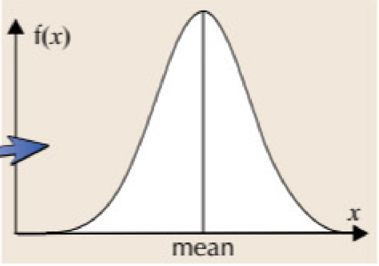
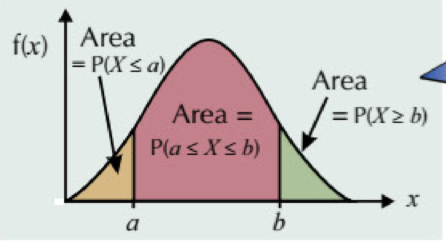
The Normal Distribution

The normal distribution is everywhere in statistics. Everywhere, I tell you. So learn this well...

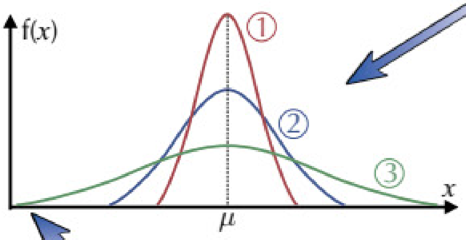
The **Normal Distribution** is '**Bell-Shaped**'



1) Loads of things in real life are most likely to fall '**somewhere in the middle**', and are much less likely to take **extremely high** or **extremely low** values. In this kind of situation, you often get a **normal distribution**.
2) If you were to draw a graph showing how likely different values are, you'd end up with a graph that looks a bit like a **bell**. There's a peak in the middle at the **mean** (or **expected value**). And the graph is **symmetrical** — so values the same distance **above** and **below** the mean are **equally likely**.



3) With a graph of a normal distribution, the probability of the random variable taking a value **between two limits** is the **area under the graph** between those limits. The **total probability** is 1, so the **total area** under the graph must also be **1**. And since the mean, μ, is in the middle, $\mathbf{P(X \le \mu) = P(X \ge \mu) = 0.5}$.
4) These three graphs all show normal distributions with the **same mean** (μ), but **different variances** (σ^2) — see page 142. Graph 1 has a **small** variance, and graph 3 has a **larger** variance — but the total area under all three curves is the **same** (= 1).
5) There are **points of inflection** (see p.90) at $\mathbf{x = \mu + \sigma}$ and $\mathbf{x = \mu - \sigma}$. About **two-thirds** of the total area lies within **1 standard deviation** of the mean (i.e. $\mu \pm \sigma$), **95%** of the total area lies within **2 standard deviations** of the mean ($\mu \pm 2\sigma$) and **almost all** of the total area lies within 3 standard deviations of the mean ($\mu \pm 3\sigma$).
6) A very useful normal distribution is the **standard normal distribution**, or Z — this has a **mean of zero** and a **variance of 1**.



As you get further from the mean, the normal distribution tends to 0, but never touches it.

Normal Distribution: $N(\mu, \sigma^2)$

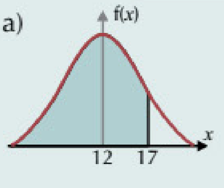
- If X is normally distributed with **mean μ** and **variance σ^2**, it's written $X \sim N(\mu, \sigma^2)$.
- The **standard normal distribution** Z has **mean 0** and **variance 1**, i.e. $Z \sim N(0, 1)$.

Use your **Calculator** to work out **Probabilities**

When working out probabilities, it's usually a good idea to draw a **sketch** showing the area you're trying to find. Then you can use the **normal cumulative distribution function** on your calculator — you'll have to enter the values of the **mean** (μ), the **standard deviation** (σ) and the **x-values** you're interested in.

