution

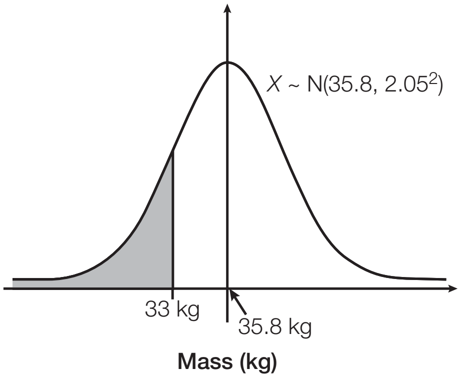
The calculation of probabilities often comes up in the context of solving practical problems. The bell-shaped distribution is a good model for a wide range of 'real world' situations, as Adolphe Quetelet found out in his quest for the 'average man'.

Suppose that one year Nando finds that his lambs have a mean mass of 35.8 kg, with a standard deviation of 2.05 kg. As usual, Nando can only send his lambs to market if they have a mass of at least 33 kg. What is the probability that a lamb has a mass of less than 33 kg? If Nando has 150 lambs, how many lambs do you expect he will have to keep back?

We assume that X, the random variable representing the mass of a lamb, follows a normal distribution. So:

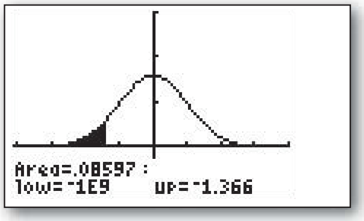
$$X \sim \mathrm{N}(35.8, 2.05^2)$$

The probability that a lamb has a mass of less than 33 kg is given by the area under the curve to the left of $x = 33$:



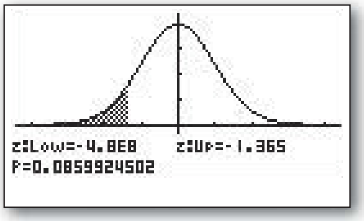
On your GDC, enter a large negative number (such as $-1E9$) for the lower boundary, 33 for the upper boundary, 35.8 for μ, and 2.05 for σ.

TEXAS



```
Area=.08597 :
low=⁻1E9      up=⁻1.366
```

CASIO



```
z:Low=-4.8E8    z:Up=-1.365
p=0.0859924502
```

Hence $P(X < 33) = 0.0860$.

As Nando has 150 lambs in total, the expected number that are below 33 kg is $150 \times 0.0860 = 13$. So Nando expects to keep 13 lambs because they are not heavy enough to go to market.

Worked example 11.3

Q. A manufacturer calculates that the lifetime of his laptop batteries is normally distributed with a mean life of 28 months and a standard deviation of 7.5 months.

The manufacturer gives a 12-month guarantee on each battery.

(a) What is the probability that a battery will last at least 28 months?

(b) What is the probability that a battery will last more than 38 months?

(c) What is the probability that a battery will last less than 12 months?

(d) The manufacturer makes 5000 batteries each month. How many batteries can he expect will be returned under the terms of his guarantee?

→

continued . . .

(e) The company accountant thinks that the figure in (d) is too high. Should the manufacturer aim to make the standard deviation larger or smaller?

Write down the distribution.

A. Let X be the lifetime of a battery; then

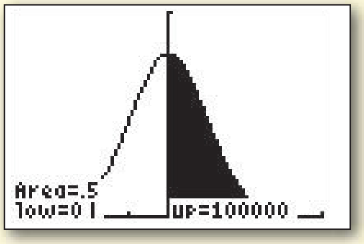
$X \sim \mathrm{N}(28, 7.5^2)$

To answer this question, you do not really need to use a calculator because you know the distribution is symmetrical about the mean.

(a) The mean of the distribution is 28, so by the symmetry of the normal distribution, $P(X \geq 28) = 0.5$

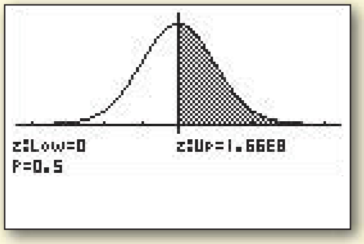
Check your answer on your GDC. See section *11.1* on pages 669-670 of the GDC chapter if you need a reminder.

TEXAS



```
Area=.5
low=0        up=100000
```

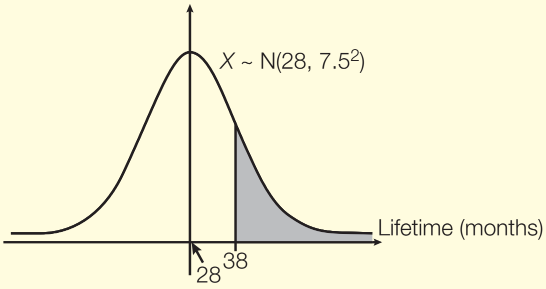
CASIO



```
z:Low=0         z:Up=1.66E8
P=0.5
```

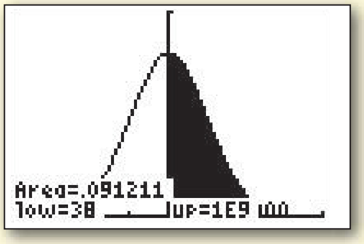
We want the area to the right of 38. Sketch the curve and this area.

(b)



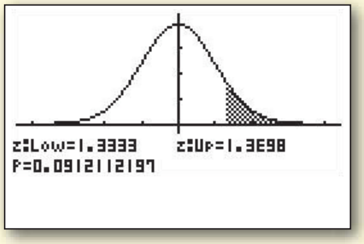
Use your GDC, as above, to calculate the probability.

TEXAS



```
Area=.091211
low=38       up=1E9
```

CASIO

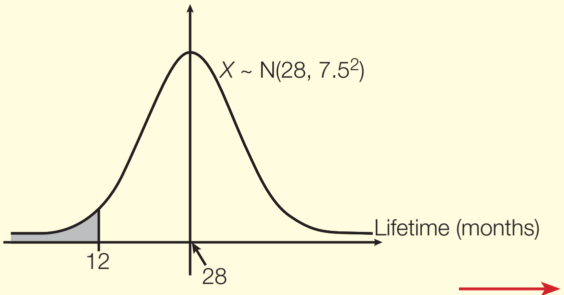


```
z:Low=1.3333    z:Up=1.3E98
P=0.0912112197
```

$P(X \geq 38) = 0.0912$

We want the area to the left of 12.

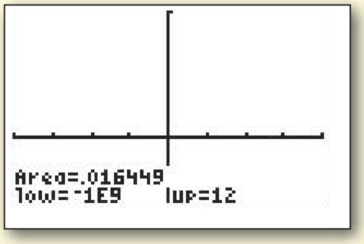
(c)



continued . . .

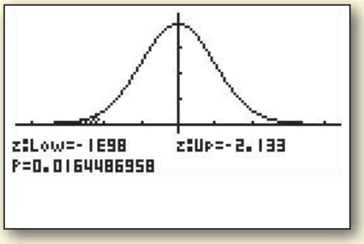
Notice that here the GDC has calculated the probability but not displayed the shaded curve. Be aware that sometimes your GDC might not give the expected output; but as long as you entered the calculation correctly, it should still produce the correct answer.

TEXAS



```
Area=.016449
low=-1E9     up=12
```

CASIO



```
z:Low=-1E98      z:Up=-2.133
P=0.0164486958
```

$P(X < 12) = 0.0164$

Remember from Chapter 10 that expected value = probability × frequency. The probability of a battery being returned is P(X < 12), which we found in (c).

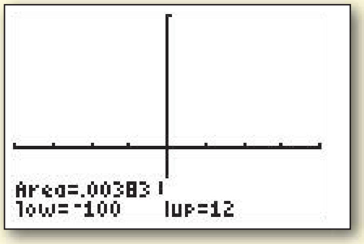
(d) Among 5000 batteries, the expected number that will be returned is $5000 \times 0.0164 = 82$.

(e) The manufacturer should aim to make the standard deviation smaller. For instance, if he reduces it to 6 months, only 19 batteries will be returned under the guarantee:

If $X \sim N(28, 6^2)$, then $P(X < 12) = 0.00383$

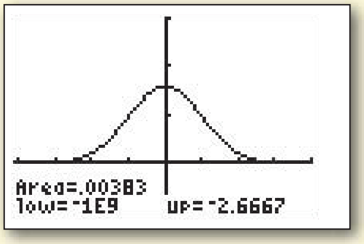
$5000 \times 0.00383 = 19.15$

TEXAS



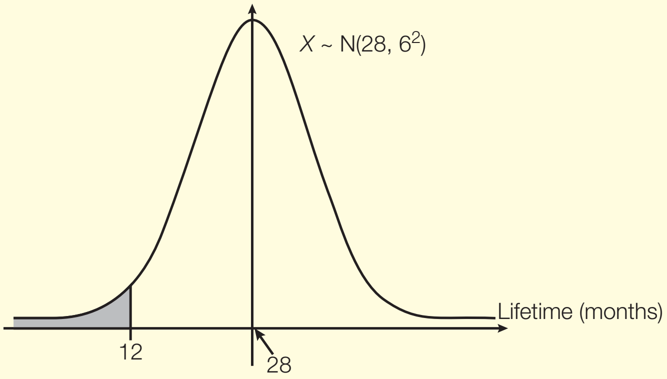
```
Area=.00383
low=-100     up=12
```

CASIO



```
Area=.00383
low=-1E9     up=-2.6667
```

The standard deviation measures the spread around the mean; if σ is smaller, values will be more tightly clustered around the mean and so the probability of getting much shorter lifetimes will be lower.



Exercise 11.3

1. A survey of a sample of Australian cattle dogs determined the mean longevity to be 13.4 years with a standard deviation of 2.4 years.

Miki is an Australian cattle dog. Estimate the probability of Miki living:

(a) beyond 10 years

(b) between 12 and 17 years

(c) no more than 18 years.

2. The masses of adult African elephants belonging to a certain herd is known to be normally distributed with a mean of 5.5 tons and a standard deviation of 0.4 tons.

(a) Bingo is an African elephant belonging to the above-mentioned herd. Estimate the probability of his mass being:

(i) less than 6 tons (ii) greater than 4.5 tons

(iii) between 5.2 and 6.2 tons.

(b) Out of a population of 100 adult elephants, estimate how many would be expected to have a mass of:

(i) between 4.8 and 6.8 tons (ii) less than 7 tons.

(c) Do you expect any of the elephants to have a mass of more than 8 tons?

3. The reaction times of a sample of students were found to be normally distributed with a mean of 222 milliseconds and a standard deviation of 41 milliseconds.

(a) What proportion of the students had a reaction time greater than 240 milliseconds?

(b) What is the percentage of students with reaction times between 200 and 300 milliseconds?

(c) If a student is chosen at random, what is the probability of his or her reaction time being less than 180 milliseconds?

(d) Out of a student population of 600, how many of them do you expect to have a reaction time of less than 180 milliseconds?

4. Milk prices in a farming community were found to be normally distributed with a mean price of 24.99 cents per litre and a standard deviation of 2.51 cents per litre.

Estimate the percentage of farms for which the price of a litre of milk is:

(a) less than 30 cents

(b) more than 20 cents

(c) between 28 and 32 cents.

Is the normal distribution relied upon too much? Is it always a good model for continuous data collected in a wide range of contexts? It can be argued that the normal distribution does not give enough importance to events that occur on the 'edges' of a population. It is used extensively in international finance, but can it predict unusual events and warn us when the system is heading towards crisis?

5. A survey concluded that the waiting times for service in a busy restaurant are normally distributed with a mean time of 12.8 minutes and a standard deviation of 3.5 minutes.

On one particular evening, the restaurant served 240 customers. Estimate the number of customers who were served after waiting:

(a) longer than 18 minutes

(b) less than 5 minutes

(c) between 10 and 15 minutes.

The manager decided to give free desserts to all customers who waited for more than 20 minutes.

(d) Estimate the number of customers out of the total of 240 who received the free desserts.

6. The IQ test scores of students in a school were found to be normally distributed with a mean of 110 and a standard deviation of 12.

(a) Estimate the **proportion** of students in the school with IQ test scores:

(i) above 90 (ii) below 80

(iii) below 125 (iv) above 130.

(b) Given that the student population is 1600, estimate the **number** of students with IQ test scores between:

(i) 85 and 125 (ii) 130 and 140.

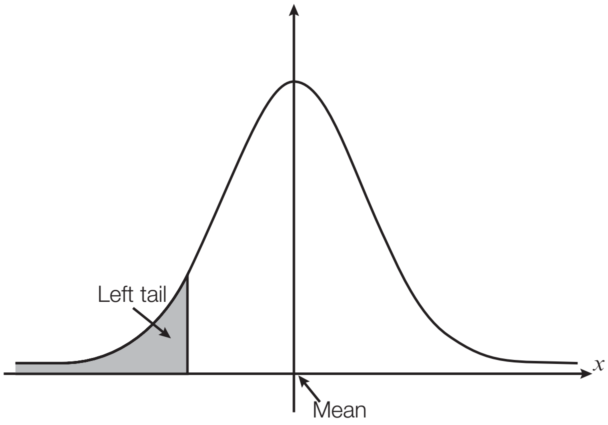
11.4 Inverse normal calculations

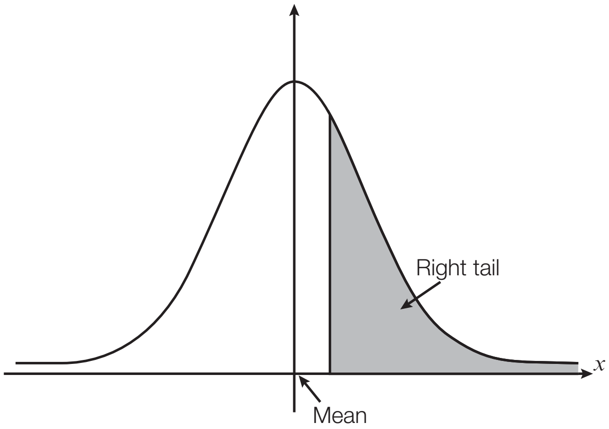
If you are given a probability value (a number between 0 and 1), your GDC can calculate the value a such that $\mathrm{P}(X \leq a)$ equals the known probability; this is called a **left tail** calculation. It can also find the value b such that $\mathrm{P}(X \geq b)$ equals the known probability, which is called a **right tail** calculation. In either case, the GDC is performing an **inverse** calculation: it is working backwards, reversing the sort of calculations you have been doing in sections 11.2 and 11.3. See '*11.2 Inverse normal calculations*' on page 671 of the GDC chapter for a reminder of how to use your GDC to calculate inverse normal calculations.

For inverse normal calculations, it is even more helpful to draw a diagram, as this enables you to predict whether the answer should be greater than or less than the mean.

First, you need to establish whether you are looking at the left or right tail of the normal distribution. Then, shade the area that represents the probability you have been given.

Remember that by the symmetry of the normal curve, exactly half the area lies on each side of the mean. So if the given probability is greater than 0.5, the area must extend across the line of symmetry. Conversely, if the given probability is less than 0.5, the area must lie entirely on one side of the mean:





If the given probability is **less than 0.5** and you are looking for a **left tail** value, the answer should be to the **left** of the mean (the value is less than the mean). If the given probability is **less than 0.5** and you are looking for a **right tail** value, the answer should be to the **right** of the mean (the value is greater than the mean).

If the given probability is **greater than 0.5** and you are looking for a **left tail** value, the answer should be to the **right** of the mean (the value is greater than the mean). If the given probability is **greater than 0.5** and you are looking for a **right tail** value, the answer should be to the **left** of the mean (the value is less than the mean).

Inverse standard normal distribution

Recall that the standard normal distribution has mean 0 and standard deviation 1, denoted by $\mathrm{N}(0, 1^2)$.

In this case, if the given probability is **less than 0.5** and you are looking for a **left tail** value, the answer should be **negative**; if the given probability is **less than 0.5** and you are looking for a **right tail** value, the answer should be **positive**.

If the given probability is **greater than 0.5** and you are looking for a **left tail** value, the answer should be **positive**; if the given probability is **greater than 0.5** and you are looking for a **right tail** value, the answer should be **negative**.

By default, your GDC does inverse calculations with the standard normal distribution, so if the question you want to answer is about the standard normal distribution, the only number you need to enter is the given probability.

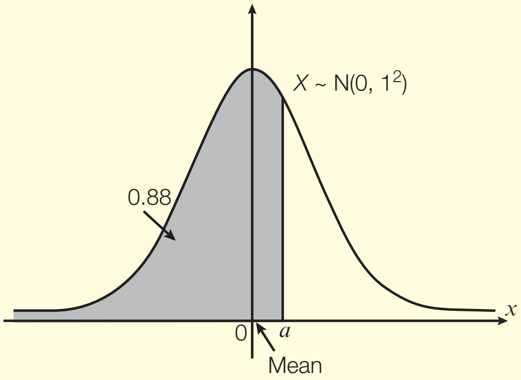
On most TEXAS GDCs, the default setting is to calculate a left tail value. On most CASIO GDCs, you have the option of choosing a right tail or a left tail.

Worked example 11.4

Q. If $X \sim N(0, 1^2)$, find the value of a for which $P(X \le a) = 0.88$.

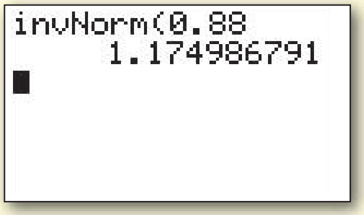
A.

Sketch the graph. $P(X \le a)$ tells you that you should be looking at a left tail. Since $0.88 > 0.5$, a must lie to the right of the mean so that the area extends across the central line of symmetry.



