Lecture 6 : The Normal Distribution

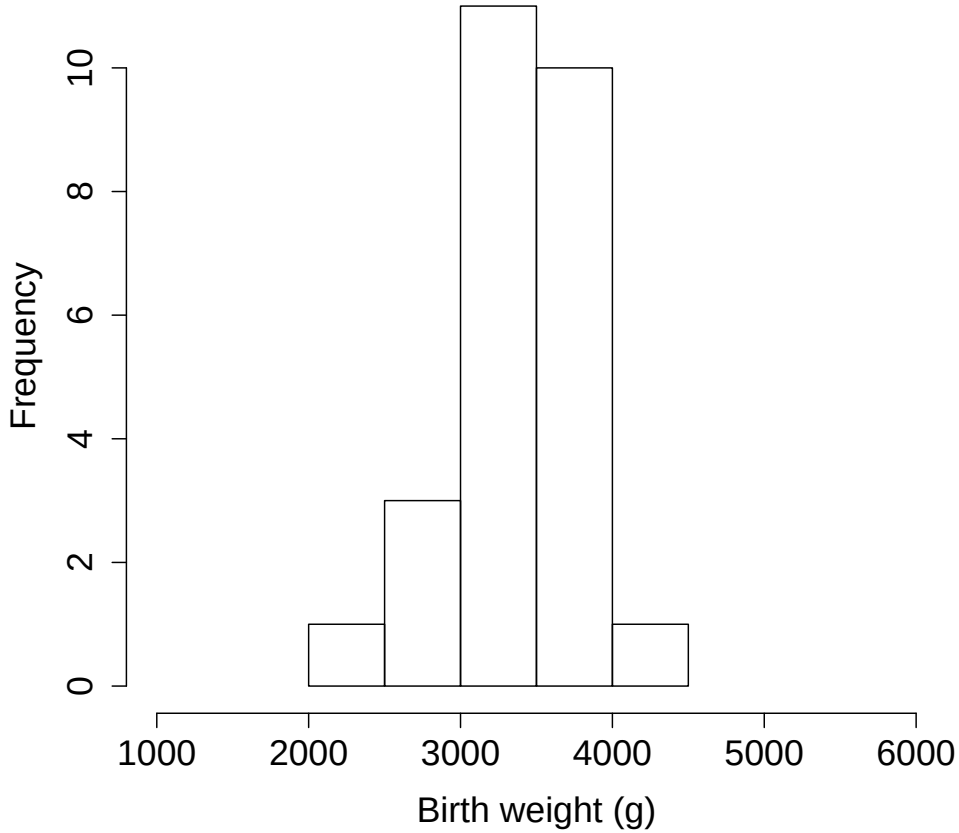
Jonathan Marchini

Continuous data

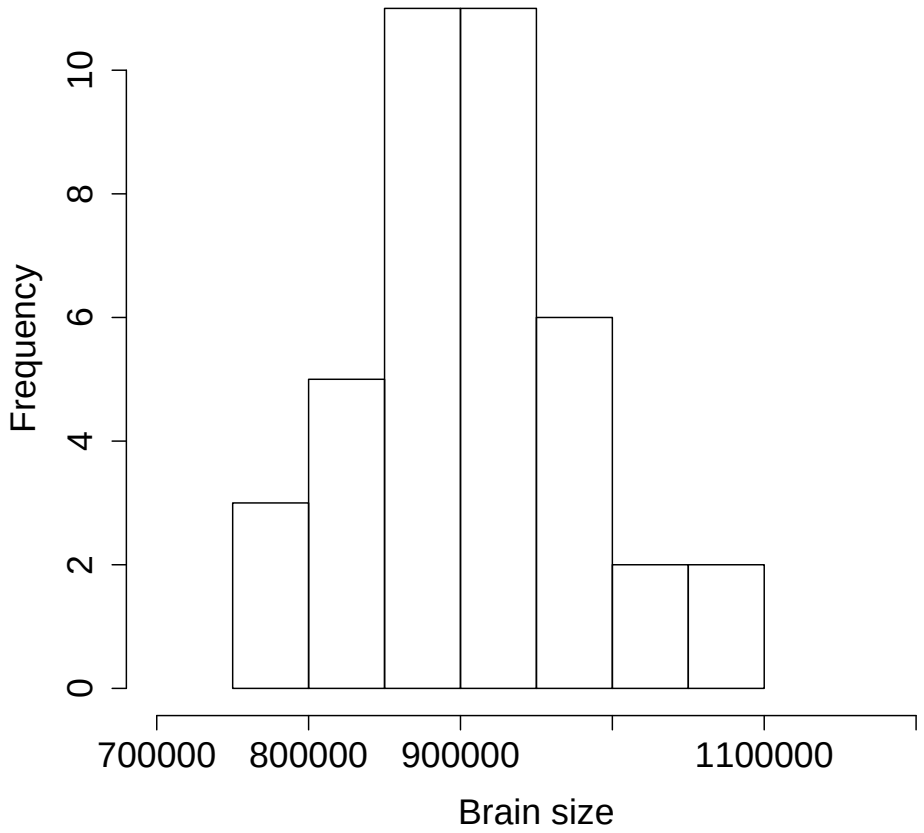
In previous lectures we have considered discrete datasets and discrete probability distributions. In practice many datasets that we collect from experiments consist of continuous measurements.

So we need to study probability models for continuous data.

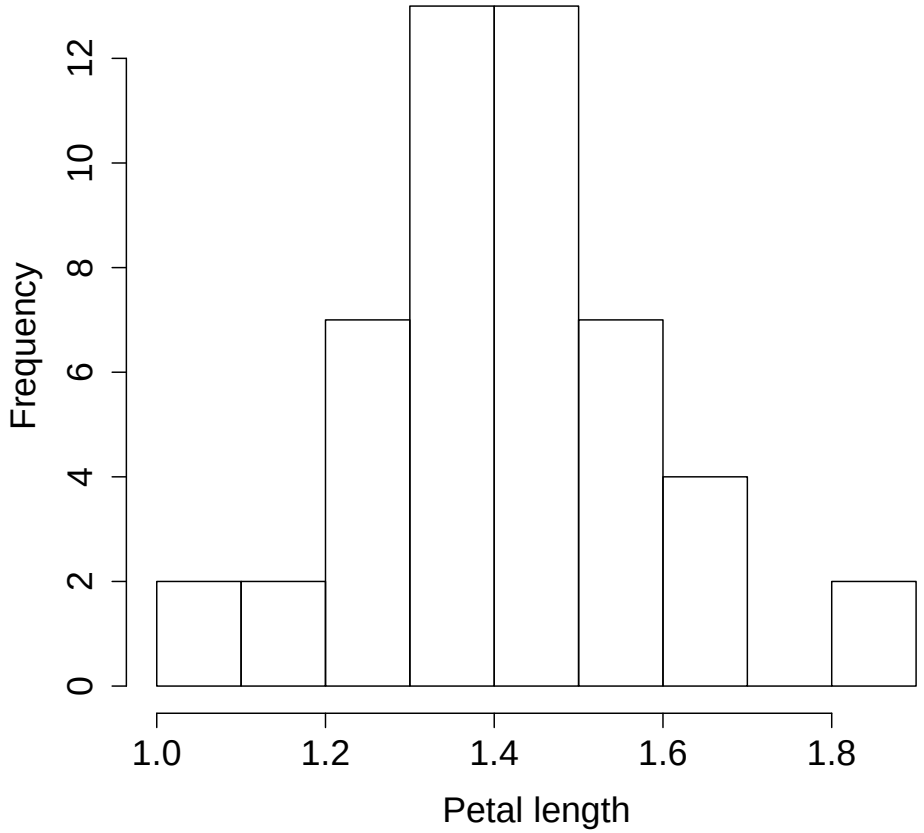
The birth weights of the babies in the Babyboom dataset