Example: $X \sim N(12, 16)$. Find: a) $P(X \le 17)$, b) $P(X > 10)$, c) $P(8 \le X \le 15)$

a)

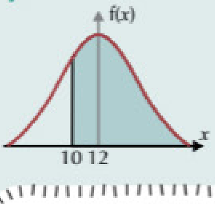


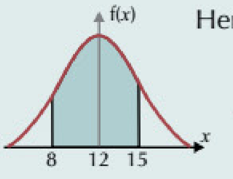
Draw a sketch, then enter $\mu = 12$, $\sigma = \sqrt{16} = 4$, the upper bound $x = 17$ and a lower bound $x = -9999$. Then $P(X \le 17) = \mathbf{0.894}$ (3 s.f.)

If you're looking for $P(X \le x)$ or $P(X \ge x)$, you might need to choose your own lower or upper bound — just pick a large negative or positive number (for parts a) and b) I've used −9999 and 9999 respectively).

b) As for part a), draw a sketch, then enter $\mu = 12$, $\sigma = 4$, the lower bound $x = 10$ and an upper bound $x = 9999$. Then $P(X > 10) = \mathbf{0.691}$ (3 s.f.)

From the fact that the area under the graph is 1, $P(X > x) = 1 - P(X \le x)$.
For continuous random variables, $P(X > x) = P(X \ge x)$ (and $P(X < x) = P(X \le x)$).





c) Here, draw a sketch then enter $\mu = 12$, $\sigma = 4$, the lower bound $x = 8$ and the upper bound $x = 15$. Then $P(8 \le X \le 15) = \mathbf{0.615}$ (3 s.f.)

You can work out this area by splitting it up and subtracting: $P(8 \le X \le 15) = P(X \le 15) - P(X \le 8)$.

If your calculator can only do the standard normal distribution, Z, you'll have to convert your distribution to Z first — see p.166.

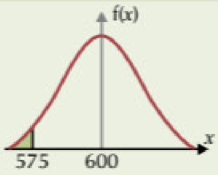
The Normal Distribution

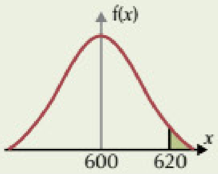
The *Normal Distribution* can be used in *Real-Life* situations

Example: The times taken by a group of people to complete an assault course are normally distributed with a mean of 600 seconds and a variance of 105 seconds. Find the probability that a randomly selected person took: a) fewer than 575 seconds, b) more than 620 seconds.

If X represents the time taken in seconds, then $X \sim N(600, 105)$.

a) You need to find $P(X < 575)$. So sketch the graph, then enter $\mu = 600$, $\sigma = \sqrt{105}$, the upper bound $x = 575$ and a lower bound $x = -9999$. Then $P(X < 575) = 0.00735$ (3 s.f.)





b) This time, you're looking for $P(X > 620)$. So sketch the graph, then enter $\mu = 600$, $\sigma = \sqrt{105}$, the lower bound $x = 620$ and an upper bound $x = 9999$. Then $P(X > 620) = 0.0255$ (3 s.f.)

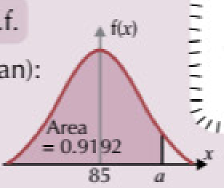
Use the *Inverse Normal Function* to find values of *x*

You might be given a **probability**, p, and asked to find the **range** of x-values where the probability of X falling in this range is p (i.e. you're told $P(X \leq a) = p$ for some value a that you have to find). For $<$ or $\leq$ questions, you can do this directly on your calculator using the **inverse normal function** — just input p, μ and σ. For $>$ or $\geq$ questions, subtract p from 1 to turn $P(X > a) = p$ into $P(X \leq a) = 1 - p$, then use your calculator as before.

You can sometimes use the standard normal distribution to answer these questions — see the next page.

Example: $X \sim N(85, 25)$. If $P(X < a) = 0.9192$, find the value of a to 2 s.f.