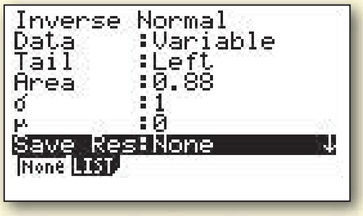
Enter the number 0.88 into the inverse normal function on your GDC. See section *11.2 Inverse normal calculations* on page 671 of the GDC chapter if you need to.

TEXAS

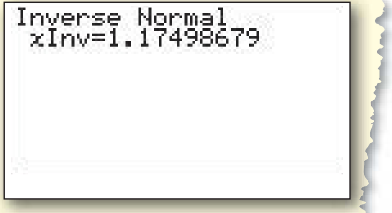


```
invNorm(0.88
      1.174986791
█
```

CASIO



```
Inverse Normal
Data    :Variable
Tail    :Left
Area    :0.88
σ       :1
μ       :0
Save Res:None      ↓
None LIST
```



```
Inverse Normal
 xInv=1.17498679
```

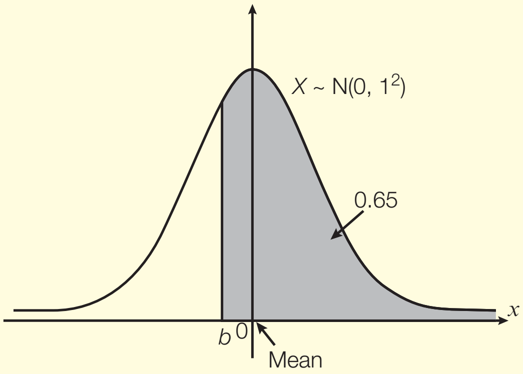
The GDC gives $a = 1.17$.

Worked example 11.5

Q. If $X \sim N(0, 1^2)$, find the value of b for which $P(X \ge b) = 0.65$.

A.

Sketch the graph. $P(X \ge b)$ tells you that you should be looking at a right tail. Since $0.65 > 0.5$, b must lie to the left of the mean so that the area extends across the central line of symmetry.

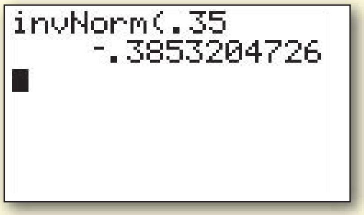


continued . . .

Use the inverse normal calculation on your GDC. On a CASIO GDC, choose 'right tail' and enter 0.65.

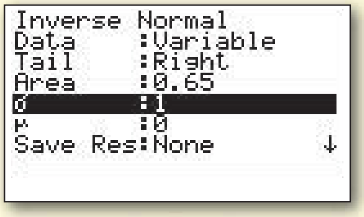
On a TEXAS GDC, because it only deals with left tails, you need to calculate the area of the left tail first. Remember that the total area under the curve is 1, and the right tail area is given as 0.65, so the left tail area = 1 – 0.65 = 0.35 and this is the number that you enter into the GDC.

TEXAS

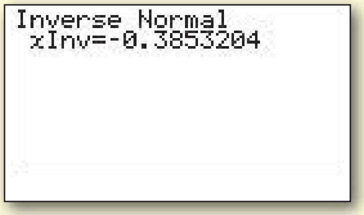


```
invNorm(.35
     -.3853204726
█
```

CASIO



```
Inverse Normal
Data      :Variable
Tail      :Right
Area      :0.65
σ         :1
μ         :0
Save Res:None        ↓
```



```
Inverse Normal
 xInv=-0.3853204
```

The GDC gives $b = -0.385$.

(Note: the negative sign shows you that the value of b is below the value of the mean.)

Inverse general normal distribution

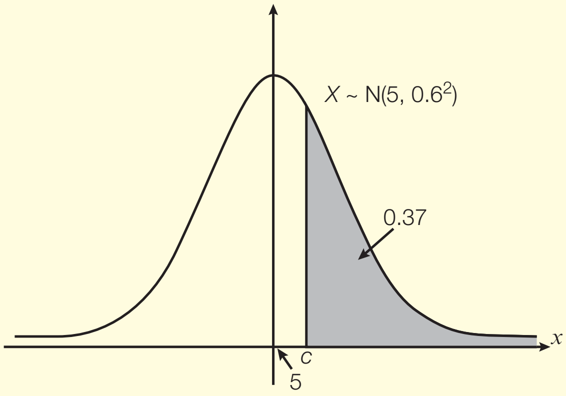
For distributions $N(\mu, \sigma^2)$ where μ is not 0 or σ is not 1, you can use your GDC for inverse calculations just as you did for the standard normal distribution, but now you also need to enter the values of the mean and standard deviation **after** the value of the probability.

Worked example 11.6

Q. If $X \sim N(5, 0.6^2)$, find the value of c for which $P(X \geq c) = 0.37$.

A.

Sketch a normal curve with mean 5 and standard deviation 0.6. $P(X \geq c)$ tells you that you are looking at a right tail. Since $0.37 < 0.5$, c must lie to the right of the mean. Shade in the relevant area.



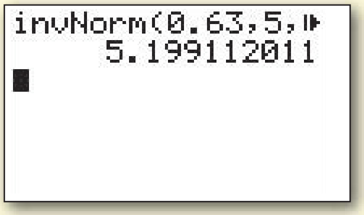
continued . . .

Again, since this is a right tail, for a TEXAS GDC you need to calculate the area of the left tail first:

left tail area = 1 – 0.37 = 0.63

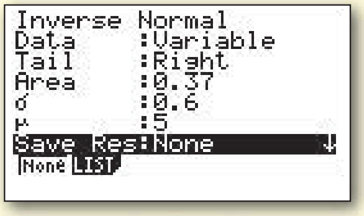
so this is the number that you enter, followed by the values of μ and σ. See section '*11.2 Inverse normal calculations*' on page 671 of the GDC chapter if you need to.

TEXAS

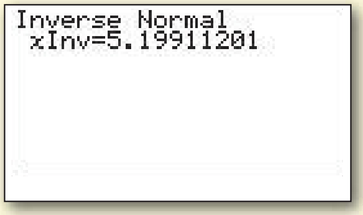


```
invNorm(0.63,5,▸
     5.199112011
```

CASIO



```
Inverse Normal
Data     :Variable
Tail     :Right
Area     :0.37
σ        :0.6
μ        :5
Save Res:None
None LIST
```



```
Inverse Normal
 xInv=5.19911201
```

From GDC: $c = 5.12$

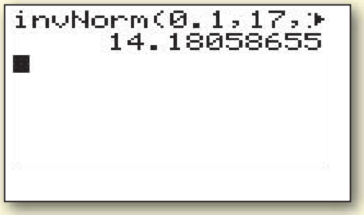
Worked example 11.7

Q. If $X \sim \mathrm{N}(17, 2.2^2)$, find the values of d and e for which $\mathrm{P}(X \le d) = 0.1$ and $\mathrm{P}(d \le X \le e) = 0.72$.

Finding d is a left tail calculation. Use the inverse nomal calculation on your GDC.

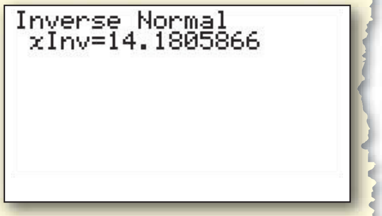
A.

TEXAS



```
invNorm(0.1,17,▸
     14.18058655
```

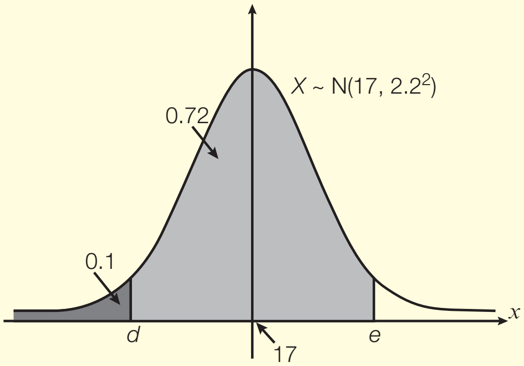
CASIO



```
Inverse Normal
 xInv=14.1805866
```

From GDC: $d = 14.2$

Sketch the graph on your GDC. From P($d \le X \le e$), we know that e should lie to the right of d. Note that P($X \le d$) + P($d \le X \le e$) = P ($X \le e$), so P($X \le e$) = 0.1 + 0.72 = 0.82. The value of e can then be found by a left tail calculation. Since 0.82 > 0.5, e must lie to the right of the mean 17.

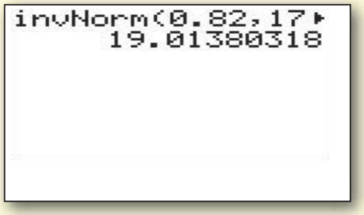


→

continued . . .

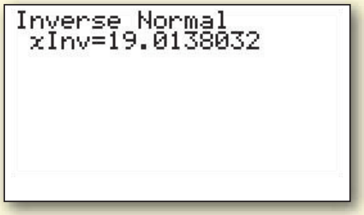
Enter 0.82, $\mu = 17$ and $\sigma = 2.2$ in a left tail calculation to find e using the inverse normal calculation on your GDC.

TEXAS



```
invNorm(0.82,17▸
      19.01380318
```

CASIO



```
Inverse Normal
 xInv=19.0138032
```

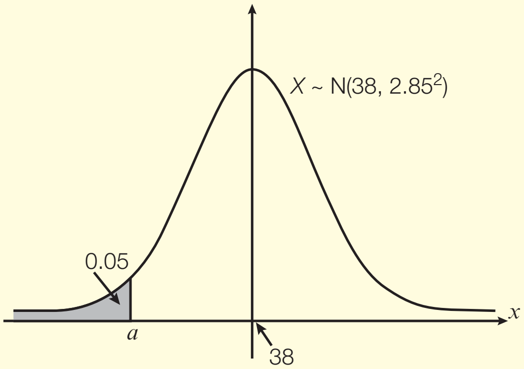
From GDC: $e = 19.0$

Worked example 11.8

Q. In a certain year, Nando estimates that 5% of his lambs have not reached the required mass for market. The mean mass of the lambs this year is 38 kg and the standard deviation is 2.85 kg. What is the minimum mass requirement for lambs to be sent to market this year?

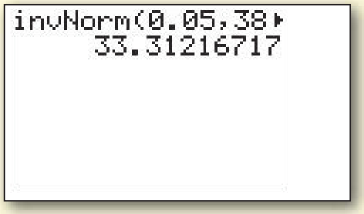
A.

Sketch the graph. Let a be the minimum mass required. If 5% of the lambs cannot be sent to market, the proportion of underweight lambs is 0.05, which means that $P(X \leq a) = 0.05$.



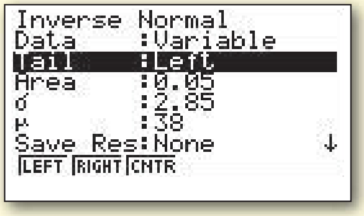
Use the inverse normal calculation on your GDC. Enter 0.05, $\mu = 38$ and $\sigma = 2.85$ on your GDC for a left tail calculation to find a.

TEXAS

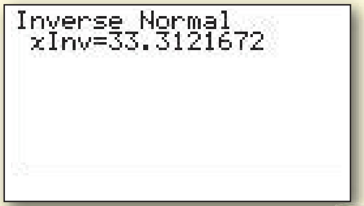


```
invNorm(0.05,38▸
      33.31216717
```

CASIO



```
Inverse Normal
Data     :Variable
Tail     :Left
Area     :0.05
σ        :2.85
μ        :38
Save Res:None        ↓
LEFT RIGHT CNTR
```



```
Inverse Normal
 xInv=33.3121672
```

The minimum mass required is 33.3 kg.

Exercise 11.4

1. It is known that a random variable X is normally distributed with mean 45 and standard deviation 9. Find $\mathrm{P}(28 \leq X \leq 66)$.

2. The BMI (body mass index) of a group of women is believed to be normally distributed with a mean of 23 and a standard deviation of 3. If a woman is chosen at random from the group, what is the probability that her BMI is no greater than 26?

3. The masses of teachers in a certain school are normally distributed with mean 65 kg and standard deviation 9.3 kg.

   (a) Mr Lee teaches at the school. What is the probability that he has a mass of more than 75 kg?

   (b) 80% of teachers in the school have a mass of less than m kg. Determine the value of m.

4. The masses of a sample of pet rabbits are normally distributed with mean 2.6 kg and standard deviation 0.4 kg. If one of the rabbits is chosen at random, what is the probability that it has a mass of between 2 kg and 3 kg?

5. Michelle has bought a sack of potatoes from the market. The masses of the potatoes are known to be normally distributed with mean of 350 g and a standard deviation of 20 g.

   (a) Michelle takes a potato randomly out of the sack. Find the probability that it has a mass of:

   (i) more than 380 g (ii) less than 390 g

   (iii) between 320 g and 405 g.

   (b) 10% of the potatoes have a mass of less than w grams.

   (i) Sketch, shade and label a normal distribution curve to represent this information.

   (ii) Hence determine the value of w.

6. Mr Gonzalez is a tomato farmer. The masses of the tomatoes from his farm are normally distributed with a mean of 160 grams and a standard deviation of 15 grams. Mr Gonzalez wants to categorise his tomatoes according to mass. The lightest 10% are to be classed as 'small', the heaviest 10% as 'large' and the rest as 'medium'. Given that the medium tomatoes weigh between x grams and y grams, find the values of x and y.

7. Zara bought a large box of table tennis balls for a regional tournament. The diameter of the balls was assumed to be normally distributed with mean 44 mm and standard deviation 1.8 mm. Zara rejected 3% of the balls for being undersized and 2% for being oversized. She decided to use the remaining balls for the tournament. What was the range of diameters of the balls which Zara declared fit for the tournament?

Learning links

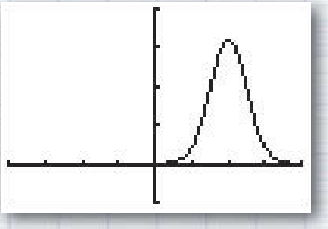
11A Transforming a normal variable to a standardised normal variable

As you have seen, it is particularly easy to do calculations with the standard normal distribution N(0, 1^2) on the GDC, as you do not need to enter the values of μ and σ. In the days before calculators and computers, people would look up normal distribution probabilities (areas under the curve) and inverse normal values from tables compiled especially for the standard normal distribution.

Any normal distribution can be **standardised**, that is, transformed to the standard normal distribution. In fact, this is what your GDC does when you give it the mean and standard deviation of a general normal distribution.

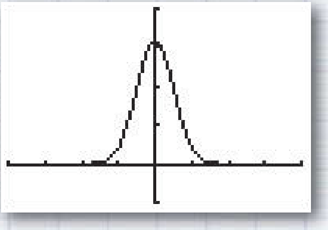
How does it do that?

Suppose that $X \sim \text{N}(2, 0.5^2)$. If we plot this normal curve on the x- and y-axes, it looks like:

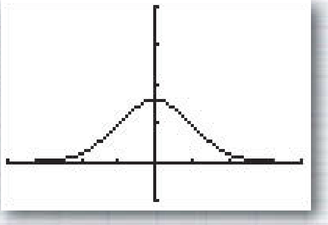


The curve is centred at the mean 2, which lies to the right of 0.

If we subtract the mean from each x value, then the curve will be shifted left by 2 units so that it is centred at 0:



This curve is narrower and taller than the standard normal distribution curve, because its standard deviation is less than 1. We can make the curve have exactly the same shape as the standard normal curve if we divide all x values by the standard deviation 0.5 (i.e. multiply them by 2):



→

continued . . .

So, by subtracting the mean from each value of the random variable X and then dividing by the standard deviation, we have obtained a variable $z = \frac{x-\mu}{\sigma}$ which follows the **standard** normal distribution. In fact, for any particular value x taken by the random variable X, the value $z = \frac{x-\mu}{\sigma}$ represents the **number of standard deviations** that x is away from the mean; $z = \frac{x-\mu}{\sigma}$ is often called the 'z-score' of x.

So, if $X \sim \text{N}(2, 0.5^2)$:

- For $x = 3.5$, $z = \frac{3.5-2}{0.5} = 3$, so the value 3.5 is 3 standard deviations above the mean.
- For $x = 1.75$, $z = \frac{1.75-3}{0.5} = -2.5$, so the value 1.75 is 2.5 standard deviations below the mean.
- Similarly, calculating $\text{P}(X \le 3.2)$ is the same as calculating $\text{P}(Z \le \frac{3.2-2}{0.5}) = \text{P}(Z \le 2.4)$ where $Z \sim \text{N}(0, 1^2)$.

Worked example

Q. The time taken to go through the security check at an airport is found to be normally distributed with a mean of 25 minutes and a standard deviation of 5.25 minutes.

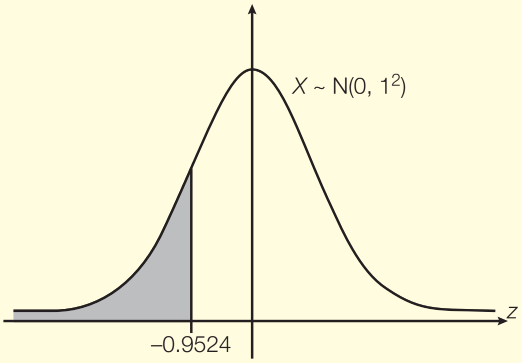
(a) What is the probability that a passenger will wait for less than 20 minutes?

(b) Ben is late and only has 34 minutes to get through security and out to the departure lounge. What is the probability that he will not be able to board his plane?