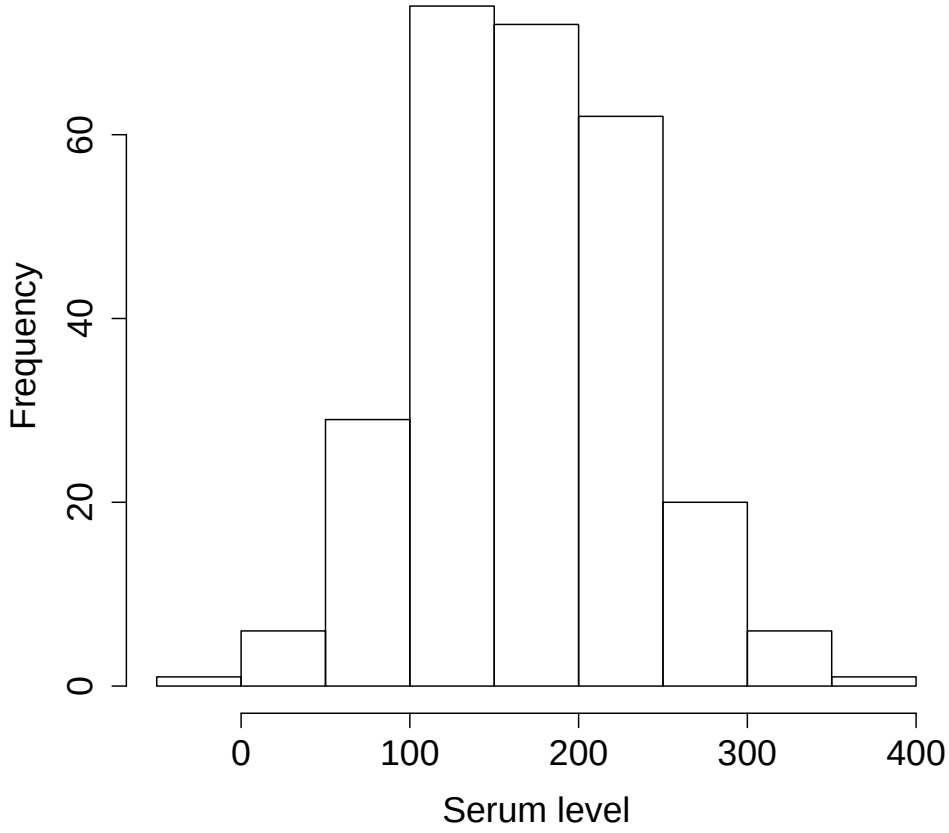
The brain sizes of 40 students



The petal length of a type of flower



Serum level measurements from healthy volunteers



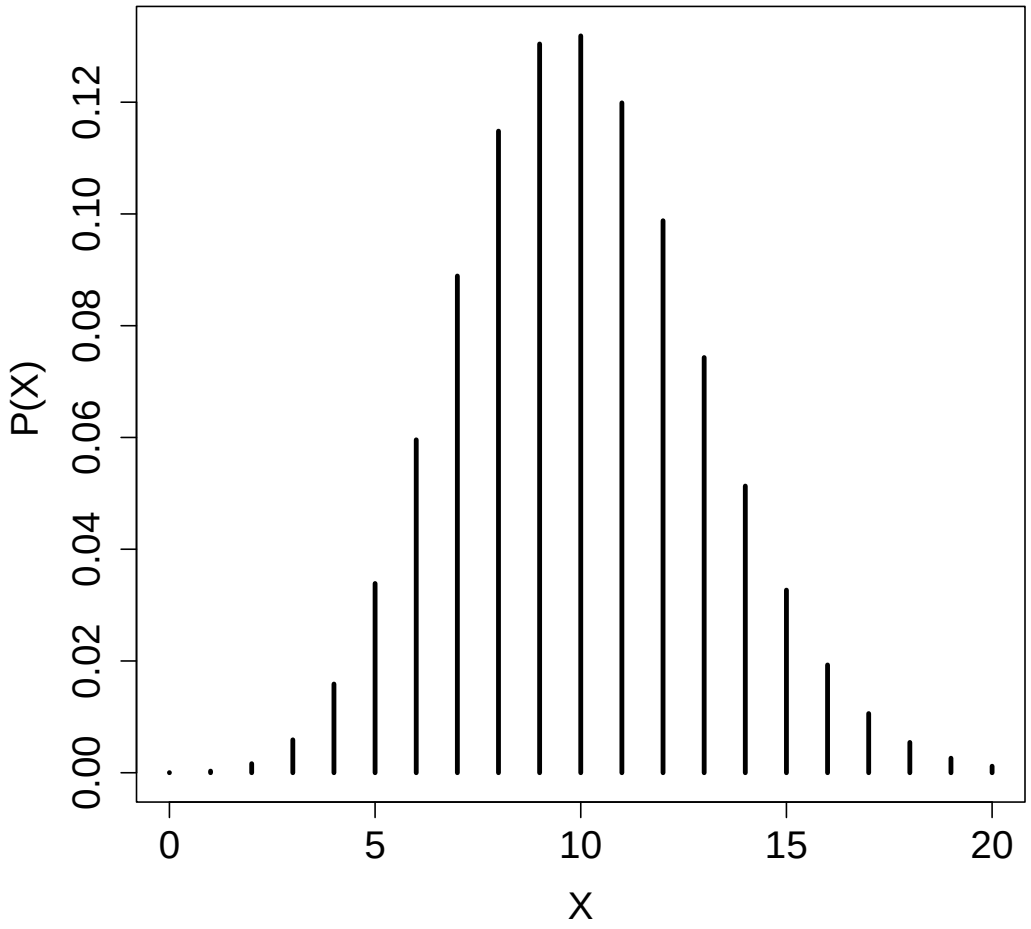
Continuous probability distributions

When we considered the Binomial and Poisson distributions we saw that the probability distributions were characterized by a formula for the probability of each possible discrete value.

All of the probabilities together sum up to 1.

We can visualize the density by plotting the probabilities against the discrete values.

A discrete probability distribution

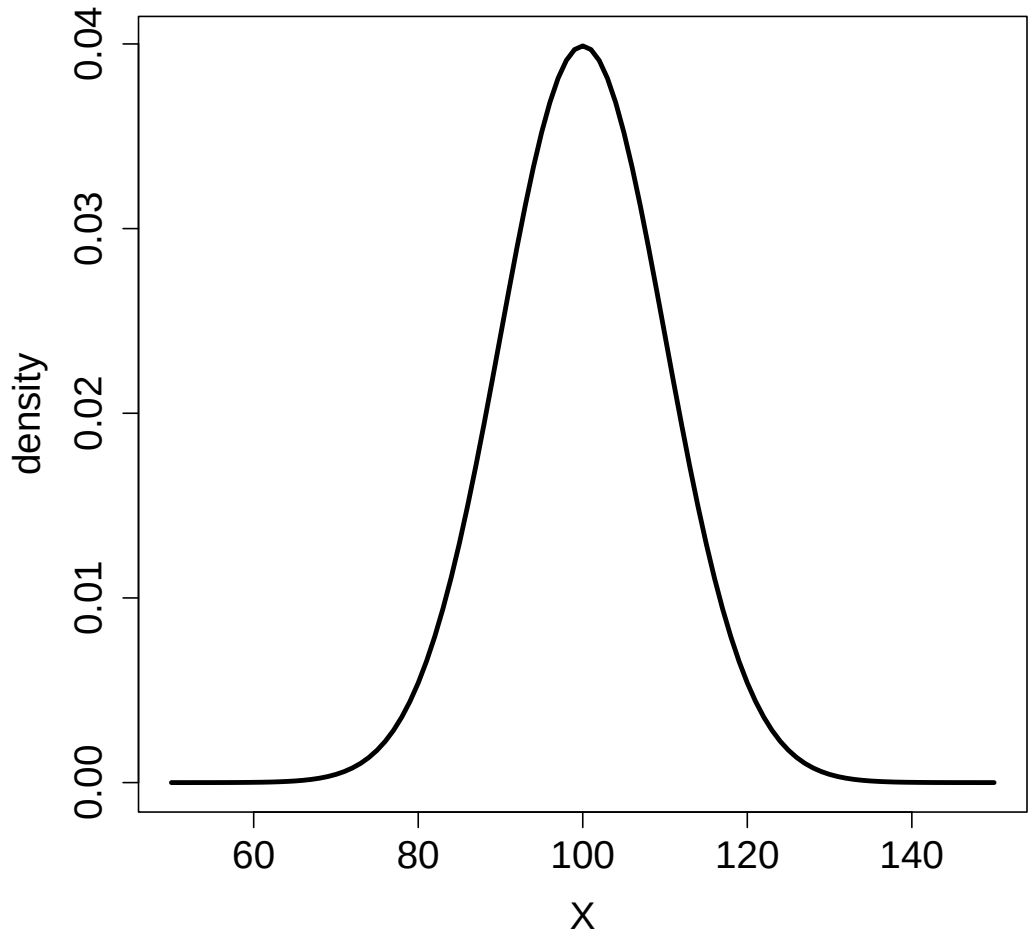


For continuous data we don’t have equally spaced discrete values so instead we use a curve or function that describes the probability *density* over the range of the distribution.

