Summary of key points

1. The area under a continuous probability distribution is equal to 1.

2. If X is a normally distributed random variable, you write $X \sim N(\mu, \sigma^2)$ where μ is the population mean and σ^2 is the population variance.

3. The normal distribution
   - has parameters μ, the population mean, and σ^2, the population variance
   - is symmetrical (mean = median = mode)
   - has a bell-shaped curve with asymptotes at each end
   - has total area under the curve equal to 1
   - has points of inflection at $\mu + \sigma$ and $\mu - \sigma$

4. The standard normal distribution has mean 0 and standard deviation 1.

   The standard normal variable is written as $Z \sim N(0, 1^2)$.

5. If n is large and p is close to 0.5, then the binomial distribution $X \sim B(n, p)$ can be approximated by the normal distribution $N(\mu, \sigma^2)$ where
   - $\mu = np$
   - $\sigma = \sqrt{np(1-p)}$

6. If you are using a normal approximation to a binomial distribution, you need to apply a **continuity correction** when calculating probabilities.

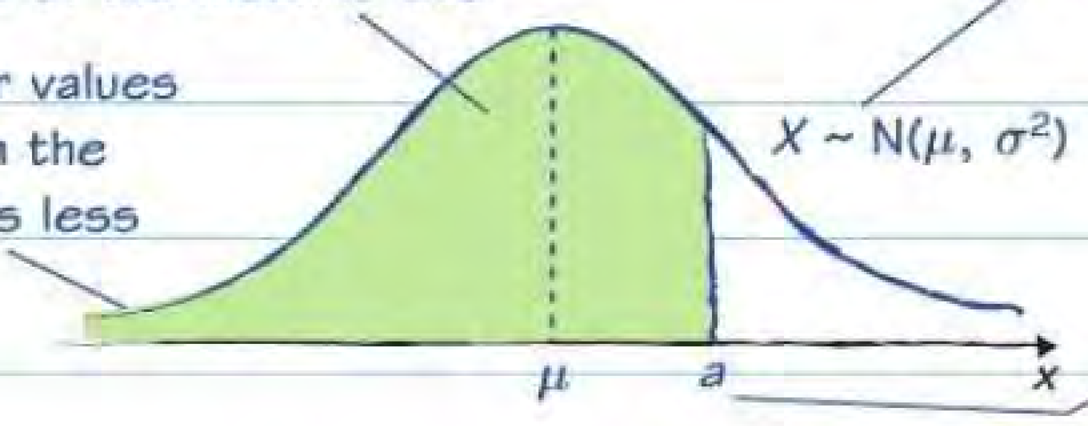
The normal distribution 1

The normal distribution is a good model for lots of **continuous** distributions in real life. A normal distribution is defined by its **mean**, μ, and its **standard deviation**, σ. You write $N(\mu, \sigma^2)$.

The shaded area represents the probability that $X < a$ (or $X \leqslant a$). The total area under the curve is 1. The curve is symmetrical, so the area to the left of the mean is 0.5

The curve never touches zero, but for values more than 4 standard deviations from the mean it is very close. $P(x < \mu - 4\sigma)$ is less than 0.0001

This means 'X is normally distributed with mean μ and standard deviation σ'.



You can find $P(X < a)$ using your calculator.

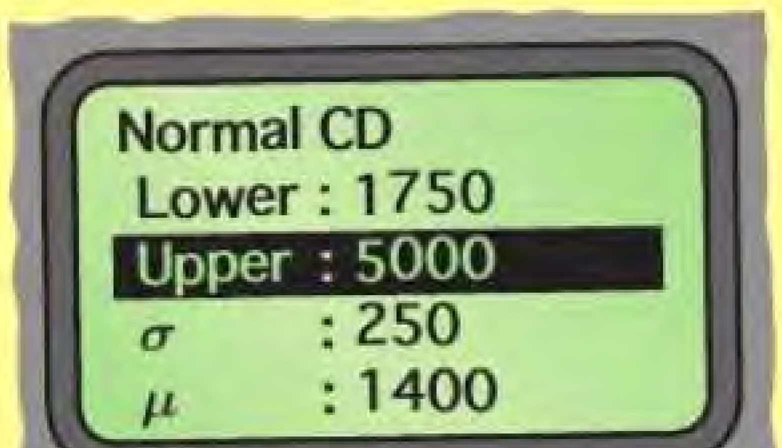
Using your calculator

You will be expected to find probabilities for a normal distribution using your calculator. Use the '*Normal cumulative distribution*', or '*Normal CD*' function.

You might have to enter lower **and** upper bounds for the probability. If you need to find the probability that a normally distributed random variable is **below** a given amount, you should enter a lower bound at least 5 standard deviations away from the mean.

- ✓ To find $P(X < a)$, enter an extreme lower bound.
- ✓ To find $P(a < X < b)$, enter the values of a and b as the lower and upper bounds.
- ✓ To find $P(X > a)$, enter an extreme upper bound.