(c) The airport wants to advertise that 90% of its passengers are through the security checks within x minutes. What is the value of x?

A. (a)

If a passenger waits for 20 minutes, the z-score is $\frac{20-25}{5.25} = -0.9524$. We want to find $\text{P}(Z < -0.9524)$.

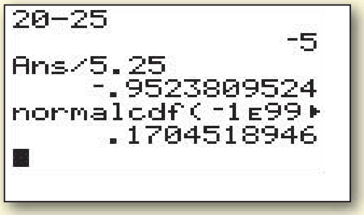


continued . . .

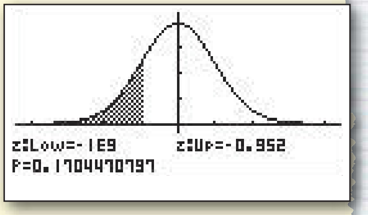
Enter the value –0.9524 into your GDC. There's no need to enter μ and σ. Use your GDC; see section '*11.1*' on pages 669-670 of the GDC chapter if you need to.

continued . . .

TEXAS **CASIO**



```
20-25
                -5
Ans/5.25
     -.9523809524
normalcdf(-1E99▸
      .1704518946
```

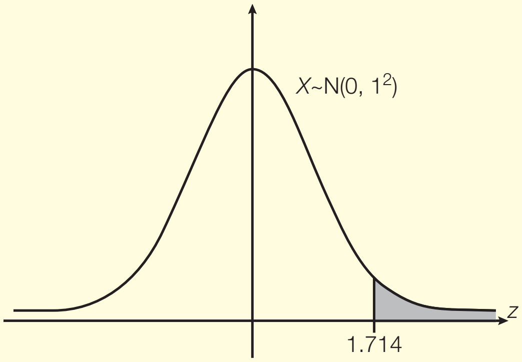


```
z:Low=-1E9    z:Up=-0.952
P=0.1704470797
```

The probability that a passenger waits for less than 20 minutes is 0.170.

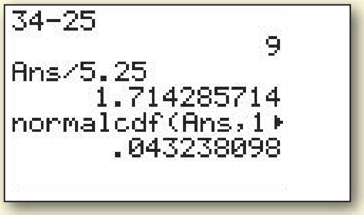
Sketch the normal curve. 34 minutes gives a z-score of $\dfrac{34 - 25}{5.25} = 1.714$. We want to find $P(Z > 1.714)$.

(b)

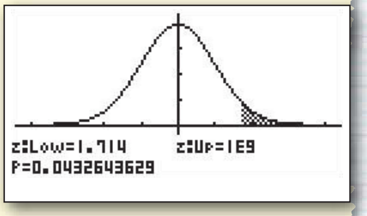


Use your GDC to calculate the probability.

TEXAS **CASIO**



```
34-25
              9
Ans/5.25
    1.714285714
normalcdf(Ans,1▸
     .043238098
```

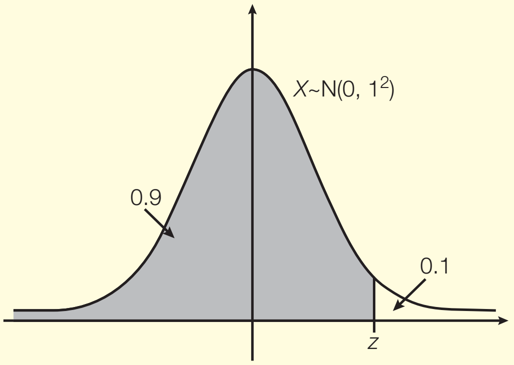


```
z:Low=1.714    z:Up=1E9
P=0.0432643629
```

The probability that Ben cannot board his plane is 0.0432.

Sketch the normal curve. We are given that $P(X \leq x) = 0.9$. This is an inverse normal calculation. In terms of $Z \sim N(0, 1^2)$, we first find $P(Z \leq z) = 0.9$.

(c)



Use the inverse normal calculation on your GDC (see section '*11.2*' on page 670 if you need to). Again, enter just the probability 0.9, without μ and σ.

The GDC gives $z = 1.282$.

continued . . .

Convert the value of z to a value of x.

$$z = \frac{x - \mu}{\sigma}$$

$$1.282 = \frac{x - 25}{5.25}$$

$$x = 1.282 \times 5.25 + 25 = 31.7$$

So 90% of passengers are through the security checks within 32 minutes.

Summary

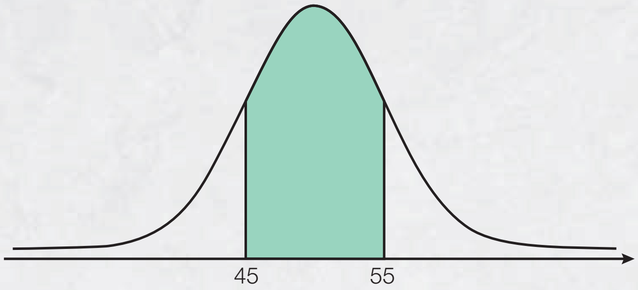
You should know:

- the concept of a random variable
- about the normal distribution, its parameters μ and σ, the bell shape, and its symmetry about $x = \mu$
- the diagrammatic representation of the normal distribution curve and areas under the curve
- how to carry out normal probability calculations using area diagrams and your GDC
- how to carry out inverse normal calculations and when they should be used.

Mixed examination practice

Exam-style questions

1. It is known that a random variable is normally distributed with a mean of 50 and standard deviation of 5.



(a) State the following probabilities:

(i) $P(45 \le X \le 55)$ (ii) $P(40 \le X \le 60)$.

(b) Find the following probabilities:

(i) $P(X \le 68)$ (ii) $P(X \ge 43)$.

2. It is known that a random variable X is normally distributed with mean 45 and standard deviation 9. Find $P(28 \le X \le 66)$.

3. A random variable X is normally distributed with mean 410 and standard deviation 29. Find:

(a) $P(X \le 380)$ (b) $P(X \ge 430)$ (c) $P(350 \le X \le 450)$.

4. The Ryder family usually book taxis from the company Quick Cabs. The waiting times for a taxi are known to be normally distributed with a mean waiting time of 19 minutes and a standard deviation of 4 minutes. Tom Ryder has just booked a taxi from Quick Cabs. Find the probability that the taxi will arrive:

(a) in less than 25 minutes (b) in more than 10 minutes (c) in 5 to 20 minutes.

5. The distances thrown in a javelin competition can be assumed to be normally distributed with a mean distance of 64.28 metres and standard deviation of 5.31 metres.

(a) Work out the probability of a randomly chosen competitor throwing the javelin beyond 70 metres.

After the first round, only 15% of the competitors progressed to the next round of the competition.

(b) Determine the minimum throwing distance needed to qualify for the next round.

Twenty-four people took part in the javelin competition. Yurek's throw was 60.7 metres.

(c) Estimate the number of competitors who performed better than Yurek.

6. The masses of drivers at a bus depot are normally distributed with mean 78 kg and standard deviation 4.7 kg.

(a) If one of the drivers is picked at random, find the probability that the driver weighs:

(i) more than 65 kg (ii) less than 85 kg.

(b) 60% of the drivers at the depot weigh more than m kg. Determine the value of m.

7. The volume of hot chocolate dispensed from a drinks machine is known to follow a normal distribution with mean 250 ml and standard deviation 8 ml. You are 'underserved' if the volume of dispensed drink is less than 230 ml.

(a) If Melanie uses the machine once, calculate the probability that she will not be underserved.

(b) Tia and Maria have just used the drinks dispenser. Calculate the probability that:

(i) neither of them has been 'underserved'

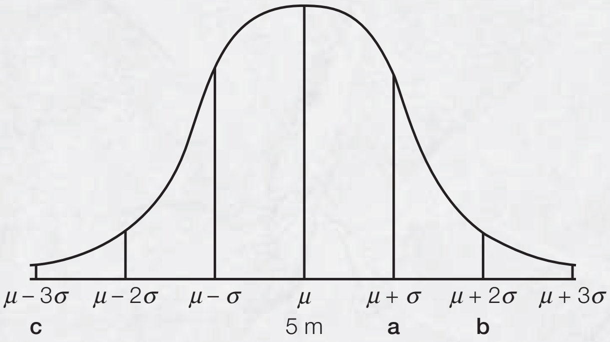
(ii) at least one of them has been 'underserved'.

(c) If the dispensing unit is adjusted and the standard deviation changes to 6 ml, do you expect more or fewer customers to be underserved? Explain your answer.

Past paper questions

1. A manufacturer makes wooden sticks with a mean length of 5 m. The lengths are normally distributed with a standard deviation of 10 cm.

(a) Calculate the values of **a**, **b** and **c** shown on the graph below.



[3 marks]

(b) What is the probability that a stick chosen at random will measure more than 4.85 m? *[3 marks]*

(c) The manufacturer sets the machine to make different sticks with a mean length of 3.5 m. It is known that 90% of the sticks will be less than 3.8 m in length. What is the standard deviation of these lengths? *[4 marks]*

[Nov 2004, Paper 2, Question 7(ii)] (© *IB Organization 2004*)

exam tip

In the examination, you will **not** be asked to calculate the mean or standard deviation using the inverse normal calculation, so part (c) is no longer within the scope of the syllabus. But we thought that you might find it interesting to try solving this problem using the methods in Learning links 11A.

2. The weights of cats form a normal distribution about a mean weight of 3.42 kg with a standard deviation of 0.82 kg.

The local veterinarian has collected data for 150 cats that have attended the surgery.

(a) (i) Write down the percentage of cats that will weigh within 1 standard deviation of the mean. *[1 mark]*

(ii) How many of the cats that visit the surgery will weigh within 1 standard deviation of the mean? *[2 marks]*

(b) (i) On a suitable bell-shaped diagram, shade in the area corresponding to all cats weighing less than 2 kg. *[1 mark]*

(ii) Calculate the standardised normal value z corresponding to 2 kg. *[2 marks]*

(iii) What percentage of cats will weigh less than 2 kg? *[2 marks]*

(c) Calculate the percentage of cats that will weigh between 2 kg and 4.8 kg. *[3 marks]*

(d) The probability of a cat weighing more than w kg is 2.5%. Find w. *[3 marks]*

[May 2004, Paper 2, Question 7(i)] (© *IB Organization 2004*)

exam tip

Part (b)(ii) is no longer on the syllabus so you will not have to answer questions like this in an examination. However, you might enjoy trying it with the help of Learning links 11A.

Chapter 12 Correlation

In this chapter you will learn:

- about bivariate data
- about the concept of correlation, scatter diagrams and the line of best fit
- how to draw lines by hand to fit data plotted on a scatter diagram
- about Pearson's product moment correlation coefficient
- how to interpret strong or weak, positive, zero or negative correlations
- how to find the regression line for y on x
- how to use the regression line to make predictions.

In Darrell Huff's book *How to Lie with Statistics*, he refers to a study that claimed that non-smoking college students gained better grades than students who smoked. The study concluded that 'smoking leads to poor college grades'. Huff questioned the findings of the study, arguing that there are other interpretations besides this obvious one. For instance, instead of smoking being the cause of poor grades, perhaps 'low grades depressed the students and caused them to smoke', or maybe 'students with lower grades were more sociable and therefore more likely to smoke'. Measuring the relationship between two sets of data is a useful tool to establish how one variable changes with another, but it is limited; it does not tell you *why* this relationship exists. The data in the study clearly showed a relationship between smoking and poor grades, but it does not explain why this relationship exists.

12.1 The concept of correlation

There are many situations in which a statistician might like to know if two variables are related and how they are related.

Data that consists of measurements of two variables taken from each individual in a sample is called **bivariate data**; the relationship between the two variables is referred to as the **correlation**. An example of bivariate data is the mass and height of boys on a basketball team; or the size and rate of photosynthesis of a leaf.

The concept of correlation is used frequently in the biological and social sciences.

For bivariate data we try to classify one of the two variables as 'independent' and the other as 'dependent'. The **independent variable** is the one that can be controlled by the person conducting the experiment or study; it is hypothesised to 'cause' some kind of effect to the dependent variable. The **dependent variable** is the variable that is just observed without being controlled, and is supposed to show the 'effect'.

For example:

- Is the stretch of someone's hand dependent on his or her height? The independent variable is the person's height, and the dependent variable is the length of their hand span.
- If a company spends more on advertising, will it sell more of its product? The independent variable could be the money spent on advertising each month, and the dependent variable would be the total monthly sales.
- Is your IB Mathematical Studies grade related to the hours of music that you listen to? The dependent variable is the IB grade and the independent variable is the average number of hours per week spent listening to music.

Exercise 12.1

1. In each of the following situations, identify which is the independent variable and which is the dependent variable and comment on whether you expect any correlation between them.

   (a) The amount of alcohol Terry consumes and his reaction time.

   (b) The number of people in a household and the monthly expenditure on food.

   (c) Saba's body mass and the number of hours she spends exercising each week.

   (d) The amount of time Dave spends exercising in the gym and his blood sugar level.

   (e) The value of a second-hand car and the mileage on the car's odometer.

   (f) The length of a person's middle finger and their sprint time/speed.

   (g) The screen size of an LED TV set and its price.

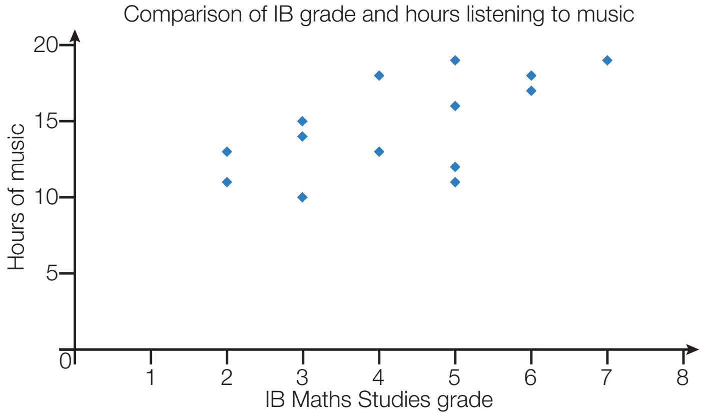
12.2 Scatter diagrams

The table below lists the Mathematical Studies grades (x) of students in an IB class and the number of hours of music that each student listens to each week (y).

x	2	3	3	5	5	7	2	3	4	6	5	5	4	6	5
y	11	14	15	16	16	19	13	10	13	18	19	12	18	17	11

Can you tell from the numbers if there is any relationship between how much music they listen to and the grade they achieve?

A diagram plotting these pairs of (x, y) values makes it easier to see a possible relationship:

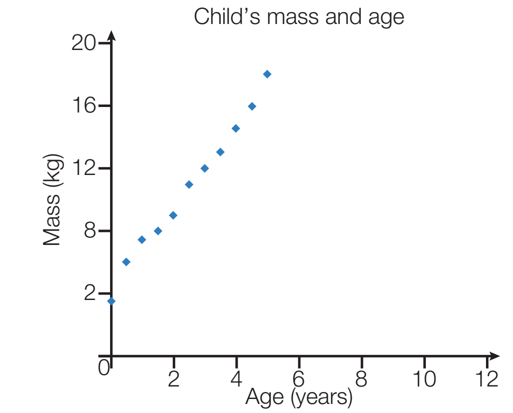


From the diagram you can see that there is a possible connection between a student's grade and the hours of music that they listen to.

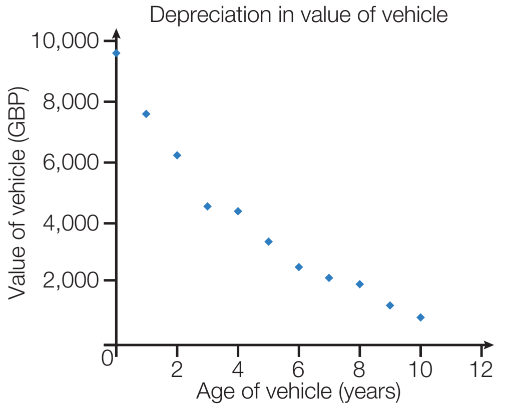
This diagram is called a **scatter diagram** (also called a scatter graph or scatter plot). Scatter diagrams help to illustrate the connection between the two variables of a bivariate data set. For each pair of data values, one value is plotted along the *x*-axis and the other along the *y*-axis.

In general, you plot the independent variable along the *x*-axis and the dependent variable along the *y*-axis; in other words, the *x*-coordinate of each plotted point is a data value of the independent variable, and the *y*-coordinate is the corresponding data value of the dependent variable.

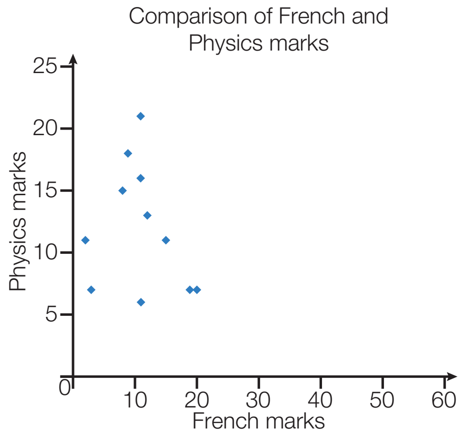
A scatter diagram shows the type of connection between the two variables. There are three principal types of correlation: positive, negative or zero (no) correlation.



Positive correlation:
The dependent variable increases as the independent variable increases (*y* increases with *x*).



Negative correlation:
The dependent variable decreases as the independent variable increases (*y* decreases with *x*).



No correlation:
The points seem to be scattered randomly and no pattern can be seen.