The curve is chosen so that the area under the curve is equal to 1.

If we observe a sample of data from such a distribution we should see that the values occur in regions where the density is highest.

A continuous probability distribution



The Normal Distribution

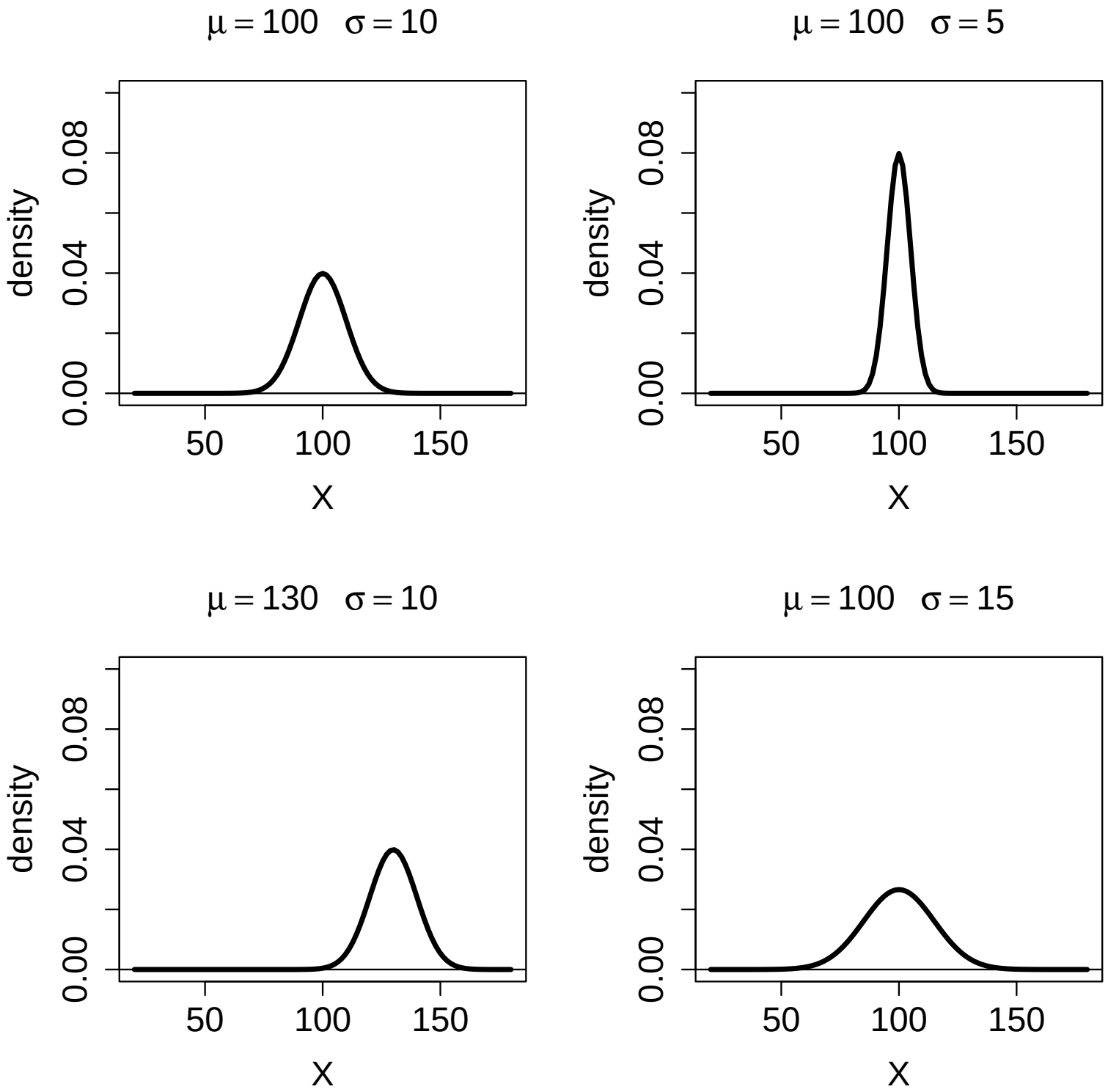
There will be many, many possible probability density functions over a continuous range of values.

The Normal distribution describes a special class of such distributions that are symmetric and can be described by two parameters

(i) μ = The mean of the distribution

(ii) σ = The standard deviation of the distribution

Changing the values of μ and σ alter the positions and shapes of the distributions.


μ = 100 σ = 10
density
0.00
0.04
0.08
50
100
150
X
μ = 100 σ = 5
density
0.00
0.04
0.08
50
100
150
X
μ = 130 σ = 10
density
0.00
0.04
0.08
50
100
150
X
μ = 100 σ = 15
density
0.00
0.04
0.08
50
100
150
X

If X is Normally distributed with mean μ and standard deviation σ, we write

$$\boxed{\mathrm{X} \sim \mathrm{N}(\mu, \sigma^2)}$$

μ and σ are the **parameters** of the distribution.

The probability density of the Normal distribution is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-(x-\mu)^2/2\sigma^2}$$

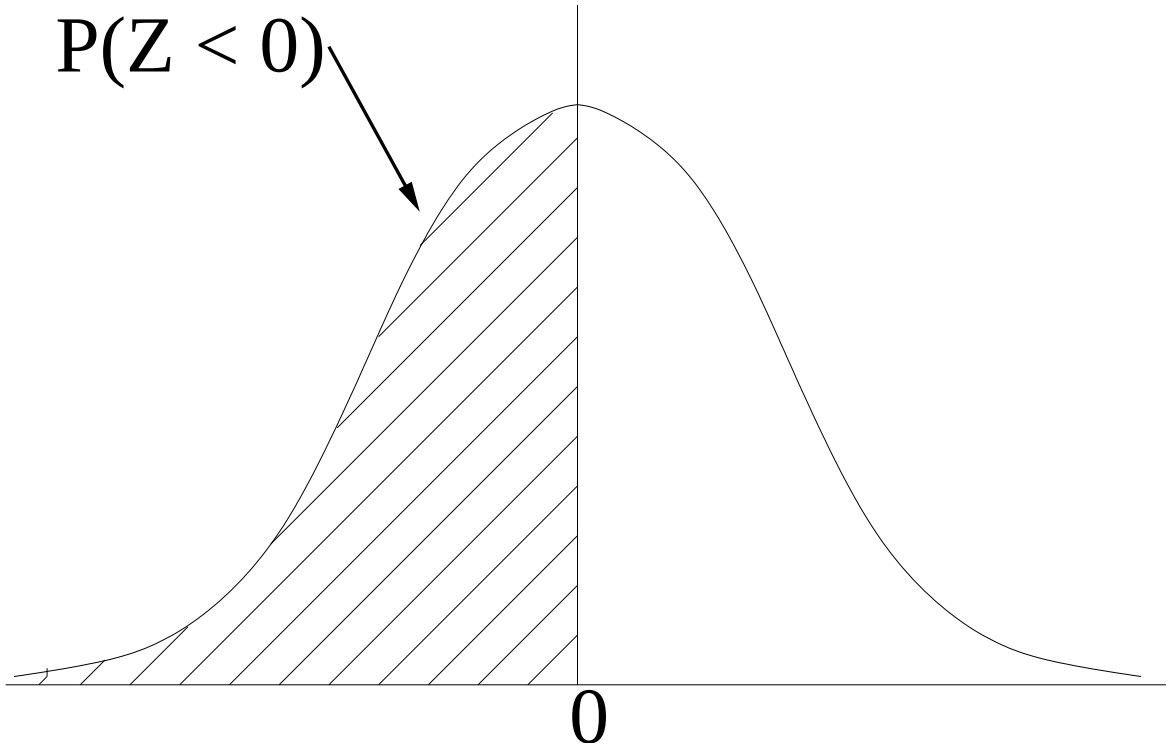
For the purposes of this course we do not need to use this expression. It is included here for future reference.

Calculating probabilities from the Normal distribution

For a discrete probability distribution we calculate the probability of being less than some value x, i.e. $P(X < x)$, by simply summing up the probabilities of the values less than x.