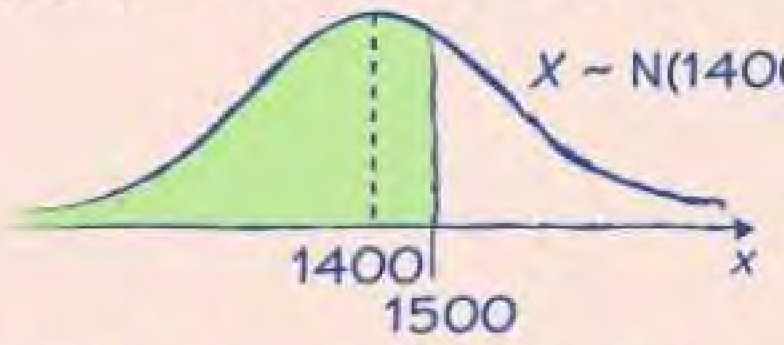
For part (c) enter a large value, such as 5000, as the upper limit in your calculator:



Worked example

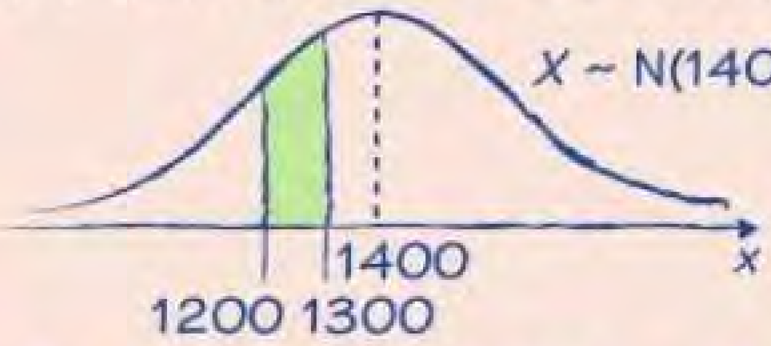
The lifetimes of fuses produced in a certain factory are normally distributed with mean 1400 hours and standard deviation 250 hours. Find the probability that a randomly chosen fuse has a lifetime of

(a) less than 1500 hours



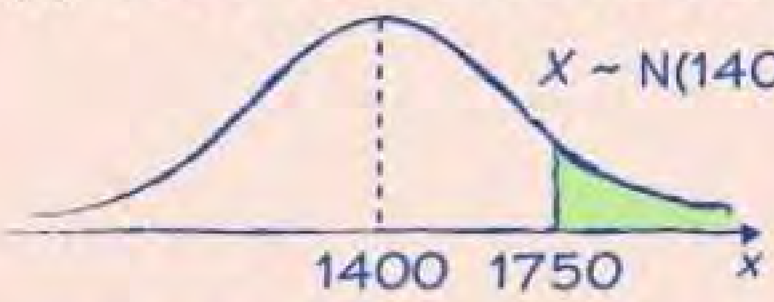
$P(X < 1500) = 0.6554$

(b) between 1200 and 1300 hours



$P(1200 < X < 1300) = 0.1327$

(c) more than 1750 hours.



$P(X > 1750) = 0.0808$

Now try this

1 The heights of a group of dogs, X mm, are normally distributed with mean 410 mm and standard deviation 125 mm. Find

(a) $P(X > 500)$ Answer **(1 mark)**

(b) $P(X < 350)$ Answer **(1 mark)**

(c) $P(380 < X < 420)$ Answer **(1 mark)**

2 The random variable $Y \sim N(2.4, 0.5)$. Find the probability that Y takes a value greater than 3 or less than 2. **(2 marks)**

Answer

$\sigma^2 = 0.5$, so $\sigma = \sqrt{0.5}$

The normal distribution 2

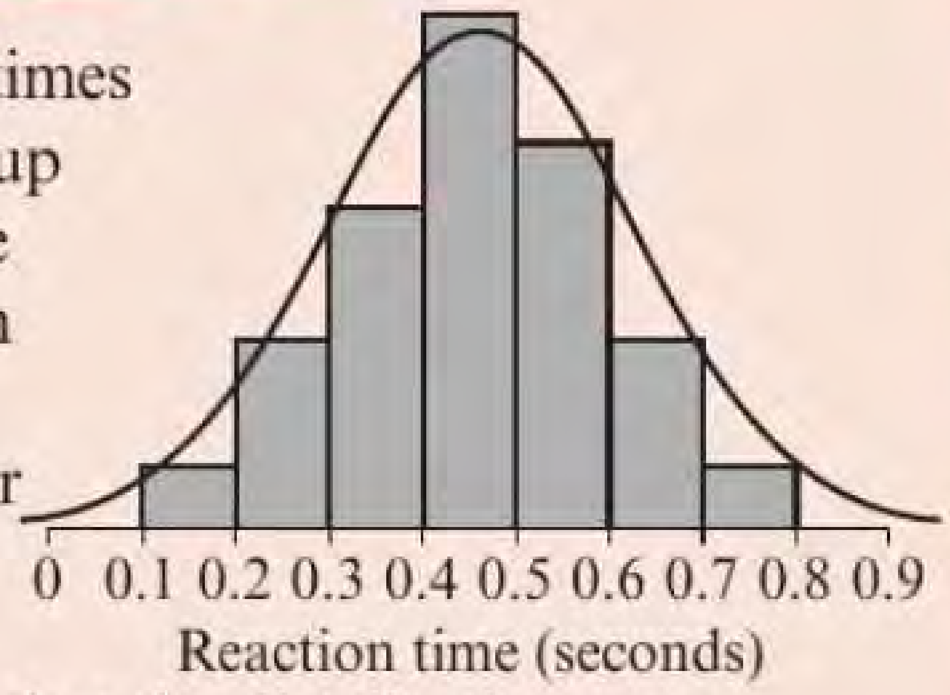
Here are three key facts you need to know about the normal distribution $X \sim N(\mu, \sigma^2)$:

1. The curve has **points of inflection** at $\mu \pm \sigma$
2. The distribution is **symmetrical**, so the mean, median and mode are equal.
3. About 68% of values lie within one standard deviation of the mean, and 95% of values lie within two standard deviations of the mean.

The points of inflection on a normal distribution curve will occur at $\mu - \sigma$ and $\mu + \sigma$. Look for the points where the curve changes from being concave to convex.

Worked example

The reaction times of a large group of adults were recorded in an experiment. The researcher drew a histogram and observed that the distribution was approximately normal.



Use the normal approximation curve drawn above to estimate

(a) the mean reaction time **(1 mark)**

0.45 seconds

(b) the standard deviation of the reaction times. **(1 mark)**

0.15 seconds

Worked example

The scores on a test are modelled as being normally distributed with mean 60% and standard deviation 7%. The pass-mark for the test is 55%.

A class of 30 students take the test.

(a) Find the probability that more than 20 students pass the test. **(4 marks)**

$X \sim N(60, 7^2)$

$P(X \geqslant 55) = 0.7625$

Let S = number of students who pass

$S \sim B(30, 0.7625)$

$P(S > 20) = 0.8461$

A student claims that the model is not realistic because in real life it is impossible to score more than 100%.

(b) Comment on the student's claim. **(1 mark)**

100% is more than 5 standard deviations from the mean, so the probability of a value greater than 100% would be virtually zero. So the model could still be realistic.

Problem solved!

You can model the number of students out of 30 who pass the test as a **binomial random variable**. You need to use the normal distribution to work out p, the probability that a single student will pass.

Revise the binomial distribution on page 131.

You will need to use problem-solving skills throughout your exam – **be prepared!**

$Z \sim N(0, 1^2)$

The normal distribution with mean 0 and standard deviation 1 is sometimes called the **standard normal distribution**. You can **standardise** a normal random variable $X \sim N(\mu, \sigma^2)$ using the coding:

$$Z = \frac{X - \mu}{\sigma}$$

You will make use of the standardised normal distribution curve on page 143.

Values from the standard normal distribution are sometimes called **z-values**.

Now try this

The weights of bags of compost filled by a machine are distributed normally with mean 20.5 kg and standard deviation 0.4 kg.

The bags are advertised as containing 20 kg of compost, and the company must refill any bags weighing less than this. A customer buys 10 bags. Find the probability that fewer than 4 need refilling.

Answer **(4 marks)**

The inverse normal function

You can use the inverse normal function on your calculator to find the value of a normal random variable associated with a particular probability.

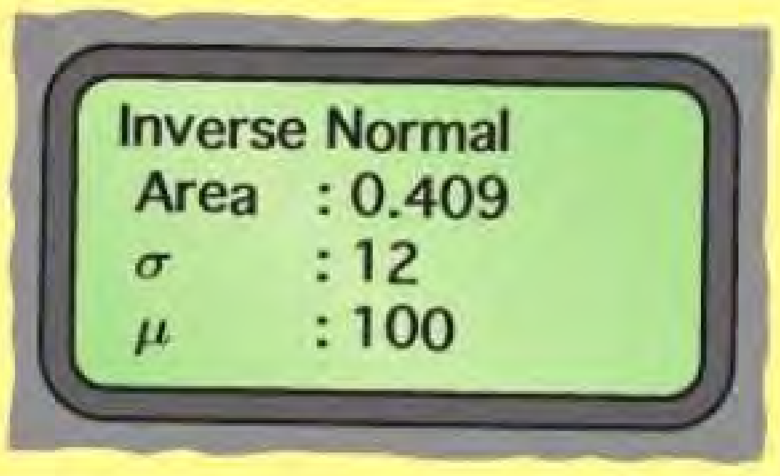
Worked example

A random variable X is normally distributed with mean 100 and standard deviation 12.

Find a such that $P(X < a) = 0.409$ **(4 marks)**

$a = 97.24$ (2 d.p.)

Use your calculator. You might need to enter the probability as 'Area'. This is because it represents the area under the normal distribution curve to the left of the value you want to find.