Correlation and causation

Although the diagram for IB Maths Studies grade and hours spent listening to music tells you there is a possible connection between the two variables, can you really say that listening to music improves your grade? Or could it be that students with higher grades generally have more time to listen to music?

It is very important not to confuse correlation with causation. This is an extremely common mistake. Does listening to music *cause* your higher maths grade? If two things tend to happen together, it does not necessarily mean that one caused the other, even if you would like that to be the case!

You should always question any correlation that you find, and ask yourself if the relationship is truly causative. Could there be a third factor that is influencing both of the variables?

For instance, suppose a doctor is claiming that his studies show that heart disease is linked to decaying teeth. His figures show that the worse someone's teeth are, the more likely that person is to have heart disease.

But does having bad teeth actually cause heart disease? Or could it be that people who do not look after their teeth also tend to take little or no exercise and have a bad diet?

How do you distinguish between correlation and causation? Firstly, to isolate the relationship you are interested in, it is important to identify all the variables that might be involved, and fix all variables other than the two being examined at the same level. In the bad teeth and heart disease example, the gender, age, diet and exercise habits of subjects would need to be kept constant across the sample. But how easy is this to achieve in practice?

Exercise 12.2

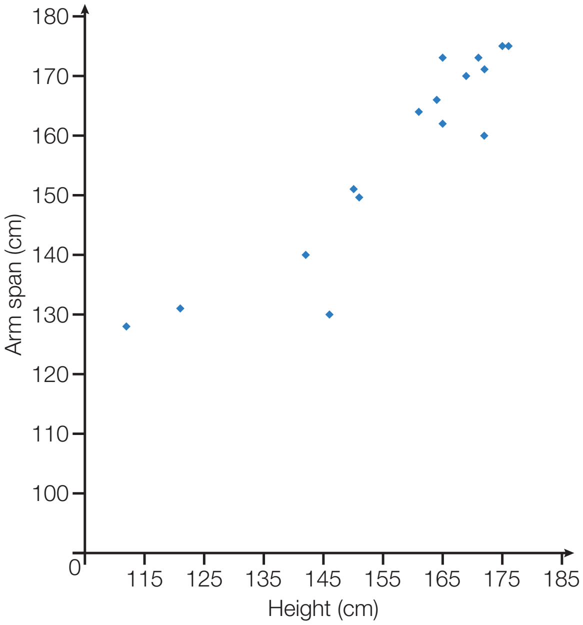
1. For each of the following scatter diagrams, identify:

(i) the type of correlation

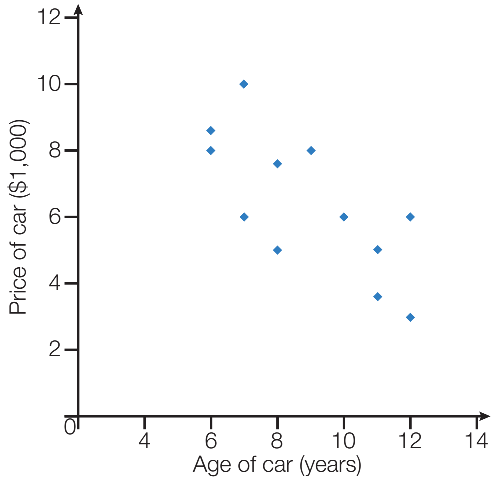
(ii) the independent variable

(iii) the dependent variable.

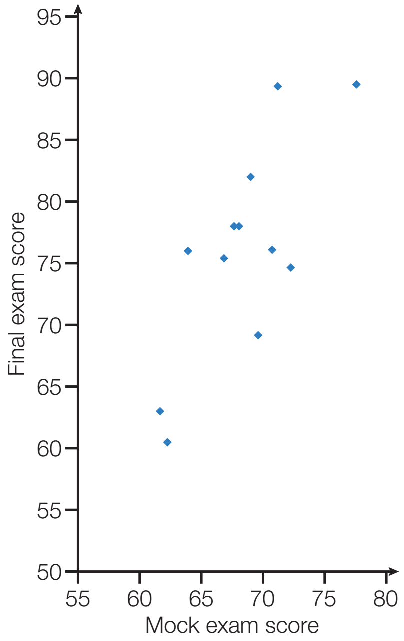
(a) The relationship between the height and arm span of a group of students:



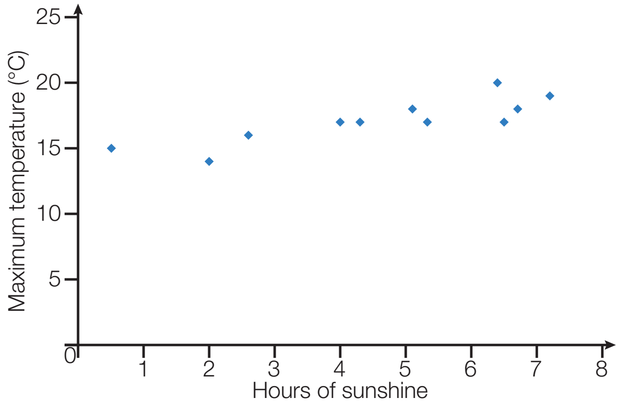
(b) The relationship between the age and price of second-hand cars.



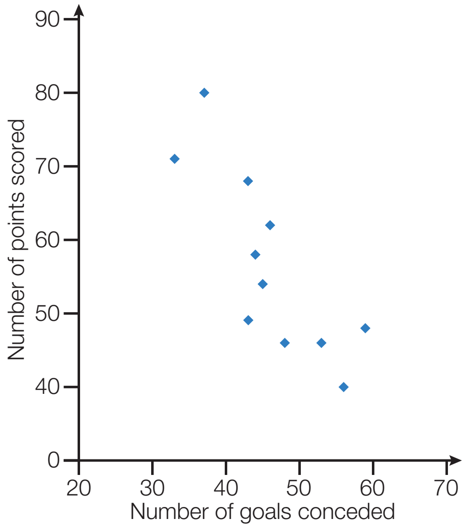
(c) The relationship between the mock and final examination scores of a group of students:



(d) The relationship between the hours of sunshine per day and the maximum temperature across eleven different cities:



(e) The relationship between the number of goals conceded and the number of points scored by teams in a football league:



12.3 Line of best fit

The **line of best fit** on a scatter diagram is drawn to give the best representation of the correlation between the two variables. The line of best fit is the one that has an approximately even spread of the data points either side of it.

Lines of best fit can be drawn by hand, on your GDC, or using a spreadsheet program on a computer.

If you have already plotted a scatter diagram on graph paper, you can draw a line of best fit by taking a long transparent ruler and adjusting its position so that the scattered points look as balanced as possible on either side of the line.

The line should be drawn so that it goes directly through the **mean point**. This is the point whose x-coordinate is the mean of the data values plotted along the x-axis and whose y-coordinate is the mean of the data values plotted along on the y-axis.

Looking again at the data on IB Mathematical Studies grade (x) and the number of hours per week (y) that a student spends listening to music:

x	2	3	3	5	5	7	2	3	4	6	5	5	4	6	5
y	11	14	15	16	16	19	13	10	13	18	19	12	18	17	11

Mean of x-values: $\bar{x} = \frac{65}{15} = 4.33$

Mean of y-values: $\bar{y} = \frac{222}{15} = 14.8$

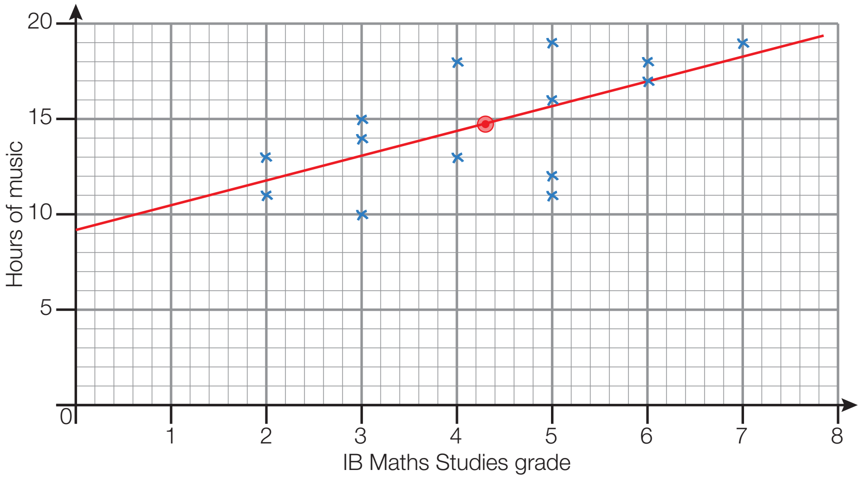
RR ***The mean is the sum of all the data values divided by the total frequency of data values. See Chapter 6.***

So the mean point of this set of bivariate data is (4.33, 14.8).

exam tip

A hand-drawn scatter diagram and line of best fit can be very helpful in your initial analysis of data collected for your project. Make sure that you always show the mean point both in the project report and in answers to examination questions.

Draw a line on the scatter diagram. It must pass through the mean point (4.33, 14.8), and the other plotted points should look approximately evenly distributed above and below it.



If you use your GDC or a computer program to draw a line of best fit, it will use similar principles to those used to draw it by hand. Using your GDC can be a quicker way to establish if there is a correlation between two variables than plotting a scatter diagram by hand, but if you are specifically asked to draw a scatter diagram, you should do it by hand. (See section '*12.1 Drawing a scatter diagram of bivariate data*' on page 672 of the GDC chapter if you need a reminder of how to use your GDC to draw a scatter diagram).

Exercise 12.3

1. For each of the following sets of data, plot a scatter diagram by hand, and draw the line of best fit. Remember to include the mean point.

(a)

Minimum temperature (°C)	10	10	11	8	7	7	11	10	10	11
Hours of sunshine	5.3	4.1	3.5	5.4	6.7	6.5	4	4.3	2.6	2

(b)

Number of goals scored	78	69	60	72	55	59	51	49	48	45	56
Number of points scored	80	71	71	68	62	58	54	49	48	47	47

(c)

Height (cm)	148	162	161	152	167	172	176	179	172	157
Circumference of neck (cm)	11	26	30	30	32	30	33	41	33	35

(d)

Population (millions)	9.4	10.8	4.9	5.6	7.2	5.8	3.4	9.8	4.1	6.3
Income per capita (US$,000)	23	20	36	31	22	27	44	14	33	21

(e)

Exam paper 1 mark (%)	86	74	74	83	64	88	88	86	70	84	85
Exam paper 2 mark (%)	94	80	73	95	42	95	85	82	71	73	89

12.4 Pearson’s product moment correlation coefficient, *r*

When you are analysing bivariate data, a hand-drawn scatter diagram can only give an approximate idea of the strength of the correlation. Moreover, fitting a line by hand is often not going to be very accurate, and different people are likely to draw different lines and come to slightly different conclusions. It is clear that a more dependable method is needed.

Pearson’s product moment correlation coefficient (PMCC) is a complicated name for a simple idea. It is a number, usually denoted by r, which can take any value between -1 and $+1$. Its sign indicates the type of correlation, and its magnitude (size) indicates the strength of correlation.

- $r = +1$ means that there is perfect positive correlation
- $r = 0$ means that there is no correlation
- $r = -1$ means that there is perfect negative correlation.

Values of r calculated from data usually lie between $+1$ and -1 and can be roughly interpreted as follows:

$r = -1$	Perfect negative correlation
$-1 < r \leq -0.75$	Strong negative correlation
$-0.75 < r \leq -0.5$	Moderate negative correlation
$-0.5 < r \leq -0.25$	Weak negative correlation
$-0.25 < r < 0$	Very weak negative correlation
$r = 0$	No correlation
$0 < r < 0.25$	Very weak positive correlation
$0.25 \leq r < 0.5$	Weak positive correlation
$0.5 \leq r < 0.75$	Moderate positive correlation
$0.75 \leq r < 1$	Strong positive correlation
$r = 1$	Perfect positive correlation

exam tip

If $-0.5 < r < 0.5$, it is difficult to draw a line of best fit that has any meaning, because the data is just too scattered. It is important to keep this in mind if you are going to use scatter diagrams in your project.

Calculating the PMCC using your GDC

It is much quicker and easier to use your GDC to calculate the value of r than it is to do it by hand. (See Learning links 12A if you want to know how it is done by hand.)

For example, let us use the data below:

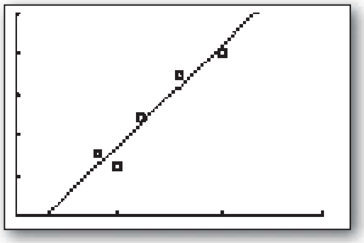
Distance, x (km)	4	8	5	10	6
Time, y (minutes)	15	35	12	40	24

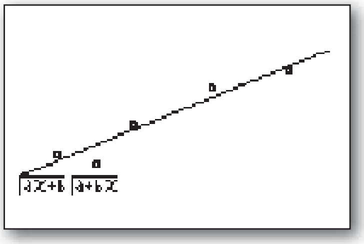
Plot the data on a scatter diagram and draw the line of best fit:

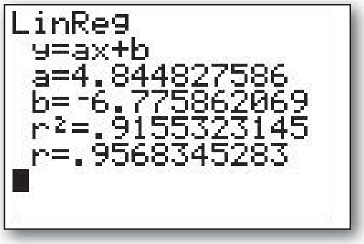
(See ‘*12.2 Finding the product moment correlation coefficient and the equation of the regression line y on x*’ on page 674 of the GDC chapter for a reminder of how to do this.)

TEXAS

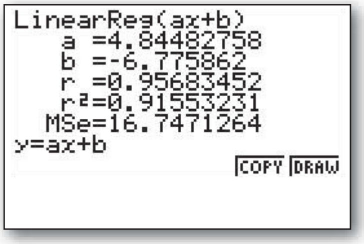
CASIO







```
LinReg
 y=ax+b
 a=4.844827586
 b=-6.775862069
 r²=.9155323145
 r=.9568345283
```



```
LinearReg(ax+b)
   a =4.84482758
   b =-6.775862
   r =0.95683452
   r²=0.91553231
  MSe=16.7471264
y=ax+b
                 COPY DRAW
```

From GDC: $r = 0.957$ (3 s.f.)

As well as the value of r, the linear regression function gives you more information:

- It tells you the equation of the line of best fit, $y = ax + b$. For this example, $a = 4.84$ and $b = -6.78$, so the equation of the line is $y = 4.84x - 6.78$.
- It also displays r^2, called the 'coefficient of determination', which measures the strength of correlation but doesn't indicate what type it is.

FF ***You will learn more about linear regression in section 12.5.***

Calculating the PMCC using a spreadsheet

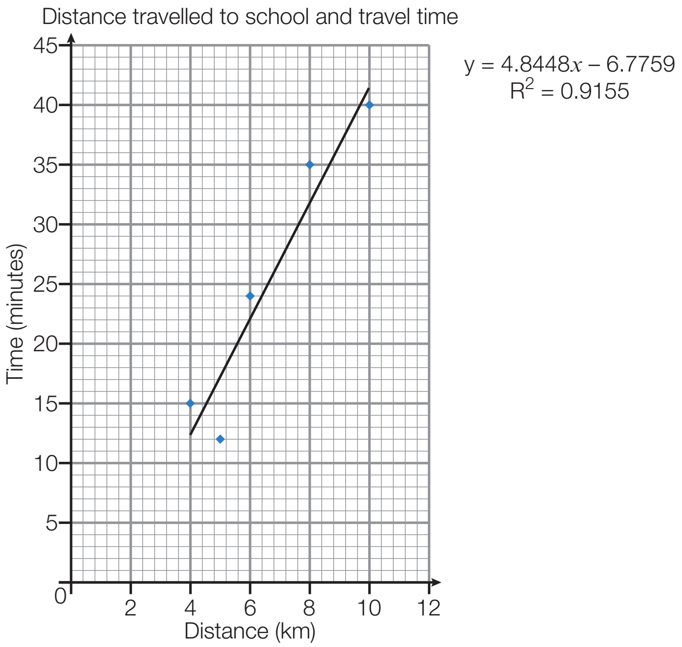
It is convenient to use a spreadsheet program on a computer to investigate the correlation between two variables. The steps are generally as follows:

1. Enter the x values in column A and the y values in Column B.
2. Select the cells containing the data and choose the scatter graph option; the scatter plot will then be drawn by the software.
3. Choose the option to add the line of best fit; this is usually called a **trend line**. If you select the options to 'display equation on chart' and to 'display R-squared value on chart', then the program will draw the line of best fit and label it.

exam tip

If an examination question asks for the PMCC, you are expected to use your GDC.

Karl Pearson (1857–1936) was interested in biometry, the use of mathematical methods to study biology. His work was founded on the concept that the scientific method is more 'descriptive' than 'explanatory', and he developed statistical methods to provide accurate mathematical descriptions of biological data. Between 1893 and 1912 he wrote 18 papers entitled *Mathematical Contributions to the Theory of Evolution*. Pearson emphasised the idea that correlation can be measured. In 1911 he established the world's first university statistics department at University College, London.



If you take the square root of the R^2 value displayed on the spreadsheet chart, you will get the PMCC, $r = 0.957$.

Learning links

12A Calculating the PMCC by hand

In most cases you will calculate r using a GDC, a statistics package on a computer, or a spreadsheet. However, it is often easier to gain understanding of an idea if you do a few calculations 'by hand' rather than relying on a calculator or computer.

The formula for the PMCC is:

$$r = \frac{s_{xy}}{s_x \times s_y}$$

where

- s_{xy} is the **covariance** of x and y
- s_x is the standard deviation of the x values
- s_y is the standard deviation of the y values.