Draw a sketch (the probability is > 0.5, so a will be to the **right** of the mean):
Then input probability $= 0.9192$, $\mu = 85$ and $\sigma = \sqrt{25} = 5$ into the **inverse normal function** on your calculator.
So $P(X < a) = 0.9192$ for $a = 92$ (2 s.f.).



Practice Questions

Q1 If $X \sim N(50, 16)$, find the following to 4 decimal places:
a) $P(X \leq 55)$, b) $P(X < 42)$, c) $P(X > 56)$, d) $P(47 < X < 57)$.

Q2 If $X \sim N(5, 7^2)$ find the following to 4 decimal places:
a) $P(X < 0)$, b) $P(X \leq 1)$, c) $P(X \geq 7)$, d) $P(2 < X < 4)$.

Q3 If $X \sim N(28, 36)$, find the value of a such that:
a) $P(X < a) = 0.8546$ b) $P(X \geq a) = 0.2418$ c) $P(a < X < 30) = 0.5842$

Exam Questions

Q1 The random variable X has a normal distribution with mean 120 and standard deviation 25.
a) Find $P(X > 145)$. [1 mark]
b) Find the value of j such that $P(120 < X < j) = 0.4641$ [2 marks]

Q2 A garden centre sells bags of compost. The volume of compost in the bags is normally distributed with a mean of 50 litres.
a) If the standard deviation of the volume is 0.4 litres, find the probability that a randomly selected bag will contain less than 49 litres of compost. [1 mark]
b) If 1000 of these bags of compost are bought, how many bags would you expect to contain more than 50.5 litres of compost? [3 marks]

The number of ghost sightings follows a paranormal distribution...

Make sure you know how to use the normal functions on your own calculator — you'll probably use the normal cdf and inverse function a lot in these questions. It's also worth drawing a quick sketch of the graph for each question.

The Standard Normal Distribution

On p.164 I mentioned the standard normal distribution — a normal distribution with a mean of 0 and a standard deviation of 1. The standard normal distribution is represented by the letter Z, so Z ~ N(0, 1).

*Transform to Z by **Subtracting μ**, then **Dividing by σ***

1) You can convert **any** normally-distributed variable to Z by:
   (i) **subtracting the mean**, and then (ii) **dividing by the standard deviation.**

$$\text{If } X \sim N(\mu, \sigma^2), \text{ then } \frac{X-\mu}{\sigma} = Z, \text{ where } Z \sim N(0, 1)$$

This is the Greek letter 'phi'.

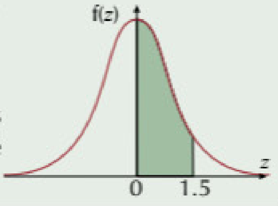
2) Values of the **cumulative distribution function** of Z can be written **Φ(*z*)** — so $P(Z \le z) = \Phi(z)$.
3) Once you've transformed a variable, you can work out the probabilities for Z using your **calculator** (some calculators can **only** work out probabilities for Z — in which case, you'll have to use this method **every time**).

Example: If $X \sim N(5, 16)$, find $P(5 < X < 11)$ by first converting to the standard normal distribution, Z.

To convert to Z, **subtract** μ (= 5) from any numbers and **divide by** σ (= $\sqrt{16} = 4$).

So $P(5 < X < 11) = P\left(\frac{5-5}{4} < Z < \frac{11-5}{4}\right) = \mathbf{P(0 < Z < 1.5)}$

Now use a **calculator** to find the probability. There are two ways you can do this — for each, enter $\mu = 0$ and $\sigma = 1$, then either use the limits 0 and 1.5, or do the calculation $P(Z < 1.5) - P(Z < 0) = 0.9331... - 0.5 = \mathbf{0.433}$ (3 s.f.)



From the symmetry of the curve, Φ(0) = 0.5.

*Use the **Percentage Points Table** if you're given a **Probability***

I know, I know — the method above seems like a bit of a faff if you can just plug the numbers straight into your calculator. But converting to Z comes in handy in a couple of different places...