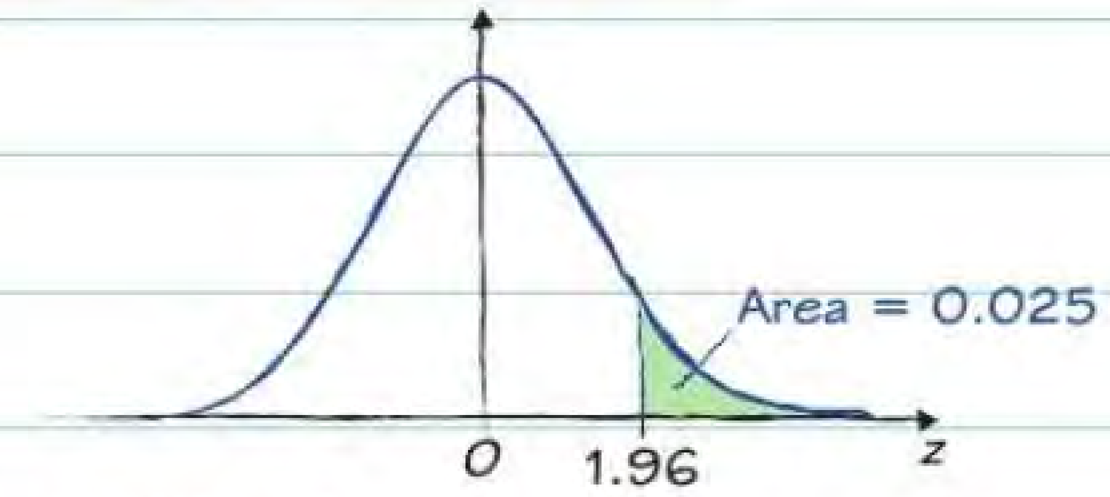


Using tables

The **percentage points table** in the formulae booklet tells you values of Z for certain probabilities, where Z is the **standard normal distribution**, $Z \sim N(0, 1^2)$.

p	z
0.0500	1.6449
0.0250	1.9600
0.0100	2.3263

This row tells you that $P(Z > 1.96) = 0.025$



Be careful. This table gives areas to the **right** of a z-value, and not all the probabilities are listed.

Worked example

The lengths of the films released in one year, L minutes, are normally distributed with

$L \sim N(128, 15^2)$

(a) Write down the median length of film. **(1 mark)**

128 minutes

(b) Find the upper quartile, Q_3, of L. **(3 marks)**

$P(L < Q_3) = 0.75$

So $Q_3 = 138$ (3 s.f.)

(c) Write down the lower quartile, Q_1, of L. **(1 mark)**

$Q_1 = 128 - (138 - 128) = 118$

(a) A normal distribution is **symmetrical** so mean = median.

(b) 25% of the values in the distribution are above the upper quartile. To use the inverse normal function on your calculator you need to enter the probability that the value lies **below** the point you are looking for, so enter 0.75 as the area.

(c) L is symmetrical so $Q_3 - 128 = 128 - Q_1$

You could also use your calculator to find Q_1 such that $P(L < Q_1) = 0.25$.

Now try this

The weights of some courgettes, W grams, were modelled by $W \sim N(450, 100^2)$.

(a) Find w such that $P(432 < W < w) = 0.3$ **(4 marks)**

Answer

Find $P(W < 432)$, then add 0.3 to find $P(W < w)$. Then use the inverse normal function on your calculator.

Two courgettes are chosen at random.

(b) Find the probability that only one weighs between 432 grams and w grams. **(3 marks)**

Answer

There are two successful outcomes:

1. First courgette ✓ Second courgette ✗
2. First courgette ✗ Second courgette ✓

Had a look ☐ Nearly there ☐ Nailed it! ☐

Finding μ and σ

You might need to use information about a normal distribution to find its **mean** (μ) and its **standard deviation** (σ). You need to make use of the **standard normal distribution**, $Z \sim N(0, 1^2)$.

Worked example

The times taken for a search engine to complete a web search are normally distributed with mean 0.63 seconds. The company states that 97.5% of searches are completed in less than 1 second.

Find the standard deviation of the times taken to complete a web search. **(4 marks)**

$P(X < 1) = P\left(Z < \frac{1 - 0.63}{\sigma}\right) = 0.975$

$\frac{1 - 0.63}{\sigma} = 1.96$

$0.37 = 1.96\sigma$

$\sigma = 0.189$ (3 s.f.)