RR ***s_x and s_y are an alternative to the notation σ_x and σ_y for standard deviations, which you may see in other statistics textbooks; recall that we also used σ to denote standard deviations in Chapters 7 and 11.***

continued . . .

The formulae for s_x and s_y are as follows:

$$s_x = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2} \qquad s_y = \sqrt{\frac{\sum y^2}{n} - \bar{y}^2}$$

RR **_Recall from Chapter 6 that the $\sum$ symbol means 'add up all the values of'. So $\sum x^2$ means 'add up all the values of x squared'._**

Covariance is another measure, related to the standard deviation, of the connection between two variables. The formula for the covariance is

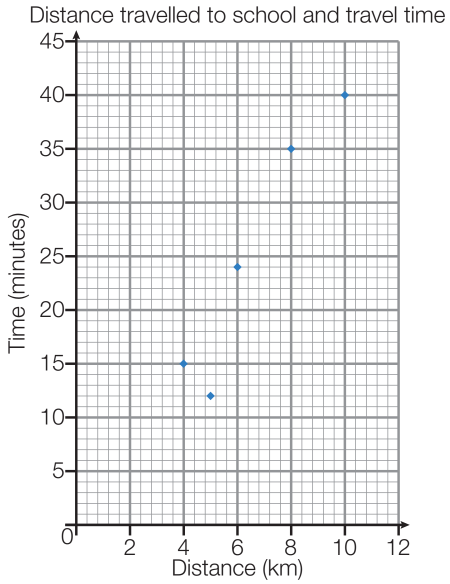
$$s_{xy} = \frac{\sum xy}{n} - \bar{x} \times \bar{y}$$

Let's take a simple example to see how the PMCC can be calculated 'by hand'.

Five IB students decided to see if there was any correlation between the distance that they travelled to school and the duration of the journey. All five students travelled by bus.

Distance, x (km)	4	8	5	10	6
Time, y (minutes)	15	35	12	40	24

The scatter diagram of their data shows that there is positive correlation between the two variables.



The most organised way of doing the calculation is to draw up a table listing all the values you need in order to apply the formulae.

→

continued . . .

	x	y	xy	x^2	y^2
	4	15	60	16	225
	8	35	280	64	1225
	5	12	60	25	144
	10	40	400	100	1600
	6	24	144	36	576
Totals	33	126	944	241	3770
	$\bar{x} = \frac{33}{5} = 6.6$	$\bar{y} = \frac{126}{5} = 25.2$			

Then substitute values into the formulae:

$$s_{xy} = \frac{\sum xy}{n} - \bar{x} \times \bar{y} = \frac{944}{5} - 6.6 \times 25.2 = 22.48$$

$$s_x = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2} = \sqrt{\frac{241}{5} - 6.6^2} = 2.154$$

$$s_y = \sqrt{\frac{\sum y^2}{n} - \bar{y}^2} = \sqrt{\frac{3770}{5} - 25.2^2} = 10.907$$

Hence

$$r = \frac{22.48}{2.154 \times 10.907} = 0.957$$

The value of r is close to +1, showing that there is strong positive correlation between the distance travelled to school and the time taken.

Exercise 12.4

1. Determine the value of Pearson's product moment correlation coefficient r in each of the following cases.

(a)

x	1	4	6	1	10	5	2	10
y	20	18	13	23	10	18	18	13

(b)

x	600	515	618	551	493	595	515
y	21	29	64	69	13	9	77

(c)

x	344	269	194	171	339	221	349	330	272
y	165	185	262	300	170	180	99	95	173

(d)

x	70	50	35	48	78	68	69	39	73	44	35	46
y	54	38	41	37	70	54	63	41	58	33	29	48

(e)

x	9.8	1.7	10.7	7.4	6.2	3.4	9.3	7.1	10.1	9.3
y	20.5	5.2	22.1	14.3	15.3	11.9	22.6	18.8	18.6	17

If data collected from a scientific experiment suggests that there is a connection between two variables, plotting a graph to make a visual check is a good starting point. However, to determine the type and strength of the correlation, and be able to make predictions based on the findings, we need to use formulae and precise calculations to obtain quantitative results such as the PMCC and the regression line. This is an approach that has been used widely across all the sciences.

12.5 Regression line of y on x

The **regression line** is another type of line of best fit for a set of bivariate data. It is obtained by minimising the sum of the squared distances between the data points and the line. It takes into account the distance of each point from the line of fit and then takes the square of this distance and minimises the total of all the squared distances so that the line represents the data as closely as possible.

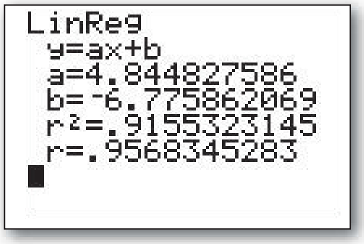
The phrase 'y on x' just means that the scatter diagram is drawn with x as the independent variable and y as the dependent variable, and that the squared distances being minimised are distances from the line in the y direction.

The regression line has an equation in the form $y = mx + c$.

Calculating the regression line using a GDC

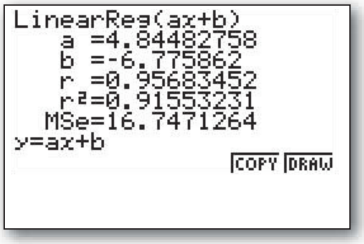
You will be expected to use a GDC to find the regression line in all exercises and examinations. When you calculate r (the PMCC) using your GDC you will get the equation of the regression line at the same time, so the instructions are identical. (See '*12.2 Finding the product moment correlation coefficient and the equation of the regression line y on x*' on page 674 of the GDC chapter if you need to.)

TEXAS



```
LinReg
 y=ax+b
 a=4.844827586
 b=-6.775862069
 r²=.9155323145
 r=.9568345283
```

CASIO



```
LinearReg(ax+b)
   a =4.84482758
   b =-6.775862
   r =0.95683452
   r²=0.91553231
  MSe=16.7471264
y=ax+b
```

When you read the equation of the regression line from the screen, remember that 'a' is the slope of the line (the gradient, m) and 'b' is the y-intercept (c).

So the equation (to three significant figures) is $y = 4.84x - 6.78$.

Learning links

12B Calculating the regression line by hand

You will be expected to use a GDC to find the regression line in all exercises and examinations; but, as usual, you might find the idea easier to understand if you practise a few calculations by hand.

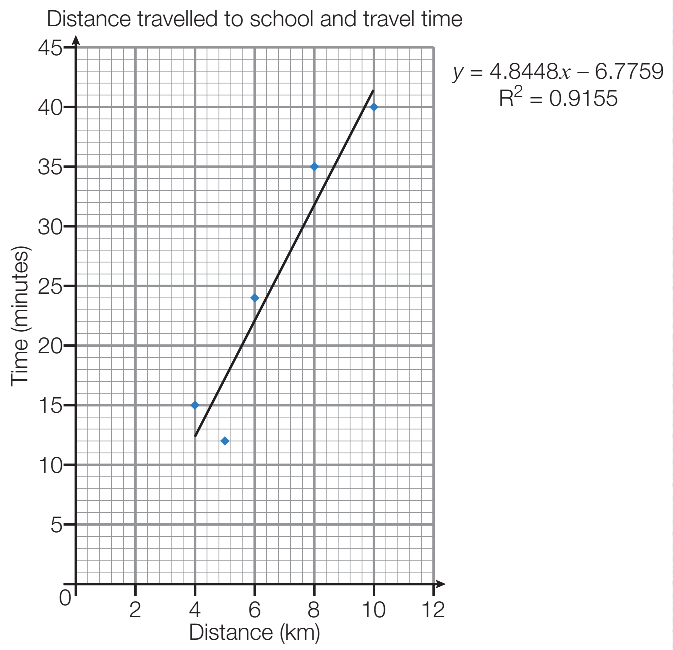
The formula for the regression line is:

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

where, as defined in Learning links 12A, s_{xy} is the covariance of x and y, s_x is the standard deviation of the x values, $\bar{x}$ is the mean of the x values, and $\bar{y}$ is the mean of the y values.

This formula is outside the IB Mathematical Studies syllabus, but if you want to know more about where it came from, you can find the explanation in any statistics textbook.

Let's revisit the example from Learning links 12A about the relationship between journey distance and travelling time:



We already calculated the following quantities in Learning links 12A:

$$\bar{x} = 6.6,\ \bar{y} = 25.2,\ s_x = 2.154,\ s_{xy} = 22.48$$

So the equation of the regression line is:

$$y - 25.2 = \frac{22.48}{(2.154)^2}(x - 6.6)$$

$$y - 25.2 = 4.845(x - 6.6)$$

$$y - 25.2 = 4.845x - 31.98$$

$$y = 4.85x - 6.78 \text{ (3 s.f.)}$$

Worked example 12.1

Q. At a local garage, Pim collected data on cars with 1.4L engines. For cars in his sample that are of the same make and model, he looked at the relationship between the age of the car and its value. He expects a negative correlation, as older cars generally cost less than newer ones.

Age of car (years)	3	5.5	4	2	6	1.5
Value of car (€)	8850	6500	7995	9150	5495	9950

Help Pim find:

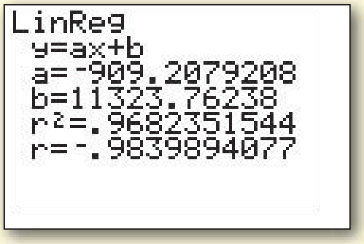
(a) the product moment correlation coefficient for his data

(b) the equation of the regression line.

Let the independent variable x be the age of a car, because we expect that it has an effect on the dependent variable, y, which would be the value of the car.

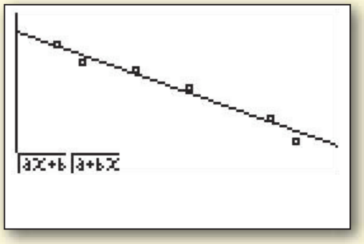
A.

TEXAS



```
LinReg
 y=ax+b
 a=-909.2079208
 b=11323.76238
 r²=.9682351544
 r=-.9839894077
```

CASIO



Use your GDC to find r and the slope and y-intercept of the regression line. See section '*12.2 Finding the product moment correlation…*' on page 674 of the GDC chapter if you need to.

(a) From GDC: $r = -0.984$, so there is a strong negative correlation between the age of the car and its re-sale value.

(b) $y = -909x + 11300$ (to 3 s.f.)

Exercise 12.5

You are expected to use your GDC to answer all the questions in this exercise.

1. Tasia has collected the following data on the ages of 12 married couples. He is investigating whether there is a correlation between the age of the husband and the age of the wife.

Age of husband (x years)	65	70	49	55	60	56	79	73	43	72	79	53
Age of wife (y years)	59	66	35	48	61	39	77	46	35	71	74	53

(a) Find the equation of the regression line of y on x.

(b) Calculate the correlation coefficient r of the data.

(c) Comment on the value of r.

2. Katarina is investigating the relation between a person's arm span and the length of their forearm. The sample data collected from 11 students is shown in the table below.

Arm span (*x* cm)	160	150	179	184	169	160	157	163	157	173	174
Length of forearm (*y* cm)	26	23	25	31	25	24	27	27	24	25	26

(a) Find the equation of the regression line of *y* on *x*.

(b) Calculate the correlation coefficient *r* of the data.

(c) Comment on the value of *r*.

3. The following table shows the population and income per capita of ten countries from the same economic zone.

Country	1	2	3	4	5	6	7	8	9	10
Population, *x* (millions)	4.49	14.86	7.5	3.76	2.59	9.15	6.70	10.18	4.67	7.95
Income per capita, *y* (USD)	682	189	353	668	950	266	355	230	491	287

(a) Write down Pearson's product moment correlation coefficient, *r*, of the data.

(b) Comment on your value of *r*.

(c) Find the equation of the regression line of *y* on *x*.

(d) Use your equation to estimate the income per capita of a country from the same zone with a population of 5.5 million.

4. Jamil is investigating the relationship between the arm span and the length of the right foot of his friends. His sample data is illustrated in the table below.

Arm span, *x* (cm)	167	171	167	152	173	126	170	160	173	182
Length of right foot, *y* (cm)	24	25	24	23	26	19	23	25	26	26

(a) Write down the equation of the regression line of length of right foot (*y*) on arm span (*x*).

(b) Use your equation from part (a) to predict the length of the right foot corresponding to an arm span of 165 cm.

(c) Write down the correlation coefficient, *r*.

(d) Describe the nature of the correlation between arm span and length of right foot, based on Jamil's data.

5. The following table shows data on the interest rate and rate of inflation of ten countries in the same region that have similar economies.

Country	1	2	3	4	5	6	7	8	9	10
Interest rate, *x* (%)	5.50	8.25	9.25	4.00	3.25	20.00	3.50	12.50	9.50	6.00
Inflation rate, *y* (%)	5.30	8.49	9.40	5.24	1.80	14.13	3.10	8.39	7.80	5.40

(a) Write down the correlation coefficient, r, of the data.

(b) The equation of the line of regression of y on x is of the form $y = mx + c$. Determine the values of m and c.

(c) A similar country has an interest rate of 16.00%. Estimate the inflation rate for this country.

12.6 Using the equation of the regression line

Can you reliably use the equation of the regression line to make predictions? Even if you are working beyond the range of the original data?

The equation of the regression line can be used to estimate values that are not in the original data. This can be very useful if the data you have collected contains insufficient detail or if you wish to make a prediction about some values that you have not collected.

However, you must be very careful: it is easy to get unreasonable values out of the equation of the regression line, leading to nonsense predictions.

In general, do **not** use the regression line to predict new values if:

- the correlation is weak: $-0.5 < r < +0.5$
- the predicted values are outside the range of the data values already collected (this is called **extrapolation**).

exam tip

This is a mistake made by many students doing a statistical project. If your bivariate data show weak correlation, make sure that you comment on this.

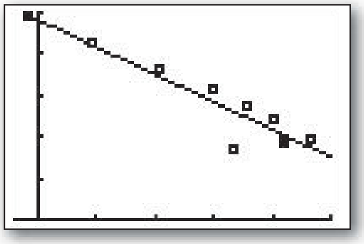
Here is an example where a seemingly good set of data can lead to a false prediction.

The table shows a subset of data collected by a weather balloon. The data consists of the different temperatures recorded as the balloon rises.

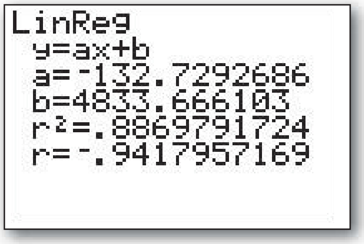
Temperature, *x* (°C)	16.6	21.0	21.0	23.4	20	17.8	15.0	10.3	4.7	–0.8
Height, *y* (m)	1700	1850	1920	1960	2440	2740	3140	3660	4270	4880

The scatter graph shows a strong correlation, and the value of r confirms this.

Scatter graph:



PMCC:



```
LinReg
 y=ax+b
 a=-132.7292686
 b=4833.666103
 r²=.8869791724
 r=-.9417957169
```

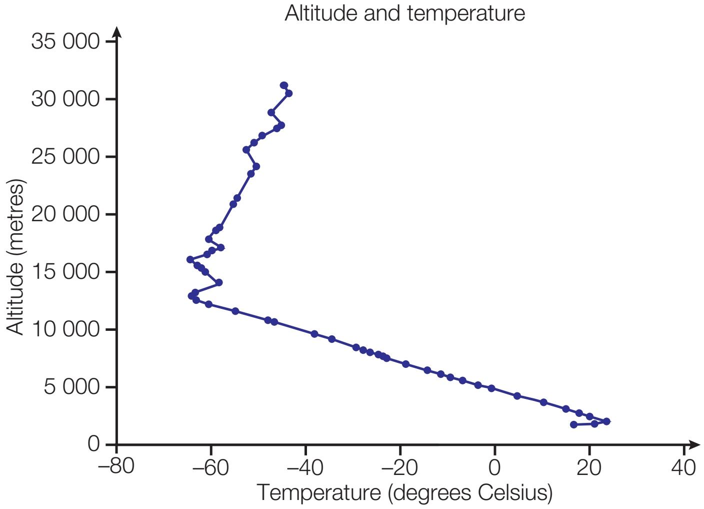
From the GDC, the correlation coefficient is $r = -0.942$ and the equation of the regression line is $y = -133x + 4830$ (3 s.f.).

For a height of 5000 m ($y = 5000$), the equation of the regression line gives $x = \dfrac{5000 - 4830}{-133} = -1.28$, i.e. a temperature of −1.28°C.

For a height of 10 000 m ($y = 10000$), the equation gives $x = \dfrac{10000 - 4830}{-133} = -38.9$, i.e. a temperature of −38.9°C.

However, when the entire data set is plotted (see graph below), you can see that the linear equation we found fits only those data points with height less than 12 000 m. At higher altitudes, this equation would give temperature values that are far from the real measurements.

Correlation does not imply causation. 'Peer review', the system whereby scientific papers are checked by researchers working in other universities or countries, is an important way in which scientists can have their results verified or critiqued and so improve their performance. A search on the internet will produce some interesting examples of correlated variables that are taken to have a causative relationship, but when another scientist tests them the connection is found to be false.



Real data; figure used with permission from Houghton Mifflin Harcourt.

Exercise 12.6

You are expected to use your GDC to answer all the questions in this exercise.

1. The performance of eleven students on two examination papers in History is shown in the table below.

Paper A (x%)	90	71	91	71	65	98	70	86	95	80	75
Paper B (y%)	82	82	86	63	81	95	76	91	98	92	71

(a) Calculate the product moment correlation coefficient, r, of the data.