You saw on the previous page how to find the value of x given a probability, p. Well, for certain 'nice' values of p (such as 0.75, 0.9, 0.95 etc.), you can use the **percentage points table** (see p.257 — it'll be on the formula sheet in your exam). The percentage points table gives you the value of z for which $\mathbf{P(Z \le z) = p}$. For example, if $P(Z \le z) = 0.75$, then from the percentage points table, $z = 0.674$.

Example: The lengths of lobsters in a lobster sanctuary are modelled by a normal distribution with a mean length of 35 cm and a standard deviation of 6 cm. Gary the lobster is l cm long, and 5% of lobsters in the sanctuary are longer than Gary. Estimate the length of Gary.

It's not immediately obvious what you're being asked here — so write down what you know:
Let L represent the length of a lobster in cm. Then $\mathbf{L \sim N(35, 6^2)}$ and $\mathbf{P(L > l) = 0.05}$.

Now, write a statement involving a probability in the percentage points table:
0.05 isn't in the percentage points table, but 1 – 0.05 = **0.95** is.
So $P(L \le l) = 1 - P(L > l) = 1 - 0.05 = 0.95$.

Look this up in the percentage points table: $P(Z \le z) = 0.95$ for $z = \mathbf{1.645}$.

Drawing a diagram might- no actually, it won't help.

Finally, transform to the standard normal distribution, Z (see above):

$\frac{l-\mu}{\sigma} = z$, so $\frac{l-35}{6} = 1.645 \Rightarrow l - 35 = 9.87 \Rightarrow l = 44.87$, so Gary is **44.9 cm long** (3 s.f.).

*Transform to Z if μ is **Unknown...***

Converting to Z is **vital** if you need to find the value of μ.

Example: $X \sim N(\mu, 2^2)$ and $P(X < 23) = 0.9015$. Find μ.

Start by transforming the probability for X into a probability for Z: $P(X < 23) = P\left(Z < \frac{23-\mu}{2}\right) = 0.9015$

Use your calculator to find the value of z for which $\Phi(z) = 0.9015$ — this gives $z = \mathbf{1.29}$ (2 d.p.)

Now form and solve an equation in μ: $\frac{23-\mu}{2} = 1.29 \Rightarrow 23 - \mu = 2.58 \Rightarrow \mu = \mathbf{20.42}$

The Standard Normal Distribution

...or if σ is Unknown

You can use a similar method if you need to find the value of σ.

Example: $X \sim N(53, \sigma^2)$ and $P(X < 50) = 0.1$. Find σ.

Again, transform the probability for X into a probability for Z:

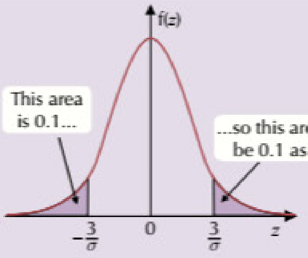
$P(X < 50) = P\left(Z < \frac{50-53}{\sigma}\right) = P\left(Z < -\frac{3}{\sigma}\right) = 0.1$.

Ideally, you'd look up 0.1 in the percentage points table to find $-\frac{3}{\sigma}$. Unfortunately, it isn't there, so you have to think a bit...

From the symmetry of the graph, $P\left(Z < -\frac{3}{\sigma}\right) = P\left(Z > \frac{3}{\sigma}\right) = 0.1$,

so $P\left(Z < \frac{3}{\sigma}\right) = 1 - 0.1 = \mathbf{0.9}$. From the percentage points table, $z = \mathbf{1.282}$.

So $\frac{3}{\sigma} = 1.282 \Rightarrow \sigma = \mathbf{2.34}$ (3 s.f.)



P(Z < –a) = P(Z > a), because Z has mean 0 and normal distributions are symmetrical.