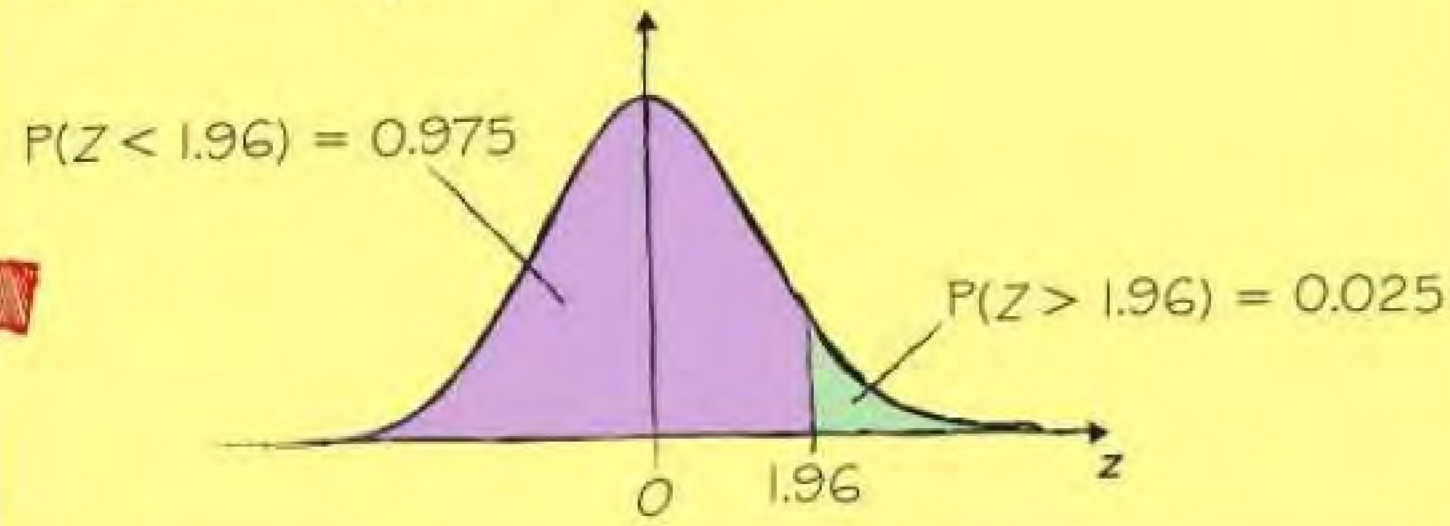
If you need to find μ or σ in your exam, you will often be able to use the **percentage points table** in the booklet. Have a look at page 142 for a reminder on how this table works.

$P(X < 1) = 0.975$ so

$P(X > 1) = 1 - 0.975 = 0.025$

The percentage points table tells you that this occurs at $z = 1.96$

So $z = \frac{1 - 0.63}{\sigma} = 1.96$



Worked example

X is a normally distributed random variable with mean μ and standard deviation σ.

$P(X > 8.6) = 0.3$ and $P(X < 7.7) = 0.05$

(a) Show that $\mu = 7.7 + 1.6449\sigma$ **(3 marks)**

$P(X < 7.7) = P\left(Z < \frac{7.7 - \mu}{\sigma}\right) = 0.05$

$\frac{7.7 - \mu}{\sigma} = -1.6449$

$\mu = 7.7 + 1.6449\sigma$

(b) Obtain a second equation and hence find the value of μ and the value of σ. **(4 marks)**

$P(X > 8.6) = P\left(Z > \frac{8.6 - \mu}{\sigma}\right) = 0.3$

$\frac{8.6 - \mu}{\sigma} = 0.5244$

$\mu = 8.6 - 0.5244\sigma$

So $7.7 + 1.6449\sigma = 8.6 - 0.5244\sigma$

$2.1693\sigma = 0.9$

$\sigma = 0.4149...$

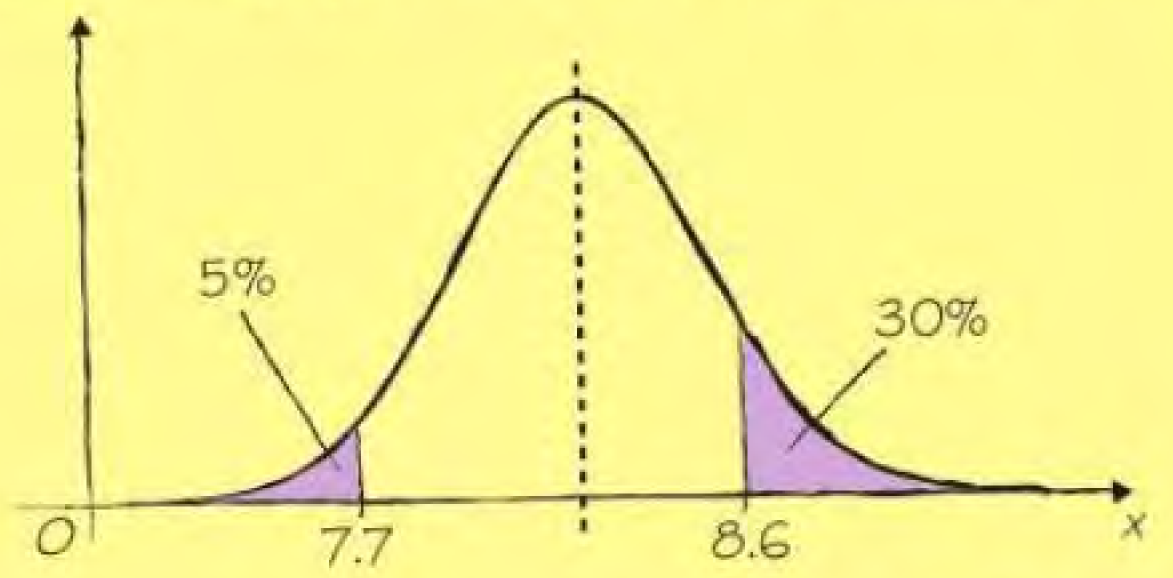
$= 0.415$ (3 s.f.)