(b) What does the value of r suggest about the relationship between scores on the two examination papers?

(c) Hayley, a student from another class, has taken only Paper A. She scored 83%. Estimate Hayley's anticipated score on Paper B.

(d) Joe's score on Paper B was 68%. Calculate an estimate of Joe's score on Paper A.

2. The table shows data on the GDP per capita and the unemployment rate (as a percentage of the labour force) of 10 countries.

Country	1	2	3	4	5	6	7	8	9	10
GDP per capita, *x* ($,000)	50.5	48.1	31.4	50.1	44.4	44.6	27.9	48.5	37.0	22.1
Unemployment rate, *y* (%)	4.10	7.89	7.40	7.82	9.52	6.00	16.40	14.30	8.20	6.30

(a) Determine the equation of the regression line of y on x.

(b) Use your equation to estimate the GDP per capita of a country with an unemployment rate of 6.8%.

(c) Estimate the unemployment rate of a country with a per capita GDP of $34,500.

(d) Find the correlation coefficient r.

(e) Comment on the value of r.

3. The total number of points gained by the top ten teams in the English Premier League for the 2009/2010 and 2010/2011 seasons is shown in the following table.

Team	**Points**	
	2009/2010 (*x*)	**2010/2011 (*y*)**
Manchester United	85	80
Chelsea	86	71
Manchester City	67	71
Arsenal	75	68
Tottenham	70	62
Liverpool	63	58
Everton	61	54
Fulham	46	49
Aston Villa	64	48
Sunderland	44	47

Jessica wants to determine whether there is a correlation between the points gained by these ten teams in the two successive seasons. She thinks there is bound to be a strong positive correlation between the points gained by each team in the two seasons.

(a) By calculating the coefficient of correlation, comment on the accuracy of Jessica's assertion.

(b) The equation of the regression line of y on x is $y = mx + c$. Determine the values of m and c.

Hull City gained 30 points in the 2009/2010 season but they were relegated at the end of the season and so did not play in the same division during the next season.

Jessica decided to use the equation from part (b) to estimate the number of points Hull City would have gained if they had stayed in the same division for the 2010/2011 season.

(c) What is Jessica's estimate?

(d) Comment on the reliability of the answer in part (c).

Summary

You should know:

- what bivariate data is and how to plot it on a scatter diagram
- the concept of correlation and the line of best fit
- how to draw a line of best fit by hand, passing though the mean point on the scatter diagram
- how to calculate Pearson's product moment correlation coefficient, *r*, using your GDC
- how to interpret values of *r* in terms of strong or weak, positive, zero or negative correlation
- how to find the regression line of *y* on *x*
- how and when to use the regression line for prediction purposes.

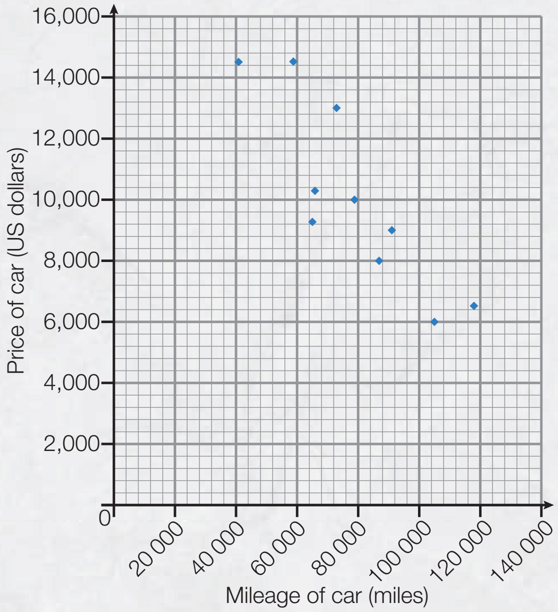
Mixed examination practice

Exam-style questions

1. Bethan is investigating the mileage of second-hand cars and their prices. The results of her investigation on a sample of ten cars are shown in the table.

Mileage	65 000	118 000	66 000	41 000	73 000	105 000	79 000	91 000	87 000	59 000
Price (US dollars)	9,295	6,495	10,295	14,495	12,995	5,995	9,995	8,995	7,995	14,495

The diagram shows the scatter graph of her data.



(a) State the independent and dependent variables.

(b) Comment on the correlation between the variables.

(c) State the mean mileage and the mean price of the cars.

(d) The point on the scatter diagram representing the mean mileage and the mean price of the cars can be represented as $(\bar{m}, \bar{p})$. Copy the scatter diagram onto graph paper and draw the line of best fit, indicating the point $(\bar{m}, \bar{p})$.

2. The table shows the attendance during a school term and the end-of-term test scores of 12 students.

Attendance (%)	76	83	76	91	54	97	48	97	89	90	84	91
Test score (%)	54	50	57	72	39	75	36	59	59	65	59	72

(a) Draw a scatter diagram by hand to illustrate the data. Include the line of best fit on your diagram.

(b) State the type of correlation.

(c) Determine the linear correlation coefficient, r, of the data and hence comment on the strength of the correlation.

3. The following table shows the distances (in metres) thrown by ten athletes who participated in both a javelin and a discus competition.

Javelin, x (m)	54.5	33.7	50.6	61.9	49.3	45.9	38	46.1	52.8	62
Discus, y (m)	44.8	37.12	51.76	62.64	34.08	26.08	38.64	31.68	48.64	62.8

(a) Find the equation of the regression line of y on x.

(b) Find the linear correlation coefficient, r, of the data.

(c) Comment on the value of r.

4. Jensen is investigating the relationship between the cross-sectional area (in mm^2) and the current rating (in amperes) of some copper conductors. His results are illustrated in the table below.

Cross-sectional area, x	1	1.5	2.5	4	6	10	16
Current rating, y	15	19.5	27	36	46	63	85

(a) Write down the correlation coefficient, r.

(b) Describe the nature of the correlation between cross-sectional area and current rating of these copper conductors.

(c) Write down the equation of the regression line of current rating (y) on cross-sectional area (x) of the copper conductors.

(d) Use your equation from part c) to predict the current rating of a copper conductor with a cross-sectional area of 3.5 mm^2.

(e) Comment on the suitability of using the regression line equation obtained in part (c) to predict the current rating of a copper conductor with cross-sectional area 18.5 mm^2.

5. The weather section of a newspaper lists the minimum and maximum daytime temperatures (in degrees Celsius) of 11 cities:

Minimum, x	1	6	3	7	6	0	0	4	3	2	8
Maximum, y	12	7	10	9	11	15	15	11	9	15	10

(a) Write down the correlation coefficient, r, of the data.

(b) The equation of the regression line of y on x is of the form $y = mx + c$. Determine the values of m and c.

(c) Comment on the suitability of using the regression line equation from part (b) to predict the maximum daytime temperature of a city from the same region with a minimum temperature of 15°C.

(d) Another city in the same region had a minimum temperature of 5°C on the same day. Estimate the maximum temperature.

6. The following table shows 52-week highest and lowest prices (in sterling pence) of the shares of nine companies listed in the FTSE 100 on 13 March 2012.

Highest, x	1228	1754	2347	1207	3344	1491	645	420	3194
Lowest, y	940	787	1412	740	2138	900	464	301	2543

Mr Lawrence believes that there is a strong positive correlation between the highest and lowest share prices.

(a) By calculating Pearson's product moment correlation coefficient for the data, comment on the validity of Mr Lawrence's assertion.

(b) Determine the equation of the regression line of y on x.

(c) One of the listed companies in the FTSE 100 had a highest share price of 1564 pence. Estimate its lowest share price over the 52-week period.

(d) Work out an estimate of the maximum share price of another company with a lowest share price of 2300 pence over the 52-week period.

7. The following table shows the finishing times (in seconds) of some athletes who competed in both the 100 m and the 200 m races in a track and field event.

100 m time, x	13.3	14.1	14.7	16.4	14.2	15.8	16.3	15.3	15.4	15.7
200 m time, y	25.9	25.4	25.6	29.8	24.9	27.1	31.0	29.2	30.1	29.4

(a) Find the correlation coefficient r.

(b) Comment on the value of r.

(c) Determine the equation of the regression line of y on x.

Serginio and Tunde participated in the same event.

Serginio ran the 100 m race in 15.12 seconds, but he did not compete in the 200 m race due to an injury.

Tunde was disqualified from the 100 m race due to a false start but finished the 200 m race in 26.1 seconds.

(d) Use your equation from part (c) to estimate:

(i) Serginio's finishing time if he had run the 200 m race

(ii) Tunde's finishing time if he had run the 100 m race.

Past paper questions

1. In an experiment a vertical spring was fixed at its upper end. It was stretched by hanging different weights on its lower end. The length of the spring was then measured. The following readings were obtained.

Load x (kg)	0	1	2	3	4	5	6	7	8
Length y (cm)	23.5	25	26.5	27	28.5	31.5	34.5	36	37.5

(a) Plot these pairs of values on a scatter diagram taking 1 cm to represent 1 kg on the horizontal axis and 1 cm to represent 2 cm on the vertical axis. *[4 marks]*

(b) (i) Write down the mean value of the load ($\bar{x}$). *[1 mark]*

(ii) Write down the standard deviation of the load. *[1 mark]*

(iii) Write down the mean value of the length ($\bar{y}$). *[1 mark]*

(iv) Write down the standard deviation of the length. *[1 mark]*

(c) Plot the mean point $(\bar{x}, \bar{y})$ on the scatter diagram. Name it L. *[1 mark]*

It is given that the covariance S_{xy} is 12.17.

(d) (i) Write down the correlation coefficient, *r*, for these readings. *[1 mark]*

(ii) Comment on this result. *[2 marks]*

(e) Find the equation of the regression line of *y* on *x*. *[2 marks]*

(f) Draw the line of regression on the scatter diagram. *[2 marks]*

(g) (i) Using your diagram or otherwise, estimate the length of the spring when a load of 5.4 kg is applied. *[1 mark]*

(ii) Malcolm uses the equation to claim that a weight of 30 kg would result in a length of 62.8 cm. Comment on his claim. *[1 mark]*

[Total 18 marks]

[Nov 2006, Paper 2, TZ0, Question 3] (© *IB Organization 2006*)

2. Tania wishes to see whether there is any correlation between a person's age and the number of objects on a tray which could be remembered after looking at them for a certain time.

She obtains the following table of results.

Age (*x* years)	15	21	36	40	44	55
Number of objects remembered (*y*)	17	20	15	16	17	12

(a) Use your graphic display calculator to find the equation of the regression line of *y* on *x*. *[2 marks]*

(b) Use your equation to estimate the number of objects remembered by a person aged 28 years. *[1 mark]*

(c) Use your graphic display calculator to find the correlation coefficient *r*. *[1 mark]*

(d) Comment on your value for *r*. *[2 marks]*

[Total 6 marks]

[Nov 2007, Paper 1, TZ0, Question 14] (© *IB Organization 2007*)

Chapter 13 Chi-squared hypothesis testing

In this chapter you will learn:

- when you can use chi-squared tests
- how to find expected frequencies
- how to calculate the chi-squared statistic
- what degrees of freedom are and how to work out the number of degrees of freedom in a given situation
- to formulate null and alternative hypotheses
- about *p*-values, what they mean and how to use them.

As you learned in Chapter 12, Pearson's product moment correlation coefficient can be used to test the strength of connection between two variables in a set of bivariate data.

There are many other situations in which you might want to test if two or more variables are dependent on each other. For example, in a trial of a new fertiliser the manufacturer wants to find out whether the fertiliser actually has an effect on the growth of crops; in other words, is the yield of a plot where the new chemical has been applied significantly higher than the yield of the control plot?

Karl Pearson developed a statistical method called the chi-squared (χ^2) test for independence, which can be used to determine if a result is significant. This test could be used to determine if the yield in the treated plot is significantly greater than that in the control plot and thus indicate if the new chemical works.

To take another example, suppose you want to know whether the choice of a second language for the IB diploma is independent of gender. Assuming this were the case, you would expect the proportions of boys and girls choosing each language to be the same. You could test your assumptions using the chi-squared (χ^2) test. You would collect the real data on how many boys and girls choose each language option and compare these numbers with the 'expected' numbers. Then you would determine if the real data deviates enough from the expected to say that students' language choices might depend on gender.

13.1 The χ^2 statistic

The χ^2 statistic is a goodness-of-fit test. In general, the **χ^2 statistic** is defined as:

$$\chi^2 = \sum \frac{(f_O - f_E)^2}{f_E}$$

where f_O is an observed frequency, f_E is the corresponding expected frequency, and the Σ symbol indicates 'add up all the separate calculations'.

At its most basic level, the χ^2 test is used to check if the frequencies of collected data values (observed frequencies) differ significantly from what you would 'expect' to find (expected frequencies). In other words, it can be used to assess how good the 'fit' is between the observed frequency distribution and a 'theoretical' distribution. The theoretical or expected distribution could be based on known probabilities, the normal distribution, complete randomness, or some other distribution that you suspect is appropriate.

The χ^2 statistic is not difficult to calculate and the calculations can be done quickly on your GDC. (See '*13.1 The χ^2 test for independence*' on page 675 of the GDC chapter for a reminder if you need to.)

Learning links

13A Calculating the χ^2 statistic

To gain an understanding of the method and what the results mean, it is best to work through the following examples by hand.

Jon needs some random numbers for his project, but has left his calculator at school. He writes down 100 single figures chosen from the numbers 0 to 9. Are the figures he has written down really random? How good is the fit between the numbers he has chosen 'at random' and numbers that are truly randomly generated?

Suppose these are Jon's results:

Number	0	1	2	3	4	5	6	7	8	9
Frequency	12	7	9	14	6	7	14	13	8	10

How close is this distribution of frequencies to the 'theoretical' one?

We will use the formula $\chi^2 = \sum \frac{(f_O - f_E)^2}{f_E}$ where f_O is an observed frequency and f_E is the corresponding expected frequency.

First, we need to work out the 'expected frequencies' assuming that the numbers are truly randomly selected. Using probability theory, under the assumption of complete randomness each of the ten numbers from 0 to 9 is equally likely to be chosen, so each number should come up $\frac{1}{10} \times 100 = 10$ times.

We then calculate the discrepancy between observed and expected frequencies as follows:

Number	0	1	2	3	4	5	6	7	8	9
Observed frequency f_O	12	7	9	14	6	7	14	13	8	10
Expected frequency f_E	10	10	10	10	10	10	10	10	10	10
Difference $f_O - f_E$	+2	–3	–1	+4	–4	–3	+4	+3	–2	0

You might think that adding up the differences will give a good representation of the discrepancy between observed and expected frequencies; however, in this case the differences add up to zero because the positive and negative values cancel each other out. To get around this cancellation problem, we square the differences, making them all non-negative, before adding them up:

→

continued . . .

Number	0	1	2	3	4	5	6	7	8	9
Observed frequency f_O	12	7	9	14	6	7	14	13	8	10
Expected frequency f_E	10	10	10	10	10	10	10	10	10	10
Difference $f_O - f_E$	+2	–3	–1	+4	–4	–3	+4	+3	–2	0
Squared difference $(f_O - f_E)^2$	4	9	1	16	16	9	16	9	4	0

We find that the sum of the squared differences, $\sum (f_O - f_E)^2$, is 84. To obtain a measure of discrepancy that is of roughly the same magnitude as the individual frequencies, we divide this sum by the expected frequencies, 10, to get 8.4.

So, in Jon's random number case, $\chi^2 = 8.4$.

The critical χ^2 value

But what does the value of the χ^2 statistic mean?

Let's use the example from Learning links 13A: Jon needs some random numbers for his project; he writes down 100 single figures chosen from the numbers 0–9.

These are Jon's results:

Number	0	1	2	3	4	5	6	7	8	9
Frequency	12	7	9	14	6	7	14	13	8	10

If all 100 figures were truly chosen at random, you would expect each number from 0–9 to have a probability of 0.1 (there are 10 numbers, so the probability of selecting any one number is 1 out of 10). In Learning links 13A we compared the frequency of Jon's numbers with the expected frequencies and calculated that $\chi^2 = 8.4$. What does this value of 8.4 mean? 8.4 is a measure of the discrepancy between the real frequencies and the expected frequencies. Is this measure of discrepancy 'significant' enough for us to say that Jon's numbers are not really random and that he was actually biased in favour of certain numbers and against others?

To decide if the χ^2 statistic is large enough to conclude that the discrepancy between the observed frequencies and the expected/theoretical frequencies is significant, we compare it with a **critical** χ^2 value. This critical value depends on the number of **degrees of freedom** and the **significance level** at which you choose to work.

You can interpret the results as follows:

- if calculated χ^2 statistic > critical value, the result is significant and the expected/theoretical distribution **is not a good fit** of the observed data

- if calculated χ^2 statistic < critical value, the result is not significant and the expected/theoretical distribution **is a good fit** of the observed data.

The significance level is usually denoted by the Greek letter α (alpha). It is the greatest acceptable **probability of incorrectly** deciding that the result is significant when actually it is insignificant, i.e. thinking the theoretical distribution is not a good fit for the observed data when it actually is. So, if you choose a significance level of 5%, then there will be at most a 5% chance that you will state your data is significant when actually it is not. The lower α is, the more stringent the criterion will be for a 'significant' discrepancy, and the more confident you can be that your decision is correct.

Tables are available that give χ^2 critical values for various degrees of freedom and significance levels, and you read off your critical χ^2 from this table. There is a small table at the end of this chapter, on page 404.

FF We will see how to calculate the number of degrees of freedom later in the chapter.