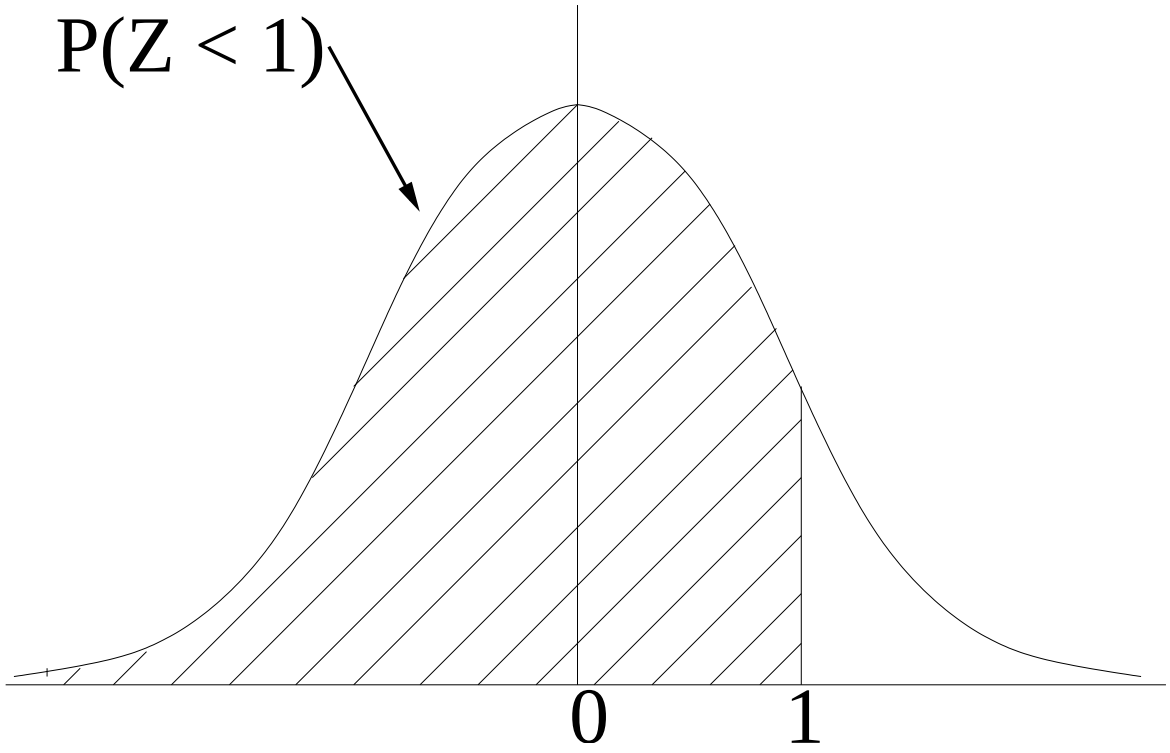
For a continuous probability distribution we calculate the probability of being less than some value x, i.e. $P(X < x)$, by calculating the area under the curve to the left of x.

Suppose $Z \sim$ N(0, 1), what is $P(Z < 0)$?



Symmetry $\Rightarrow P(Z < 0) = 0.5$

What about $P(Z < 1.0)$?

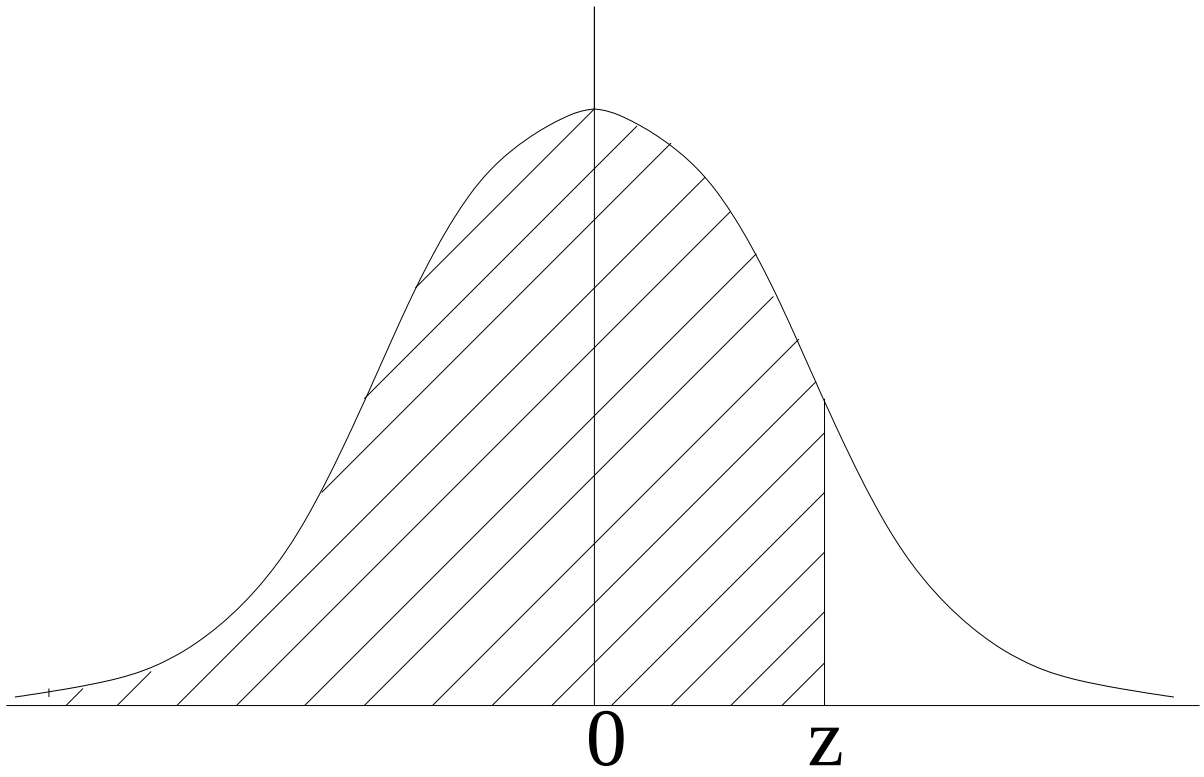


Calculating this area is not easy and so we use probability tables. Probability tables are tables of probabilities that have been calculated on a computer. All we have to do is identify the right probability in the table and copy it down!

Only one special Normal distribution, N(0, 1), has been tabulated.

The N(0, 1) distribution is called the **standard Normal distribution**.

The tables allow us to read off probabilities of the form $P(Z < z)$.

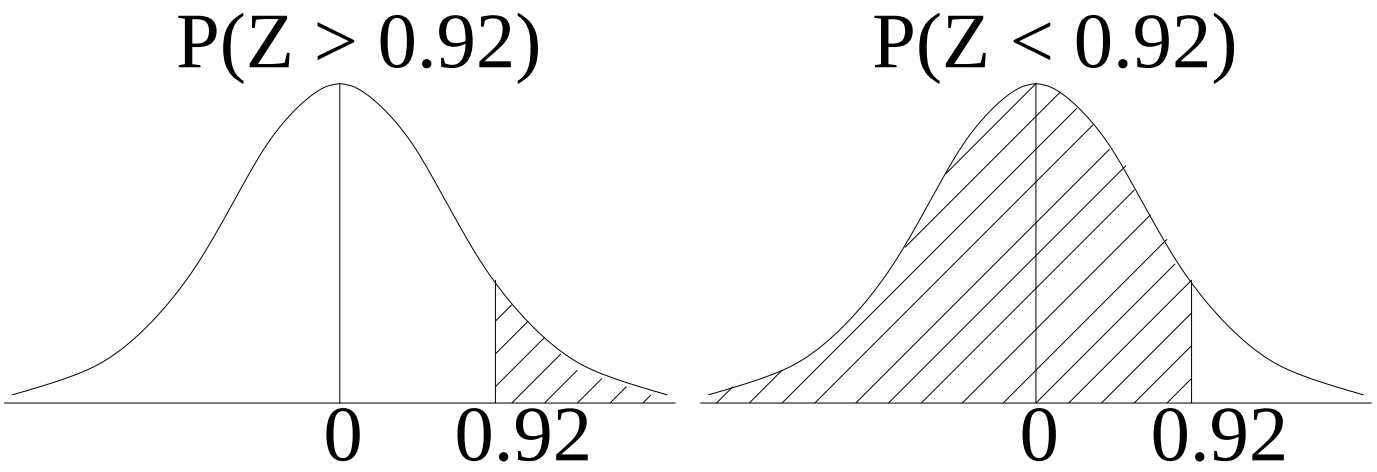


z	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	5040	5080	5120	5160	5199	5239	5279	5319	5359
0.1	0.5398	5438	5478	5517	5557	5596	5636	5675	5714	5753
0.2	0.5793	5832	5871	5910	5948	5987	6026	6064	6103	6141
0.3	0.6179	6217	6255	6293	6331	6368	6406	6443	6480	6517
0.4	0.6554	6591	6628	6664	6700	6736	6772	6808	6844	6879
0.5	0.6915	6950	6985	7019	7054	7088	7123	7157	7190	7224
0.6	0.7257	7291	7324	7357	7389	7422	7454	7486	7517	7549
0.7	0.7580	7611	7642	7673	7704	7734	7764	7794	7823	7852
0.8	0.7881	7910	7939	7967	7995	8023	8051	8078	8106	8133
0.9	0.8159	8186	8212	8238	8264	8289	8315	8340	8365	8389
1.0	0.8413	8438	8461	8485	8508	8531	8554	8577	8599	8621
1.1	0.8643	8665	8686	8708	8729	8749	8770	8790	8810	8830

From this table we can identify that $P(Z < 1.0) = 0.8413$

Example 1

If $Z \sim$ N(0, 1) what is $P(Z > 0.92)$?