If you have to find μ and σ, you'll need to solve Simultaneous Equations

Example: The random variable $X \sim N(\mu, \sigma^2)$. If $P(X < 9) = 0.5596$ and $P(X > 14) = 0.0322$, find μ and σ.

$P(X < 9) = P\left(Z < \frac{9-\mu}{\sigma}\right) = 0.5596$.

Using your calculator, this tells you that $\frac{9-\mu}{\sigma} = 0.15$, or $\mathbf{9 - \mu = 0.15\sigma}$ ①

$P(X > 14) = P\left(Z > \frac{14-\mu}{\sigma}\right) = 0.0322$, which means that $P\left(Z < \frac{14-\mu}{\sigma}\right) = 1 - 0.0322 = 0.9678$.

Using your calculator, this tells you that $\frac{14-\mu}{\sigma} = 1.85$, or $\mathbf{14 - \mu = 1.85\sigma}$ ②

② – ① gives: $(14 - \mu) - (9 - \mu) = 1.85\sigma - 0.15\sigma$, or $5 = 1.7\sigma$. This gives $\sigma = 5 \div 1.7 = \mathbf{2.94}$ (3 s.f.)

Now use either one of the equations to find μ: $\mu = 9 - (0.15 \times 2.94...) = \mathbf{8.56}$ (3 s.f.) ← So $X \sim N(8.56, 2.94^2)$.

Practice Questions

Q1 Find the value of z if: a) $P(Z < z) = 0.99$, b) $P(Z \leq z) = 0.0005$.

Q2 Find the value of μ if: a) $X \sim N(\mu, 10)$ and $P(X < 8) = 0.8925$,
b) $X \sim N(\mu, 8^2)$ and $P(X < 213) = 0.3085$.

Q3 Find the value of σ if: a) $X \sim N(11, \sigma^2)$ and $P(X < 13) = 0.6$,
b) $X \sim N(108, \sigma^2)$ and $P(X \geq 106) = 0.9678$.

Q4 The random variable $X \sim N(\mu, \sigma^2)$.
If $P(X < 15.2) = 0.9783$ and $P(X > 14.8) = 0.1056$, then find μ and σ.

Exam Questions

Q1 A sweet shop sells giant marshmallows. The mass of a marshmallow, in grams, is described by the random variable Y, where $Y \sim N(75, \sigma^2)$. It is found that 10% of the marshmallows weigh less than 74 grams.
Find σ. [3 marks]

Q2 The lifetimes of a particular type of battery are normally distributed with mean μ hours and standard deviation σ hours. A student using these batteries finds that 40% last less than 20 hours and 80% last less than 30 hours.
Find μ and σ. [7 marks]

The Norman Distribution came to England in 1066...

It's always the same — to transform to Z, subtract the mean and divide by the standard deviation. Just make sure you don't use the variance by mistake — remember, in $N(\mu, \sigma^2)$, the second number always shows the variance.

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(μ, σ^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 ~ 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.

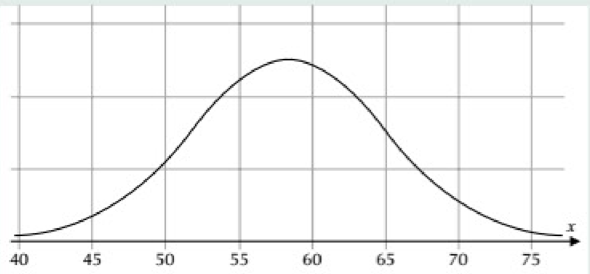
Neither — the data is **discrete** so it can't be modelled by the normal distribution, but the **probability of 'success' changes** each time (as the cards aren't replaced) so it can't be modelled by the binomial distribution.

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 ~ N(μ, σ^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 $p = 0.15$:
$P(X \geq 5) = 1 - P(X < 5) = 1 - P(X \leq 4) = 1 - 0.8298... =$ **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:
$1000 \times 0.961572... =$ **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...