In Jon's random number example, the number of degrees of freedom is 9. He can look up the χ^2 critical values for 9 degrees of freedom in a set of tables. The critical values for significance levels 10%, 5% and 1% ($\alpha = 0.1$, 0.05 and 0.01) are as follows:

$$\chi^2_{10\%} = 14.68 \quad \chi^2_{5\%} = 16.92 \quad \chi^2_{1\%} = 21.67$$

At the 5% significance level, Jon's $\chi^2_{\text{calc}} = 8.4$ and $\chi^2_{5\%} = 16.92$.

Since $8.4 < 16.92$, the result is not significant, which means that there is **not** enough evidence to claim that Jon's numbers are not random. Therefore, the χ^2 test has informed us that Jon's chosen numbers are likely to be random.

exam tip

In an examination you will always be given the critical value of χ^2 that you need.

p-value

As well as comparing the calculated χ^2 statistic with a critical value, you can also find the 'p-value' associated with the χ^2 statistic and the number of degrees of freedom. The ***p*-value** is the **probability** of getting a discrepancy as large as the calculated χ^2 statistic **if** the theoretical distribution were correct. Therefore, the smaller the p-value, the less likely it is that the observed frequencies and the theoretical distribution are a good fit and the more likely it is that the discrepancy between the observed and theoretical frequencies is significant.

You can interpret the results as follows:

- when the p-value < significance level, the result is significant and the theoretical distribution is **not a good fit** of the observed distribution
- when the p-value > significance level, the result is not significant and the theoretical distribution **is a good fit** of the observed distribution.

For Jon's χ^2_{calc} value of 8.4 and 9 degrees of freedom, $p = 0.494$. (You can find this value using a GDC or a spreadsheet program on a computer; see '*13.1 The χ^2 test for independence*' on page 675 of the GDC chapter for a reminder if you need to.) Since $0.494 > 0.05$, at the 5% significance level there is **not** enough evidence to claim that Jon's numbers are not random. (This agrees with our earlier conclusion obtained by comparing with the critical value.)

In summary

For a goodness-of-fit test:

- If $\chi^2_{\text{calc}} > \chi^2_{\text{critical}}$, then you can say that the result is significant and the theoretical distribution is not a good fit for the data; if $\chi^2_{\text{calc}} < \chi^2_{\text{critical}}$, then you can say that there is not enough evidence to suggest the result is significant, and the theoretical distribution is a good fit for the data.
- If p-value < significance level, then you can say that the result is significant and the theoretical distribution is not a good fit for the data; if p-value > significance level, then you can say that there is not enough evidence to suggest the result is significant, and the theoretical distribution is a good fit for the data.

Worked example 13.1

Q. Finn and Edda are playing a game with an eight-sided die. Edda wins if she rolls an odd number; Finn wins if he rolls an even number. Edda keeps losing the game, and is sure that the dice is biased towards even numbers. From the results below, is there enough evidence to support Edda's claim?

Number	1	2	3	4	5	6	7	8
Frequency	4	5	6	9	6	9	0	3

This is a goodness-of-fit test where the 'theoretical distribution' is based on assuming that the die is balanced. First, calculate the expected frequencies.

A. If the die is balanced, each number should appear with equal probability $\frac{1}{8}$.

The total frequency is $4 + 5 + 6 + 9 + 6 + 9 + 0 + 3 = 42$,

so each number is expected to occur $42 \times \frac{1}{8} = 5.25$ times.

Calculate the squared differences between the observed and expected frequencies.

Number	1	2	3	4	5	6	7	8
f_o	4	5	6	9	6	9	0	3
f_E	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25
$f_o - f_E$	−1.25	−0.25	0.75	3.75	0.75	3.75	−5.25	−2.25
$(f_o - f_E)^2$	1.5625	0.0625	0.5625	14.0625	0.5625	14.0625	27.5625	5.0625

→

continued . . .

Use the formula to find the χ^2 statistic.

$$\sum (f_O - f_E)^2 = 63.5$$

$$\chi^2_{calc} = \sum \frac{(f_O - f_E)^2}{f_E} = \frac{63.5}{5.25} = 12.1$$

Look up the critical value from a table. The number of degrees of freedom is 8 – 1 = 7.

For 7 degrees of freedom, $\chi^2_{5\%} = 14.1$.

Since $12.1 < 14.1$, at the 5% significance level there is not enough evidence to say that the die is unbalanced; maybe Edda is just having an unlucky day.

Let's try a different significance level.

For 7 degrees of freedom, $\chi^2_{10\%} = 12.017$.

Since $12.1 > 12.017$, at the 10% significance level we can conclude that the die is unbalanced.

Check the *p*-value as well.

For $\chi^2_{calc} = 12.1$ and 7 degrees of freedom, $p = 0.0975$.

As $0.0975 > 0.05$, the evidence for bias is not strong enough at the 5% significance level.

But, as $0.0975 < 0.1$, the evidence is strong enough at the 10% significance level.

Degrees of freedom

In the example where Jon was writing down random numbers chosen from 0–9, the table had 10 'cells' containing the observed frequencies. Since he wrote down a total of 100 numbers, the sum of the frequencies is 100. For the first 9 cells to be filled, any frequency values could have gone into them; but once they have been filled, the value of the final cell is completely determined as it must contain a frequency that would give a total frequency of 100 numbers. Therefore, the final cell has 'no freedom' so the number of degrees of freedom in this situation is 9 (10 cells – 1 cell).

In general, for a goodness-of-fit test where you are comparing a list of observed frequencies against a theoretical distribution:

degrees of freedom = number of frequencies to be compared – 1

Now let's consider more complicated situations.

A car dealer has collected data about the age of his customers and the colour of car they choose. Is colour choice independent of age, or is there any relationship between the two variables? His data, collected from 63 customers, can be presented in a **two-way contingency table** as follows:

		Age group				Total
		20–30	31–40	41–50	50+	
Colour	Green	2	3	6		14
	Red	7	4	4		18
	Grey	2	4	8		16
	Blue					15
Total		13	15	20	12	63

You learned about two-way tables in Chapter 10, section 10.2.

The cells containing the observed frequencies are arranged in four columns and four rows; each column represents a different age group, and each row represents a different choice of car colour. Note that if the totals of every row and column are known, then once the cells in the first three rows and first three columns have been filled, the values in the remaining (blank) cells are completely determined, because they need to give the correct total for each row and column. Thus there is no freedom in choosing the values to go into the cells of the fourth row and fourth column. The number of degrees of freedom in this case is the number of cells in the first three rows and first three columns: $3 \times 3 = 9$.

The general formula for calculating the degrees of freedom for a two-way contingency table, provided that the total frequency of each row and each column is known, is similar to the formula for a single list of observed frequencies:

$$\text{degrees of freedom} = (\text{number of rows} - 1) \times (\text{number of columns} - 1)$$

hint

Don't confuse contingency tables with the tables of critical values for the χ^2 test.

In the next section you will see how to use the χ^2 test for independence on a two-way table.

Exercise 13.1

1. For each of the following (two-way) contingency tables, calculate the number of degrees of freedom.

(a)

	B_1	B_2
A_1	64	134
A_2	91	174

(b)

	B_1	B_2	B_3
A_1	48	24	14
A_2	34	16	10

(c)

	B_1	B_2	B_3	B_4
A_1	45	95	63	28
A_2	20	109	57	18
A_3	45	50	104	29

(d)

	B_1	B_2
A_1	24	35
A_2	37	52
A_3	57	41
A_4	79	58

13.2 The χ^2 test for independence

In Nariko's town, people say that it is better if you get Ms B as your examiner for the driving test; she is believed to pass more people than Miss C or Mr A.

Nariko decides to collect some actual results from the test centre, and analyse them to determine whether there is any truth to the common belief.

To check if pass/fail rates in the driving test are dependent on the examiner, Nariko uses a χ^2 test. First, she puts the data she collected in a two-way table:

Observed frequencies	Mr A	Ms B	Miss C	Total
Pass	28	38	35	101
Fail	20	10	18	48
Total	48	48	53	149

At first glance, Mr A and Ms B conducted the same number of tests (48), but Ms B passed 28 more people than she failed whereas Mr A passed only 8 more people than he failed. This seems to suggest that the rumours in town are true. However, Nariko is a statistician and knows that it is important to avoid being swayed by opinion; so she starts by assuming that there is no dependence or bias.

All χ^2 tests start from the assumption that the two factors being tested are **independent**. This independence is stated as the **null hypothesis**, usually denoted by H_0.

exam tip

In examinations, all your calculations for the χ^2 test may be done on the GDC. If you use the χ^2 test in your project, you can use either the formula or your GDC.

Nariko's null hypothesis is:

H_0: the pass rate is independent of the examiner who conducts the test.

In contrast, the **alternative hypothesis** (H_1) is:

H_1: the pass rate is dependent on the examiner who conducts the test.

To apply the χ^2 test, Nariko then needs to calculate the **expected frequencies**, which are the frequencies that should occur if the factors were truly independent. These expected frequencies are calculated using probability theory (see Chapter 10).

From the rightmost column of Nariko's table above, you can see that the pass rate (proportion of passes) is $\frac{101}{149}$ and the failure rate is $\frac{48}{149}$.

Therefore, if these rates are **independent** of which examiner conducts the driving test, then all three examiners would be expected to pass (and fail) the same proportion of candidates.

Since Mr A examined 48 candidates, $48 \times \frac{101}{149} = 32.5$ of them would be expected to pass, and $48 \times \frac{48}{149} = 15.5$ would be expected to fail.

Ms B examined the same number of people as Mr A, so her expected frequencies of pass and fail are the same as for Mr A.

Miss C examined 53 candidates, so $53 \times \frac{101}{149} = 35.9$ would be expected to pass, and $53 \times \frac{48}{149} = 17.1$ would be expected to fail.

Expected frequencies	Mr A	Ms B	Miss C
Pass	32.5	32.5	35.9
Fail	15.5	15.5	17.1

The table of expected frequencies is shown on the left.

In general, the expected frequency to go in each cell of the table can be found from the following formula:

$$\text{expected frequency} = \frac{\text{row total} \times \text{column total}}{\text{total}}$$

For instance, the cell for 'Mr A, Pass' has row total 101 and column total 48, so the expected frequency is $\frac{101 \times 48}{149} = 32.5$ (3 s.f.). The cell for 'Miss C, Fail' has row total 48 and column total 53, hence the expected frequency is $\frac{48 \times 53}{149} = 17.1$.

Now Nariko has two tables, one showing the observed frequencies from the data she collected, and the other showing the expected values that she calculated using probability theory.

Observed frequencies	Mr A	Ms B	Miss C
Pass	28	38	35
Fail	20	10	18

Expected frequencies	Mr A	Ms B	Miss C
Pass	32.5	32.5	35.9
Fail	15.5	15.5	17.1

Each table contains two rows and three columns of numbers, so the number of degrees of freedom $= (2-1) \times (3-1) = 1 \times 2 = 2$.

Using her GDC, Nariko finds that $\chi^2_{\text{calc}} = 4.89$.

For 2 degrees of freedom, the tables give the critical value at the 0.05 significance level as $\chi^2_{5\%} = 5.991$. Since $\chi^2_{\text{calc}} < \chi^2_{5\%}$, Nariko concludes that she does **not** have enough evidence to reject H_0. In other words, the data does not support the claim that the pass rate is dependent on the examiner.

To confirm the result obtained by comparing χ^2_{calc} and χ^2_{critical}, Nariko also finds the p-value using her GDC: it is $p = 0.087$. Because $0.087 > 0.05$, this confirms that she should accept the null hypothesis of pass rates and examiners being independent.

hint

If χ^2_{calc} had been large enough that Nariko could reject the null hypothesis, then she would be accepting the alternative hypothesis, i.e. that there is a relationship between the examiner and the pass rate on the driving test.

Remember that even when the chi-squared test indicates that there is dependence between two factors, it does not tell you what kind of relationship it is. So even if Nariko's data did show enough evidence of dependence, the χ^2 test would not tell Nariko that Ms B was more likely to pass people than the other examiners — just that there was some relationship between the examiner and the likelihood of passing.

Significance level

You have already seen in Worked example 13.1 that the significance level chosen can make a difference to your conclusion.

For Nariko's driving test data, the p-value of 0.087 was greater than 0.05, so at the 5% significance level Nariko accepts the null hypothesis. However, since $0.087 < 0.1$, at the 10% significance level Nariko would **reject** the null hypothesis and conclude that there is dependence between pass rate and examiner. If you look up the critical χ^2 value for $\alpha = 0.1$ and 2 degrees of freedom, you will find that $\chi^2_{10\%} = 4.605$. As $\chi^2_{\text{calc}} = 4.89 > 4.065$, this also tells us that we can reject the null hypothesis.

The most commonly chosen significance level is 5%, but 1%, 2.5% and 10% are also used.

Knowing how it can drastically affect the conclusions of the test, make sure that you always state the significance level (or check that you are using the right level given in the question) when you are analysing data with the chi-squared test.

exam tip

5% is the significance level most commonly used in IB examination questions.

Exercise 13.2

1. Calculate the expected frequencies for each of the following contingency tables.

(a)

	B_1	B_2
A_1	64	134
A_2	91	174

(b)

	B_1	B_2	B_3
A_1	48	24	14
A_2	34	16	10

(c)

	B_1	B_2	B_3	B_4
A_1	45	95	63	28
A_2	20	109	57	18
A_3	45	50	104	29

(d)

	B_1	B_2
A_1	24	35
A_2	37	52
A_3	57	41

In applying the χ^2 test, you use probability theory to do calculations. Can this be considered a rigorous application of mathematics to science? In statistics, is it ever possible to give a definite answer, or is it enough that you always remain aware of the restrictions, biases or flaws that may exist? When you read the results of a statistical survey, do you keep this in mind and take it into account while you interpret the conclusions?

13.3 Using your GDC to calculate the χ^2 statistic

The questions in Exercises 13.1 and 13.2 can be worked through without using any statistical functions on your GDC. In this section, we show you how to find the χ^2 statistic and the p-value with your GDC. See '*13.1 The χ^2 test for independence*' on page 675 of the GDC chapter for details. As you will see, it is very convenient to use your GDC when you need to perform a χ^2 test for a project or assignment or to answer an examination question.

Suppose you want to find out whether there is a link between hair colour and eye colour. Do most people with black hair have brown eyes?

The following data is collected from a sample of 63 people.

	Black hair	**Brown hair**	**Blonde hair**
Blue eyes	5	7	12
Brown eyes	15	10	2
Green eyes	3	4	5

1. State the hypotheses. Remember that the null hypothesis, H_0, always assumes **independence** of the factors.

   H_0: the colour of a person's eyes is independent of the colour of their hair.
   H_1: the colour of a person's eyes is dependent on the colour of their hair.

2. Decide on the significance level that you are going to use.

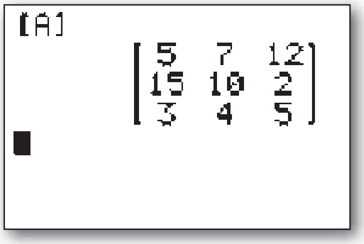
   The data will be tested at the 5% significance level.

3. Enter the table of observed values into matrix A on your GDC. The table of data has 3 rows and 3 columns, so you need a matrix with the same dimensions.

exam tip

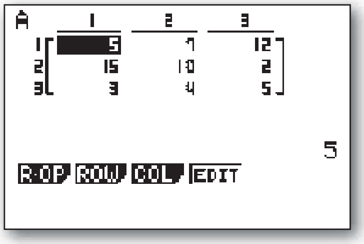
Matrices (plural of 'matrix') are not in the IB Mathematical Studies syllabus. You only need them for entering data values from two-way tables into your GDC.

TEXAS



```
[A]
        [5   7  12]
        [15 10  2 ]
        [3   4  5 ]
```

CASIO



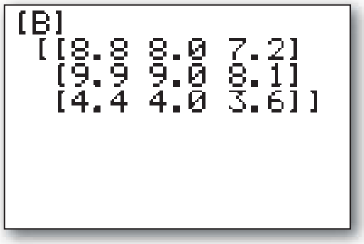
```
A      1      2      3
 1[    5      7     12]
 2[   15     10      2]
 3[    3      4      5]

                        5
R·OP ROW COL EDIT
```

Take a look at matrix B, which contains the expected frequencies. Note that the GDC sets this up automatically; you don't need to do any calculations.

GDC

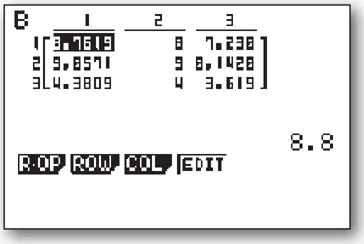
See section '*13.2 viewing the contents of a matrix*' on page 677 of the GDC chapter if you need to.

TEXAS



```
[B]
 [[8.8 8.0 7.2]
  [9.9 9.0 8.1]
  [4.4 4.0 3.6]]
```

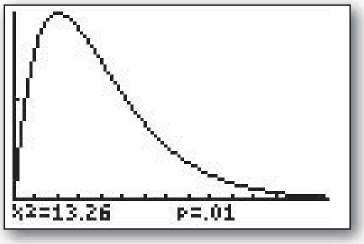
CASIO

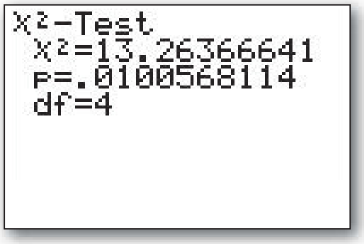


```
B      1       2      3
 1[8.7619      8  7.238]
 2[9.8571      9 8.1428]
 3[4.3809      4  3.619]

                      8.8
R·OP ROW COL EDIT
```

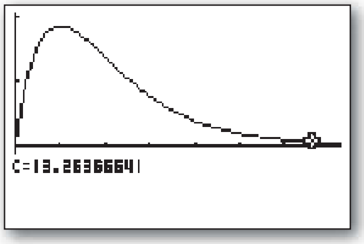
4. Ask your GDC to calculate the χ^2 statistic.