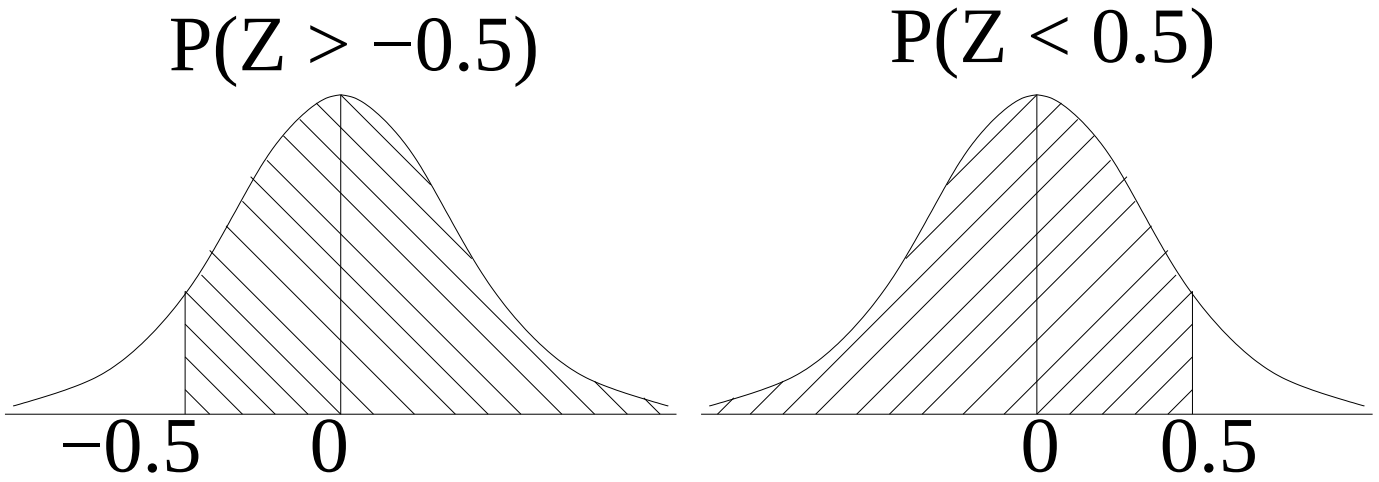


We know that $P(Z > 0.92) = 1 - P(Z < 0.92)$ and we can calculate $P(Z < 0.92)$ from the tables.

Thus, $P(Z > 0.92) = 1 - 0.8212 = 0.1788$

Example 2

If $Z \sim$ N(0, 1) what is $P(Z > -0.5)$?

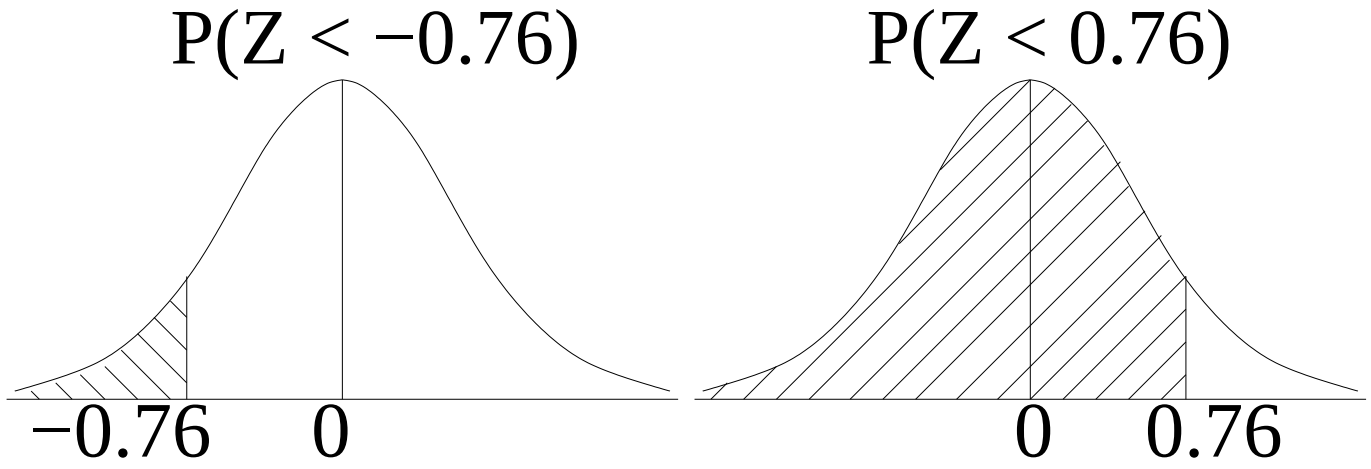


The Normal distribution is symmetric so we know that

$P(Z > -0.5) = P(Z < 0.5) = 0.6915$

Example 3

If $Z \sim$ N(0, 1) what is $P(Z < -0.76)$?

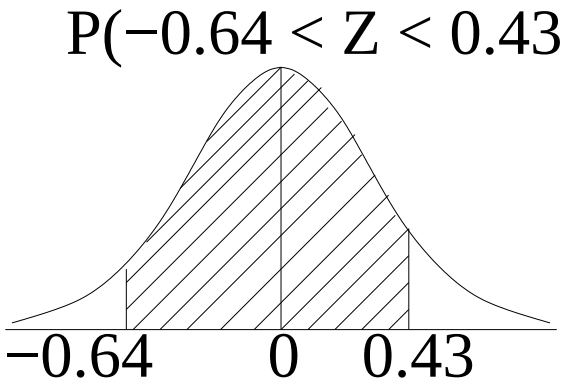


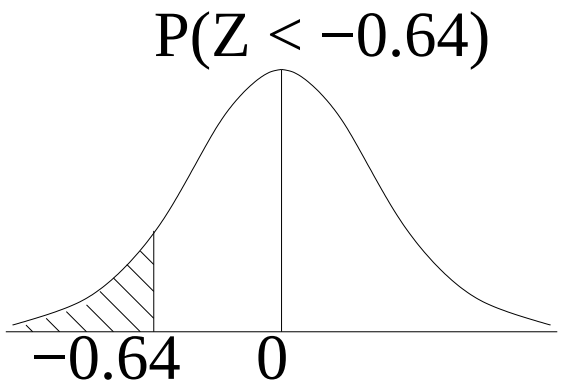
By symmetry

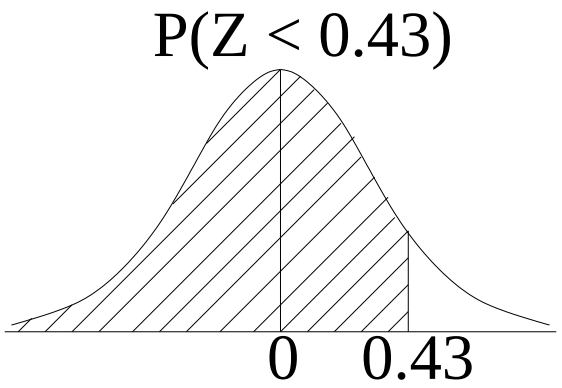
$$\begin{aligned} P(Z < -0.76) = P(Z > 0.76) &= 1 - P(Z < 0.76) \\ &= 1 - 0.7764 \\ &= 0.2236 \end{aligned}$$

Example 4

If $Z \sim$ N(0, 1) what is $P(-0.64 < Z < 0.43)$?







We can calculate this probability as

$$\begin{aligned} P(-0.64 < Z < 0.43) &= P(Z < 0.43) - P(Z < -0.64) \\ &= 0.6664 - (1 - 0.7389) \\ &= 0.4053 \end{aligned}$$

Example 5

Consider $P(Z < 0.567)$?