$\mu = 8.6 - 0.5244 \times 0.4149...$

$= 8.3824...$

$= 8.38$ (3 s.f.)

You can show this information on a sketch. This can help you visualise the problem.



The sketch makes it clear that $x = 7.7$ will give a **negative** z-value, that $x = 8.6$ will give a **positive** z-value, and that μ is between 7.7 and 8.6

If you don't know μ **or** σ, and you are given **two** probability facts, then you will have to solve a pair of **simultaneous equations** to find μ and σ.

Now try this

The weights of the oranges in a crate are normally distributed with mean μ grams and standard deviation σ grams. 20% of the oranges are lighter than 175 grams and 10% are heavier than 230 grams.

Find the value of μ and the value of σ.

Answer **(6 marks)**

Normal approximations

Binomial probabilities can be difficult to calculate for **large values of n**. In some situations, you can use a normal distribution to approximate a binomial distribution. This approximation is valid provided that **n is large** and that **p is close to 0.5**

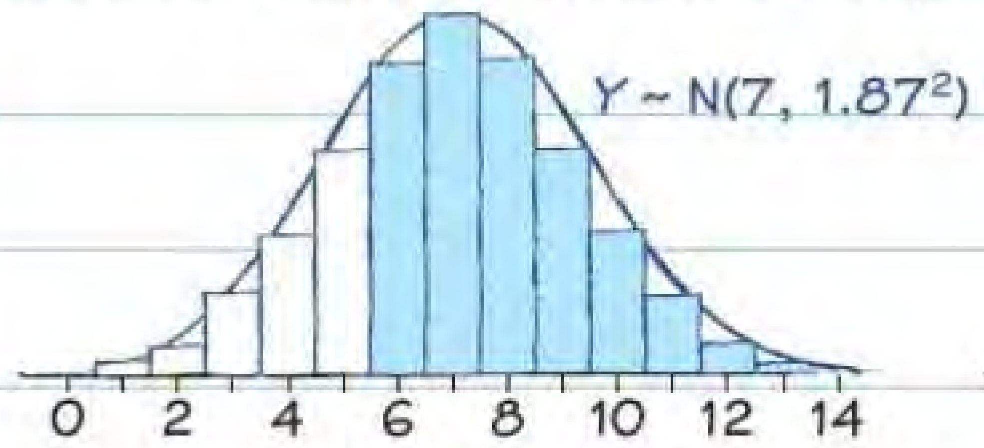
Golden rule

For large n, and p close to 0.5, you can approximate the binomial distribution $B(n, p)$ with the normal distribution $N(np, np(1-p))$.

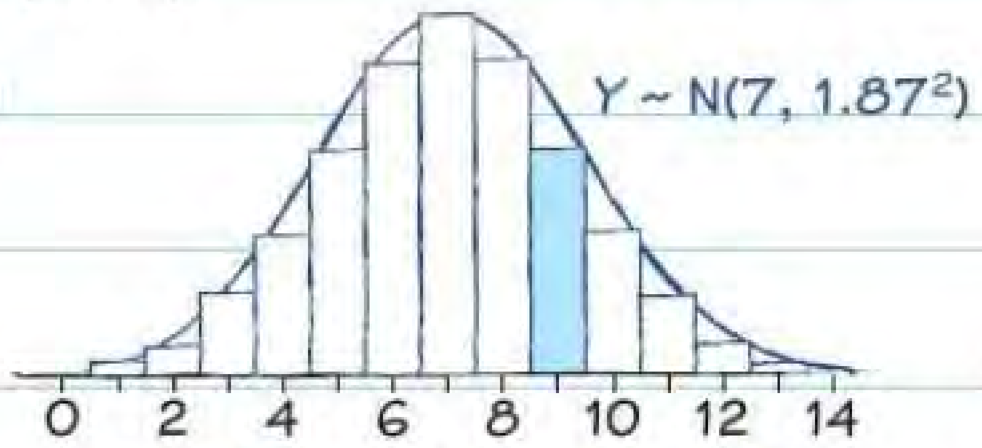
The **standard deviation** is $\sqrt{np(1-p)}$.