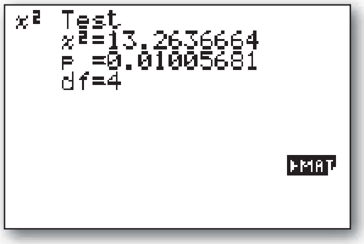
TEXAS


χ²=13.26 p=.01


χ²-Test
χ²=13.26366641
p=.0100568114
df=4

CASIO


C=13.2636664


χ² Test
χ²=13.2636664
p =0.01005681
df=4

Your GDC will give you χ^2_{calc}, the *p*-value and the degrees of freedom 'df'.

5. Compare χ^2_{calc} and $\chi^2_{5\%}$. Since df = 4, you can look up the critical value $\chi^2_{5\%}$ from a table (or, if this is an examination question, $\chi^2_{5\%}$ will be given to you). In this case:

   $\chi^2_{5\%} = 9.49$

   $\chi^2_{\text{calc}} = 13.3$

   $\chi^2_{\text{calc}} > \chi^2_{5\%}$

   So reject H_0 and accept H_1: the colour of someone's eyes is dependent on the colour of their hair.

6. As an alternative to step 5, or to confirm the conclusion drawn there, compare the *p*-value with the significance level.

   Here $p = 0.01 < 0.05$, so reject H_0 and conclude that eye colour is dependent on hair colour.

Remember that the χ^2 statistic is a measure of the **deviation** between the observed and expected frequencies. Therefore:

- If the observed and expected values are all exactly the same, then $\chi^2_{\text{calc}} = 0$.
- If the observed and expected values are close together, the value of χ^2_{calc} will be small. If χ^2_{calc} is **smaller** than the value of χ^2_{critical} for the associated degrees of freedom and significance level, then **accept** H_0.
- If the observed and expected values are very different, the value of χ^2_{calc} will be large. If χ^2_{calc} is **larger** than the value of χ^2_{critical}, then **reject** H_0.

The *p*-value is a measure of how **likely** you are to get your value of χ^2_{calc} if H_0 were true. Therefore:

- If the *p*-value is **more** than the significance level, you don't have enough evidence to doubt the validity of H_0, so **accept** H_0.
- If the *p*-value is **less** than the significance level, then **reject** H_0.

The χ^2 test is a widely used technique in the social sciences as well as in biology, psychology, geography and many other fields. It can be used to answer questions such as: Do men and women tend to vote for different political parties? How effective is a new vaccine? Is there a connection between the colour of cars and the number of accidents in which they are involved?

exam tip

In an examination, you will expected to do the calculations for a χ^2 test on your GDC, so the questions will test your **understanding** of the results that your calculator produces.

Worked example 13.2

Q. Katya has collected data about the types of movie that students in her year group particularly enjoy. She predicts that the preferences will be the same for boys and girls.
Here are her results:

	Adventure	Romance	Comedy	Animation
Male	11	3	9	8
Female	6	9	7	7

(a) State Katya's null hypothesis and her alternative hypothesis.

(b) Find the expected frequency for the number of females who prefer adventure movies.

(c) Using your GDC, find the chi-squared statistic for Katya's data.

(d) Using your GDC, find the p-value for this data.

(e) Show that the number of degrees of freedom for this data is 3.

(f) If $\chi^2_{5\%} = 7.815$, give Katya's conclusion.

Remember that the null hypothesis always assumes independence.

A. (a) H_0: the choice of favourite movie type is independent of gender.

H_1: the choice of favourite movie type depends on gender.

Look at the cell for 'Female, Adventure' and calculate its row total and column total. Also find the total number of students surveyed, which is the sum of the row totals or the sum of the column totals (these should be the same!).

(b) Row total $= 6 + 9 + 7 + 7 = 29$

Column total $= 11 + 6 = 17$

Total frequency $= 60$

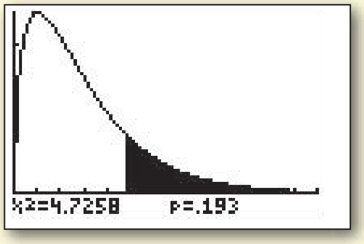
so expected frequency $= \dfrac{29 \times 17}{60} = 8.22$

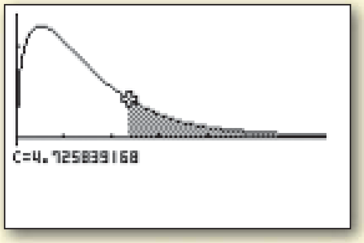
Use a matrix with 2 rows and 4 columns to enter the data into your GDC. Then calculate the χ^2 statistic. See section '*13.1 The χ^2 test for independence*' on page 675 if you need a reminder.

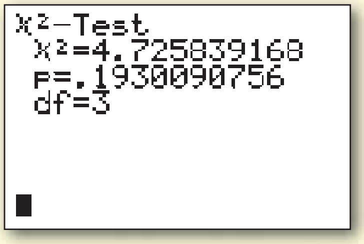
(c)

TEXAS

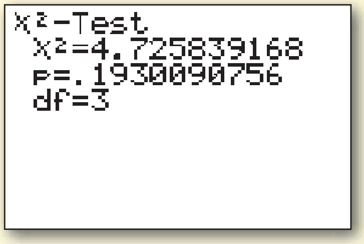
CASIO







```
χ²-Test
 χ²=4.725839168
 p=.1930090756
 df=3
```



```
χ²-Test
 χ²=4.725839168
 p=.1930090756
 df=3
```

$\chi^2_{calc} = 4.73$

continued . . .

Read off the p-value from the screen.

(d) From GDC: $p = 0.193$

Use the formula for a two-way table.

(e) $\text{df} = (4 - 1) \times (2 - 1) = 3 \times 1 = 3$

Compare χ^2_{calc} with the given critical value $\chi^2_{5\%}$. Or compare the p-value with the significance level 5%.

(f) $4.73 < 7.815$, i.e. $\chi^2_{calc} < \chi^2_{5\%}$

so Katya accepts the null hypothesis.

$p = 0.193 > 0.05$ leads to the same conclusion.

Therefore Katya concludes that the choice of favourite movie type is independent of gender.

Exercise 13.3

1. For each of the following contingency tables, calculate:

(i) the degrees of freedom (ii) the χ^2 statistic (iii) the p-value.

(a)

	B_1	B_2
A_1	64	134
A_2	91	174

(b)

	B_1	B_2	B_3
A_1	48	24	14
A_2	34	16	10

(c)

	B_1	B_2	B_3	B_4
A_1	45	95	63	28
A_2	20	109	57	18
A_3	45	50	104	29

(d)

	B_1	B_2
A_1	24	35
A_2	37	52
A_3	57	41
A_4	79	58

2. Calculate the chi-squared statistic for each of the following sets of data.

(a)

	B_1	B_2	B_3
A_1	259	280	388
A_2	130	219	210

(b)

	P	Q	R
A	86	153	52
B	79	73	35
C	33	21	28

(c)

	P	Q	R
A	23	25	30
B	33	51	46
C	61	73	30
D	86	104	46

(d)

	M	N
P	50	25
Q	80	38
R	45	37
S	61	40

3. Jocelyn wants to investigate whether high performance (at grade A* or A) in GCSE is dependent on gender. She has collected the following data to help her with her investigation.

	A*	A
Boys	77	161
Girls	109	209

She decides to set up a chi-squared test.

(a) State the null hypothesis.

(b) Find the number of degrees of freedom.

(c) Determine the chi-squared statistic for this data.

The χ^2 critical value at 5% level of significance is 3.84.

(d) State whether or not Jocelyn should reject the null hypothesis, justifying your answer clearly.

4. The table below shows data on a sample of drivers involved in motor vehicle accidents, categorised according to gender and age. Karim wants to test whether there is an association between the age and gender of drivers involved in accidents.

	Younger drivers	Older drivers
Male	217	508
Female	115	360

A chi-squared test at the 10% level of significance is performed.

(a) State the null hypothesis.

(b) State the number of degrees of freedom.

(c) Determine the chi-squared statistic for this data.

The χ^2 critical value at 10% level of significance is 4.61.

(d) What conclusion can Karim draw from this test? Give a reason for your answer.

5. A librarian carried out a survey on the genre of books borrowed from the library by readers of different ages. The results are summarised below.

	Fiction	Non-fiction	Other
Under 21	32	18	50
22–40	61	19	36
Over 40	23	37	44

Set up a suitable null hypothesis and test it at the 5% level of significance. You may take the χ^2 critical value for this test to be 9.49.

13.4 Restrictions on using the χ^2 test

When you are analysing data you have collected yourself or data that is given to you, there are some issues to bear in mind before applying the χ^2 test for independence.

- The categories for each variable must be mutually exclusive, i.e. they cannot occur at the same time.
- The **expected** frequencies should not be too small:
  - In a 2 × 2 contingency table, no cell should contain an expected frequency of 5 or less.
  - In a larger contingency table (one with more than two rows or two columns), no cell should contain an expected frequency of 1 or less, and no more than 20% of the expected frequencies should be less than 5.
  - If there are expected frequencies in a table that do not satisfy the above conditions, combine cells so that the required limits are reached.

RR *You learned about mutually exclusive events in Chapter 10.*

exam tip

The Yates continuity correction is outside the syllabus for this course. In examinations, questions will always be set so that the number of degrees of freedom is greater than 1, so you will not be expected to use the Yates correction in the examinations or your project. However, you can always explore it if you are interested.

- If the number of degrees of freedom is 1, the usual formula for the χ^2 statistic is believed to overestimate the amount of deviation between observed and expected values. This overestimation is reduced by applying the **Yates continuity correction**.

Summary

You should know:

- what the χ^2 test for independence is and how to calculate it
- what the critical χ^2 value and p-value are and how to use them
- how to use contingency tables to analyse dependency between two variables
- how to formulate a null and an alternative hypothesis
- how to calculate expected frequencies and degrees of freedom
- the meaning of the significance level
- how to determine if you should reject or accept the null hypothesis

Below is an example of a table of chi-squared values. In the examination you will always be given the values you need to use, and will not have to read them off a table like this.

Table of chi-squared critical values

		Significance level				
		0.10	**0.05**	**0.025**	**0.01**	**0.001**
Degrees of freedom	**1**	2.706	3.841	5.024	6.635	10.828
	2	4.605	5.991	7.378	9.210	13.816
	3	6.251	7.815	9.348	11.345	16.266
	4	7.779	9.488	11.143	13.277	18.467
	5	9.236	11.070	12.833	15.086	20.515
	6	10.645	12.592	14.449	16.812	22.458
	7	12.017	14.067	16.013	18.475	24.322
	8	13.362	15.507	17.535	20.090	26.125
	9	14.684	16.919	19.023	21.666	27.877
	10	15.987	18.307	20.483	23.209	29.588

Mixed examination practice

Exam-style questions

1. The table below shows cell phone ownership for different age groups.

	18–34	35–54	55+
Own cell phone	440	590	707
Do not own cell phone	28	86	362

Use a χ^2 test at the 5% significance level to test whether cell phone ownership is independent of age. Show all stages of your test, including the null and alternative hypotheses, number of degrees of freedom and the expected frequencies.

2. The table below shows the voting behaviour of a sample of workers. Omar wants to test whether voting behaviour is associated with the type of work that voters do.

	Party		
	PPP	CPP	SLP
Manual workers	37	24	44
Non-manual workers	57	30	19

A chi-squared test at the 10% level of significance is performed.

(a) State the null hypothesis.

(b) State the number of degrees of freedom.

(c) Determine the chi-squared statistic for this data.

The χ^2 critical value at the 10% level of significance is 4.61.

(d) What conclusion can Omar draw from this test? Give a reason for your answer.

3. Mrs Elwood, the Director of Studies in a secondary school, is analysing the school's GCSE results to investigate whether low performance in Mathematics (at grade F or lower) is independent of gender. The table shows the number of low grades and failures in the most recent examination session.

	F	G	U
Boys	43	22	13
Girls	31	14	9

Mrs Elwood uses the chi-squared test at the 5% level of significance.

(a) State a suitable null hypothesis, H_0.

(b) State the number of degrees of freedom.

(c) Determine the chi-squared statistic for this data.

The χ^2 critical value at the 5% significance level is 5.99.

(d) State whether or not Mrs Elwood will reject the null hypothesis, justifying your answer clearly.

4. A salesman has collected the following data on the sales of smart phones and other cell phones.

		Smart phones	Other cell phones
Age group	18–34	252	188
	35–54	235	355
	55+	165	542

He wishes to test whether ownership of smart phones depends on age.

To do so he performs a chi-squared test at the 5% level of significance.

(a) State the null hypothesis.

(b) State the number of degrees of freedom.

(c) Determine the chi-squared statistic for this data.

The χ^2 critical value at the 5% level of significance is 5.99.

(d) What conclusion can be drawn from this test? Give a reason for your answer.

5. A study was conducted by a polling agency to determine whether there is a relationship between voting behaviour and age of voters. The data gathered is shown below.

		Age of voters			
		18–24	25–34	35–54	55+
Political party	Action Party	204	343	354	429
	Progress Party	412	380	462	406
	National Party	227	124	280	154

Set up a suitable null hypothesis and test it at the 5% level of significance. You may take the χ^2 critical value for the data to be 12.59.

6. The performance of 200 students in their IB Mathematics examinations is shown in the table.
A chi-squared test is applied to the data to test whether high performance is related to the level at which the subject was studied.

	Level 7	Level 6	Level 5
Mathematical Studies	50	23	11
Mathematics SL	29	26	9
Mathematics HL	20	19	13

(a) State the null hypothesis, H_0, for this data.

(b) Calculate the expected numbers of Level 5, Level 6 and Level 7 grades for Mathematics HL.

Using a 5% level of significance the p-value was found to be 0.0806 correct to 3 significant figures.

(c) State whether the null hypothesis, H_0, should be accepted. Justify your answer.

7. The Road Safety Committee of a Local Authority has collected the following data:

	Age in years				
	≤ 19	20–24	25–44	45–64	≥ 65
Drivers in accidents	21	40	46	14	10
Drivers not in accidents	42	60	68	69	31

Set up a suitable null hypothesis and test it at the 5% level of significance. Perform your test using the *p*-value and confirm your result using the chi-squared test statistic. You may take the χ^2 critical value for the data to be 7.78. Show all the relevant stages of your test.

Past paper questions

1. Manuel conducts a survey on a random sample of 751 people to see which television programme type they watch most from the following: Drama, Comedy, Film, News. The results are as follows.

	Drama	Comedy	Film	News
Males under 25	22	65	90	35
Males 25 and over	36	54	67	17
Females under 25	22	59	82	15
Females 25 and over	64	39	38	46

Manuel decides to ignore the ages and to test at the 5% level of significance whether the most watched programme type is independent of **gender**.

(a) Draw a table with 2 rows and 4 columns of data so that Manuel can perform a chi-squared test. *[3 marks]*

(b) State Manuel's null hypothesis and alternative hypothesis. *[1 mark]*

(c) Find the expected frequency for the number of females who had "Comedy" as their most-watched programme type. Give your answer to the nearest whole number. *[2 marks]*

(d) Using your graphic display calculator, or otherwise, find the chi-squared statistic for Manuel's data. *[3 marks]*

(e) (i) State the number of degrees of freedom available for this calculation.

(ii) State the critical value for Manuel's test.

(iii) State his conclusion. *[3 marks]*

[Total 12 marks]

[Nov 2007, Paper 2, TZ0, Question 4(ii)] (*© IB Organization 2007*)

2. The local park is used for walking dogs. The sizes of the dogs are observed at different times of the day. The table below shows the numbers of dogs present, classified by size, at three different times last Sunday.

$$\begin{array}{ll} & \begin{array}{ccc}\text{Small} & \text{Medium} & \text{Large}\end{array} \\ \begin{array}{l}\text{Morning}\\ \text{Afternoon}\\ \text{Evening}\end{array} & \begin{pmatrix} 9 & 18 & 2 \\ 11 & 6 & 13 \\ 7 & 8 & 9 \end{pmatrix}\end{array}$$

(a) Write a suitable null hypothesis for a χ^2 test on this data.

(b) Write down the value of χ^2 for this data.

(c) The number of degrees of freedom is 4. Show how this value is calculated.

The critical value, at the 5% level of significance, is 9.488.

(d) What conclusion can be drawn from this test? Give a reason for your answer. *[Total 6 marks]*

[May 2007, Paper 1, TZ0, Question 8] (© *IB Organization 2007*)

3. 200 people of different ages were asked to choose their favourite type of music from the choices Popular, Country and Western and Heavy Metal. The results are shown in the table below.

Age/Music choice	**Popular**	**Country and Western**	**Heavy Metal**	**Totals**
11–25	35	5	50	90
26–40	30	10	20	60
41–60	20	25	5	50
Totals	85	40	75	200

It was decided to perform a chi-squared test for independence at the 5% level on the data.

(a) Write down the null hypothesis. *[1 mark]*

(b) Write down the number of degrees of freedom. *[1 mark]*

(c) Write down the chi-squared value. *[2 marks]*

(d) State whether or not you will reject the null hypothesis, giving a clear reason for your answer. *[2 marks]*
[Total 6 marks]

[May 2008, Paper 1, TZ1, Question 5] (© *IB Organization 2008*)