From tables we know that $P(Z < 0.56) = 0.7123$ and $P(Z < 0.57) = 0.7157$
To calculate $P(Z < 0.567)$ we *interpolate* between these two values

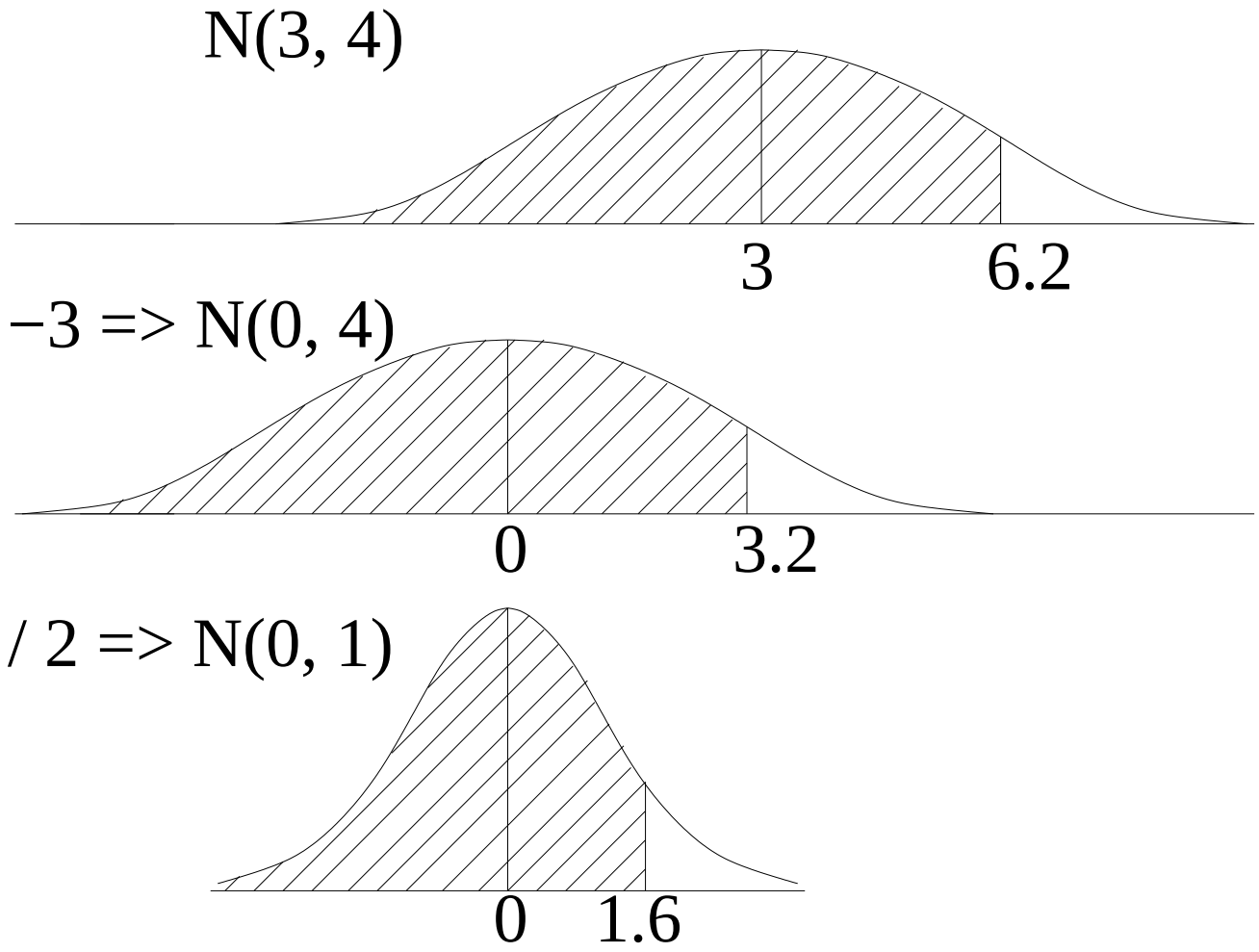
$$P(Z < 0.567) = 0.3 \times 0.7123 + 0.7 \times 0.7157 = 0.71468$$

Standardization

All of the probabilities above were calculated for the standard Normal distribution N(0, 1). If we want to calculate probabilities from different Normal distributions we convert the probability to one involving the standard Normal distribution.

This process is called **standardization**.

Suppose $X \sim$ N(3, 4), what is $P(X < 6.2)$?



We convert this probability to one involving the N(0, 1) distribution by

(i) Subtracting the mean μ

(ii) Dividing by the standard deviation σ

Subtracting the mean re-centers the distribution on zero. Dividing by the standard deviation re-scales the distribution so it has standard deviation 1. If we also transform the boundary point of the area we wish to calculate we obtain the equivalent boundary point for the N(0, 1) distribution.

$\Rightarrow P(X < 6.2) = P(Z < 1.6) = 0.9452$ where $Z \sim$ N(0, 1)

This process can be described by the following rule

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

$$Z \sim \text{N}(0,\ 1)$$

Example 6

Suppose we know that the birth weight of babies is Normally distributed with mean 3500g and standard deviation 500g. What is the probability that a baby is born that weighs less than 3100g?

That is $X \sim$ N(3500, 500^2) and we want to calculate $P(X < 3100)$?

We can calculate the probability through the process of standardization.

Drawing a rough diagram helps

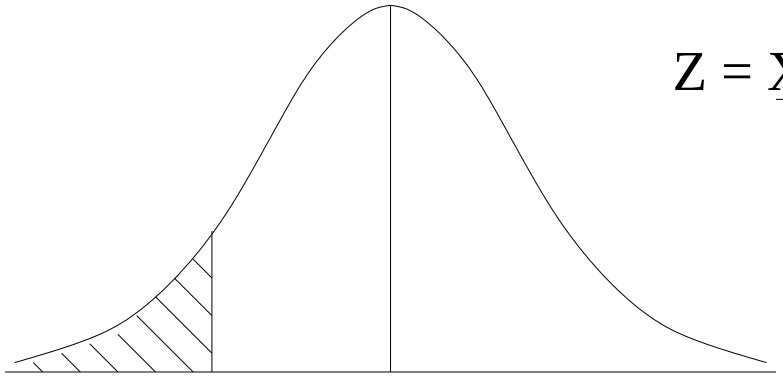
$X \sim N(3500, 500^2)$

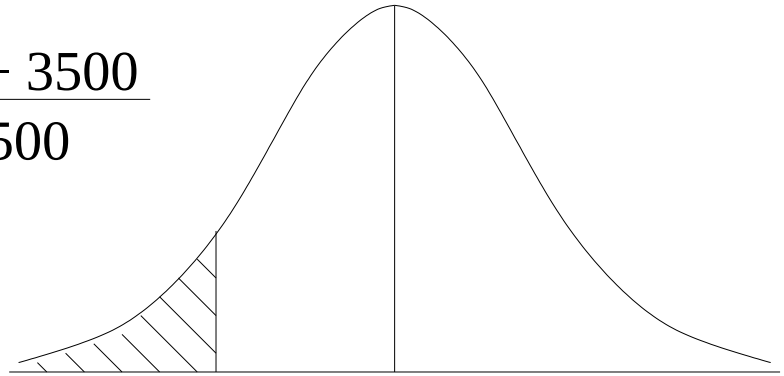
$Z \sim N(0, 1)$

$P(X < 3100)$

$P(Z < -0.8)$

$$Z = \frac{X - 3500}{500}$$





$$
\begin{aligned}
P(X < 3100) &= P\left(\frac{X - 3500}{500} < \frac{3100 - 3500}{500}\right) \\
&= P(Z < -0.8) \qquad \text{where } Z \sim \mathbf{N}(0, 1) \\
&= 1 - P(Z < 0.8) \\
&= 1 - 0.7881 \\
&= 0.2119
\end{aligned}
$$

Linear combinations of Normal random variables

Suppose two rats A and B have been trained to navigate a large maze.

X = Time of run for rat A $\quad X \sim \text{N}(80, 10^2)$

Y = Time of run for rat B $\quad Y \sim \text{N}(78, 13^2)$

On any given day what is the probability that rat A runs the maze faster than rat B?

Let $D = X - Y$ be the difference in times of rats A and B

If rat A is faster than rat B then $D < 0$ so we want $P(D < 0)$?