Continuity corrections

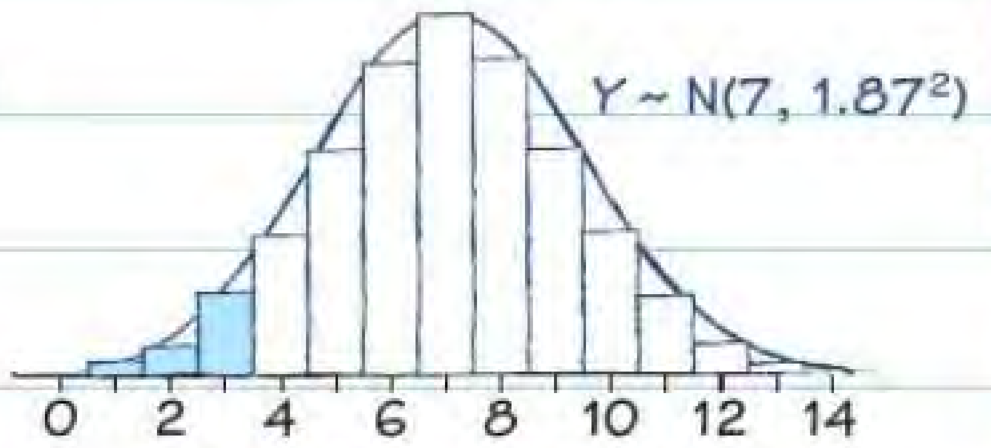
Because the normal distribution is a **continuous distribution**, the probability that a normal random variable takes **exactly** one particular value is 0. When using a normal distribution to approximate a binomial distribution you need to consider a range of values instead. This is called a **continuity correction**. Here the normal random variable $Y \sim N(7, 1.87^2)$ is being used to approximate the binomial random variable $X \sim B(14, 0.5)$:



$P(X > 5) \approx P(Y > 5.5)$



$P(X = 9) \approx P(8.5 < Y < 9.5)$



$P(X \leqslant 3) \approx P(Y < 3.5)$

Worked example

A dice is biased, so that the probability of rolling an even number is 0.46. The dice is rolled 300 times.

(a) Write down a binomial model for X, the number of times the dice lands on an even number. **(1 mark)**

$X \sim B(300, 0.46)$

(b) Explain why X can be approximated with a normal distribution, and state its mean and standard deviation. **(3 marks)**

The number of trials is large, and the probability on each trial is close to 0.5, so X can be approximated by $N(\mu, \sigma^2)$ where

$\mu = 300 \times 0.46 = 138$

$\sigma = \sqrt{300 \times 0.46(1 - 0.46)} = 8.632$ (3 d.p.)

(c) Estimate the probability that the dice lands on an even number at least 140 times. **(1 mark)**

$Y \sim N(138, 8.632^2)$

$P(X \geqslant 140) \approx P(Y > 139.5) = 0.4310$

Read the question carefully before applying your continuity correction. Part (c) says 'at least' so 140 is included. This means you need to consider values of the normal random variable greater than 139.5

In part (b), choose your continuity correction carefully: 100 and 120 should both be included.

Now try this

Two computer artificial intelligence programs, Deep Thought and Grandmaster, play backgammon against each other. The probability that Deep Thought wins each game is 0.56.

20 games are played.

(a) Calculate the probability that Deep Thought wins exactly 10 of these games. **(1 mark)**

A further 200 games are played.

(b) Use a suitable approximation to estimate the probability that Deep Thought wins between 100 and 120 of these games inclusive. **(3 marks)**

(c) Justify the validity of your approximation. **(1 mark)**

Choosing a Distribution

By now, you should be familiar with both the binomial and normal distributions. If you're not, it's worth having another read through this section until it's all clear in your head. Then come back to this page — I'll wait for you.

*Learn the **Conditions** for **Binomial** and **Normal Distributions***

You might be given a situation and asked to choose which distribution would be suitable.

Conditions for a Binomial Distribution

1) The data is **discrete**.
2) The data represents the number of '**successes**' in a **fixed number of trials** (n), where each trial results in **either** 'success' or 'failure'.
3) All the trials are **independent**, and the probability of success, p, is **constant**.

If these conditions are met, the data can be modelled by a **binomial distribution**: $B(n, p)$.

You saw these conditions on p.162.

Conditions for a Normal Distribution

1) The data is **continuous**.
2) The data is roughly **symmetrically distributed**, with a **peak** in the middle (at the **mean**, μ).
3) The data '**tails off**' either side of the mean — i.e. data values become **less frequent** as you move further from the mean. Virtually **all** of the data is within **3 standard deviations** (σ) of the mean.

If these conditions are met, the data can be modelled by a **normal distribution**: $N(\mu, \sigma^2)$.

Example: For each random variable below, decide if it can be modelled by a binomial distribution, a normal distribution or neither.

a) The number of faulty items (T) produced in a factory per day, if items are faulty independently with probability 0.01 and there are 10 000 items produced every day.

Binomial — there's a **fixed number** of **independent** trials (10 000) with **two possible results** ('faulty' or 'not faulty'), a **constant probability of 'success'**, and **T is the total number** of 'faulty' items. So $T \sim B(10\,000, 0.01)$.

b) The number of red cards (R) drawn from a standard 52-card deck in 10 picks, not replacing the cards each time.