To calculate this probability we need to know the distribution of D. To do this we use the following rule

If X and Y are two independent normal variable such that

$$X \sim \text{N}(\mu_1, \sigma_1^2) \text{ and } Y \sim \text{N}(\mu_2, \sigma_2^2)$$

then $$X - Y \sim \text{N}(\mu_1 \text{ - } \mu_2, \sigma_1^2 + \sigma_2^2)$$

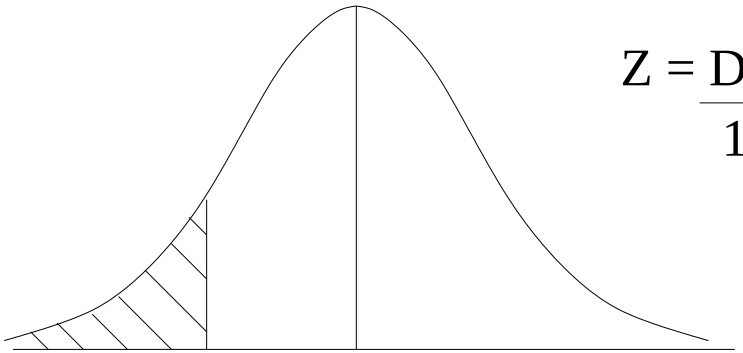
In this example,

$$D = X - Y \sim \text{N}(80 - 78, 10^2 + 13^2) = N(2, 269)$$

We can now calculate this probability through standardization

$D \sim N(2, 269)$

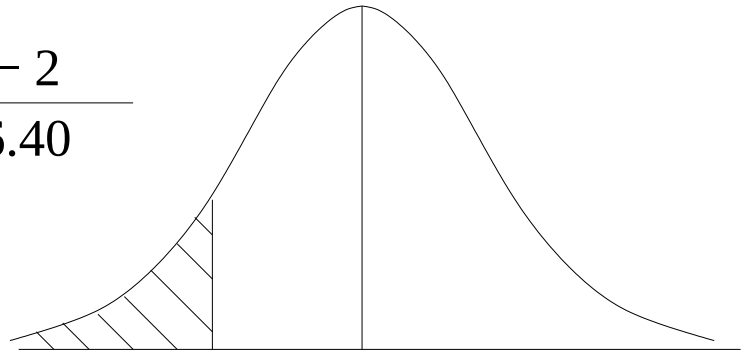
$P(D < 0)$



$$Z = \frac{D - 2}{16.40}$$

$Z \sim N(0, 1)$

$P(Z < -0.122)$



$$\frac{0 - 2}{16.40} = -0.122$$

0

$$
\begin{aligned}
P(D < 0) &= P\left(\frac{D-2}{\sqrt{269}} < \frac{0-2}{\sqrt{269}}\right) \\
&= P(Z < -0.122) \qquad\qquad Z \sim N(0,1) \\
&= 1 - (0.2 \times 0.5478 + 0.8 \times 0.5517) \\
&= 0.45142
\end{aligned}
$$

Other rules that are often used are

If X and Y are two independent normal variables such that

$$X \sim \mathrm{N}(\mu_1, \sigma_1^2) \text{ and } Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$$

then

$$\begin{aligned} X + Y &\sim \mathrm{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \\ aX &\sim \mathrm{N}(a\mu_1, a^2\sigma_1^2) \\ aX + bY &\sim \mathrm{N}(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2) \end{aligned}$$

Using the Normal tables backwards

The marks of 500 candidates in an examination are normally distributed with a mean of 45 marks and a standard deviation of 20 marks.

If 20% of candidates obtain a distinction by scoring x marks or more, estimate the value of x.

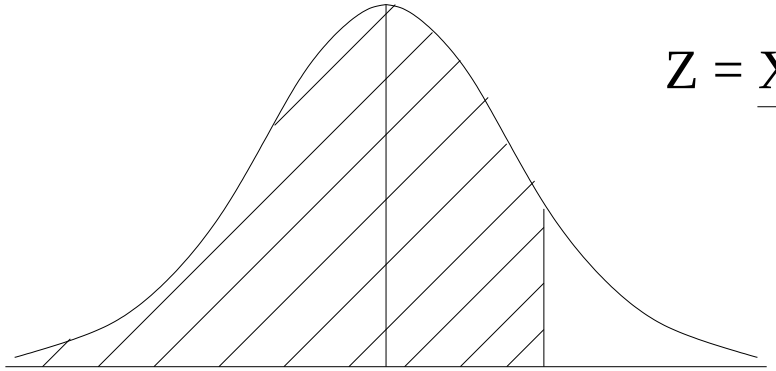
We have $X \sim$ N(45, 20^2) and we want x such that $P(X > x) = 0.2$

$\Rightarrow P(X < x) = 0.8$

$X \sim N(45, 400)$

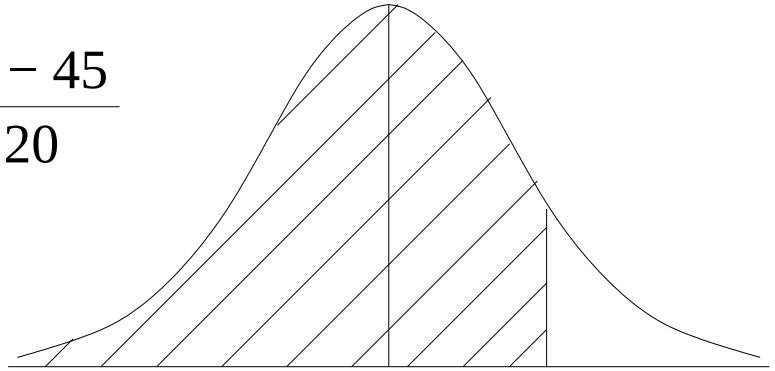
$Z \sim N(0, 1)$

$P(X < x) = 0.8$



$$Z = \frac{X - 45}{20}$$

$P(Z < 0.84) = 0.8$



$$\frac{x - 45}{20} = 0.84$$

Standardizing this probability we get

$$P\left(\frac{X-45}{20} < \frac{x-45}{20}\right) = 0.8$$

$$\Rightarrow P\left(Z < \frac{x-45}{20}\right) = 0.8$$

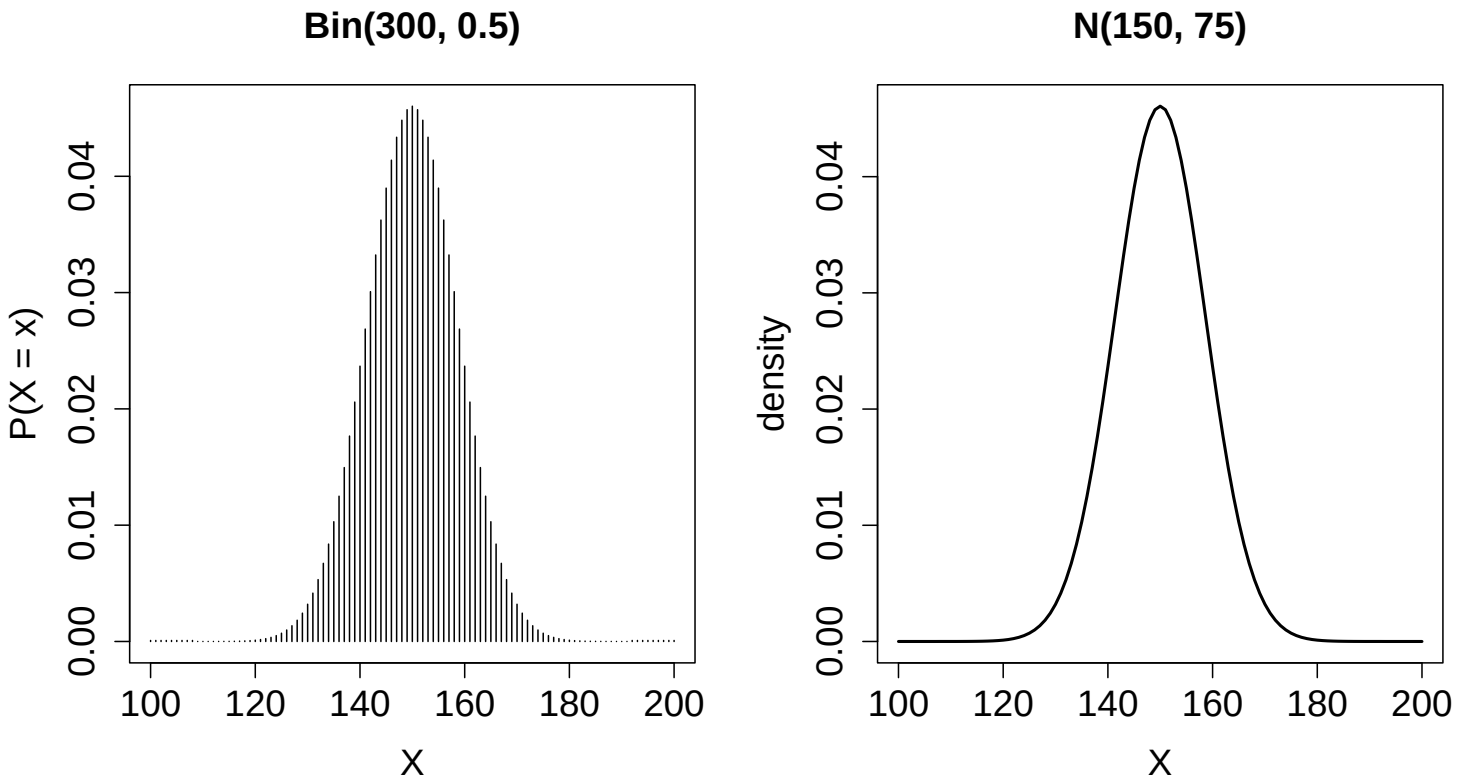
From the tables we know that $P(Z < 0.84) \approx 0.8$ so