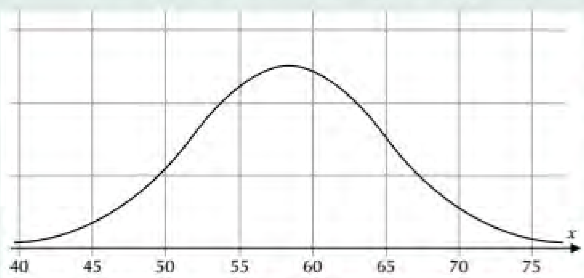
Neither — the data is **discrete** so it can't be modelled by the normal distribution, but the **probability of 'success' changes** each time (as the cards aren't replaced) so it can't be modelled by the binomial distribution.

c) The heights (H) of all the girls in a Sixth Form college.

Normal — the data is **continuous**, and you would expect heights to be distributed **symmetrically**, with most girls' heights close to the mean and a few further away. So $H \sim N(\mu, \sigma^2)$ (where μ and σ are to be calculated).

*Use **Facts** about the distribution to **Estimate Parameters***

Example: The times taken by runners to finish a 10 km race, x minutes, are normally distributed. Data from the race is shown on the diagram below. Estimate the mean and standard deviation of the times.



The mean is in the middle, so $\mu \approx 58$ **minutes**.

For a normal distribution there is a **point of inflection** at $x = \mu + \sigma$ (see p.164). Use the diagram to estimate the point of inflection. This is where the line changes from **concave** to **convex** (see p.90) — it looks like this at about $x \approx 65$.

Use your values for x and μ to estimate σ.

$65 = 58 + \sigma \Rightarrow \sigma \approx 7$ **minutes**

You could also use the point of inflection at $x = \mu - \sigma$.

Choosing a Distribution

*Once you've **Chosen** a distribution, use it to **Answer Questions***

Example: A restaurant has several vegetarian meal options on its menu. The probability of any person ordering a vegetarian meal is 0.15. One lunch time, 20 people order a meal.

a) Suggest a suitable model to describe the number of people ordering vegetarian meals.

b) Use this model to find the probability that at least 5 people order a vegetarian meal.

a) There are a **fixed number of trials** (20 meals), with probability of success (i.e. vegetarian meal) **0.15**. If X is the number of people ordering a vegetarian meal, then $X \sim B(20, 0.15)$.

b) Use your calculator, with $n = 20$ and $p = 0.15$:
$P(X \geq 5) = 1 - P(X < 5) = 1 - P(X \leq 4) = 1 - 0.8298... = \mathbf{0.170}$ (3 s.f.)

Example: The heights of 1000 sunflowers from the same field are measured. The distribution of the sunflowers' heights is symmetrical about the mean of 9.8 ft, with the shortest sunflower measuring 5.8 ft and the tallest measuring 13.7 ft. The standard deviation of the sunflowers' heights is 1.3 ft.

a) Explain why the distribution of the sunflowers' heights might reasonably be modelled using a normal distribution.

b) From these 1000 sunflowers, those that measure 7.5 ft or taller are harvested. Estimate the number of sunflowers that will be harvested.

c) Explain why you shouldn't use your answer to part b) to estimate the number of sunflowers harvested from a crop of 1000 sunflowers from a different field.

a) The data collected is **continuous**, and the distribution of the heights is **symmetrical** about the **mean**. This is also true for a normally distributed random variable X. **Almost all** of the data is within **3 standard deviations** of the mean: $9.8 - (3 \times 1.3) = 5.9$ and $9.8 + (3 \times 1.3) = 13.7$.
So the random variable $X \sim N(9.8, 1.3^2)$ seems like a reasonable model for the sunflowers' heights.

b) Using a calculator: $P(X \geq 7.5) = 0.961572...$
Multiply the total number of sunflowers by this probability:
$1000 \times 0.961572... = \mathbf{962}$ (to the nearest whole number).

c) The **mean** and **standard deviation** of another crop of sunflowers could be **different** (because of varying sunlight, soil quality etc.), so you shouldn't use 962 as an estimate. However, it would still be reasonable to assume that their heights were normally distributed — just with different values of μ and σ.

Practice Question

Q1 Explain whether each random variable can be modelled by a binomial or normal distribution or neither.

a) The number of times (T) I have to roll a fair standard six-sided dice before I roll a 6.

b) The distances (D) of a shot put thrown by a class of 30 Year 11 students in a PE lesson.

c) The number of red cars (R) in a sample of 1000 randomly chosen cars, if the proportion of red cars in the population as a whole is 0.08.

Exam Question

Q1 A biologist tries to catch a hedgehog every night for two weeks using a humane trap. She either succeeds in catching a hedgehog, or fails to catch one.

a) The biologist believes that this situation can be modelled by a random variable following a binomial distribution.

(i) State two conditions needed for a binomial distribution to arise here. [2 marks]

(ii) State which quantity would follow a binomial distribution (assuming the above conditions are satisfied). [1 mark]

b) If the biologist successfully catches a hedgehog, she records its weight. Explain why a normal distribution might be a suitable model for the distribution of these times. [2 marks]

You can't choose your family, but you can choose your distribution...

These two pages are really just bringing together everything you've learnt in this section — there shouldn't be anything about choosing a distribution that surprises you. I've saved all the surprises for the next section — read on, read on...