$$\frac{x-45}{20} \approx 0.84$$

$$\Rightarrow \quad x \approx 45 + 20 \times 0.84 = 61.8$$

The Normal approximation to the Binomial

Under certain conditions we can use the Normal distribution to approximate the Binomial distribution.



In general

If $X \sim$ Bin(n, p) then

$$\mu = np$$
$$\sigma^2 = npq \quad \text{where} \quad q = 1 - p$$

For large n and p not too small or too large

$$X \sim \text{N}(np, npq)$$

$n > 10$ and $p \approx \frac{1}{2}$ OR $n > 30$ and p moving away from $\frac{1}{2}$

Example

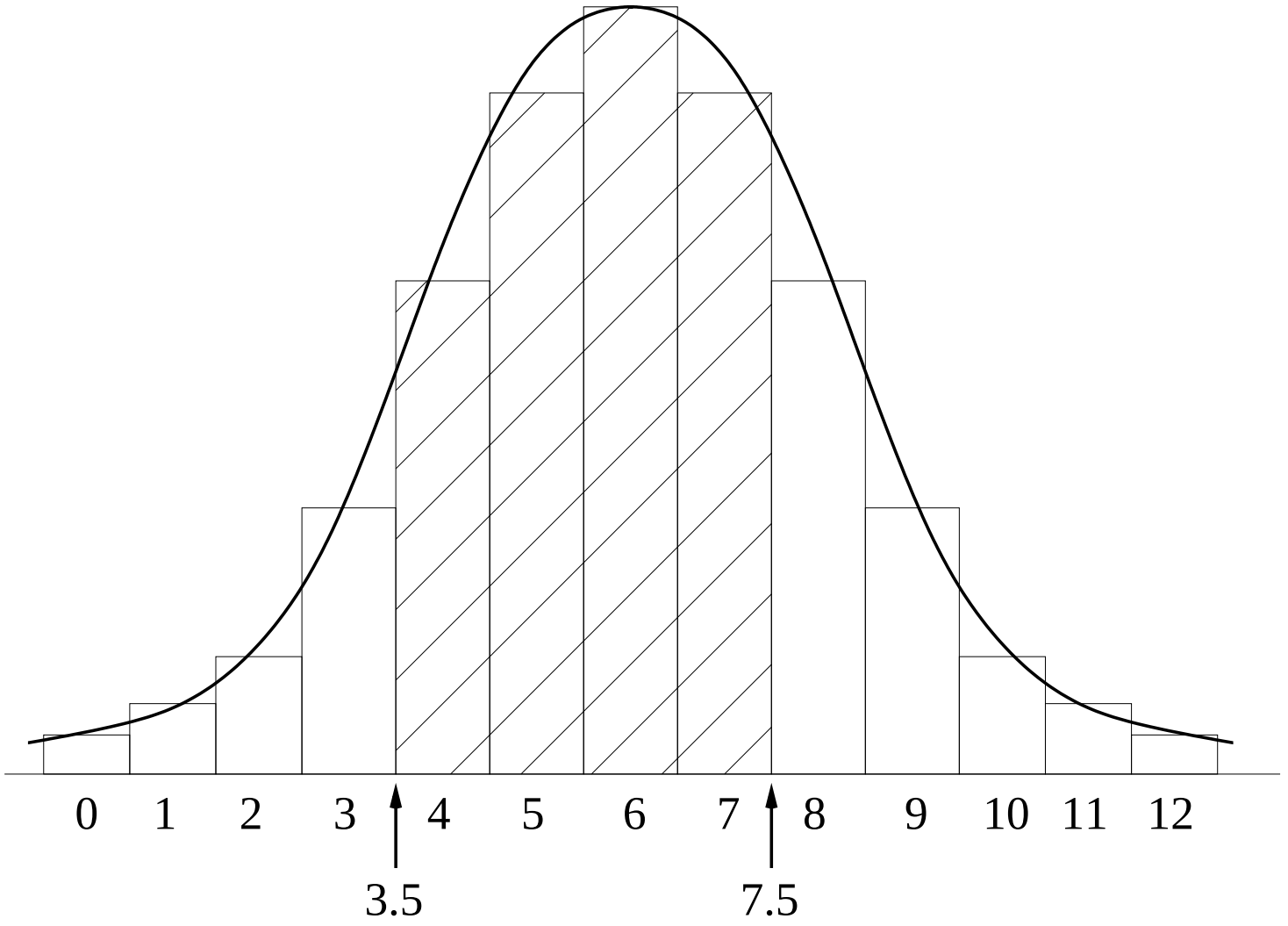
Suppose $X \sim$Bin(12, 0.5) what is $P(4 \leq X \leq 7)$?

For this distribution we have

$$\mu = np = 6$$
$$\sigma^2 = npq = 3$$

So we can use a N(6, 3) distribution as an approximation.

Unfortunately, it's not quite so simple. We have to take into account the fact that we are using a *continuous* distribution to approximate a *discrete* distribution. This is done using a **continuity correction**.


0
1
2
3
4
5
6
7
8
9
10
11
12
3.5
7.5

$P(4 \leq X \leq 7)$ transforms to $P(3.5 < X < 7.5)$

$$\begin{aligned} P(3.5 < X < 7.5) &= P\left(\frac{3.5-6}{\sqrt{3}} < \frac{X-6}{\sqrt{3}} < \frac{7.5-6}{\sqrt{3}}\right) \\ &= P(-1.443 < Z < 0.866) \text{ where } Z \sim \mathbf{N}(0,1) \\ &= 0.732 \end{aligned}$$

The exact answer is 0.733 so in this case the approximation is very good.

The Normal approximation to the Poisson

We can also use the Normal distribution to approximate a Poisson distribution under certain conditions.

In general,

If $X \sim \text{Po}(\lambda)$ then

$$\mu = \lambda$$
$$\sigma^2 = \lambda$$

For large λ (say $\lambda > 20$)

$$X \sim \text{N}(\lambda, \lambda)$$

Example

A radioactive source emits particles at an average rate of 25 particles per second. What is the probability that in 1 second the count is less than 27 particles?

X = No. of particles emitted in 1s $\quad X \sim$ Po(25)

So, we can use a N(25, 25) as an approximate distribution.

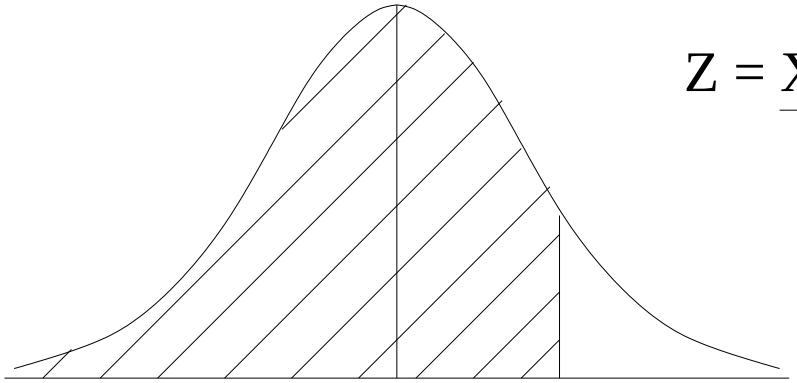
Again, we need to make a continuity correction

So $P(X < 27)$ transforms to $P(X < 26.5)$

$X \sim N(25, 25)$

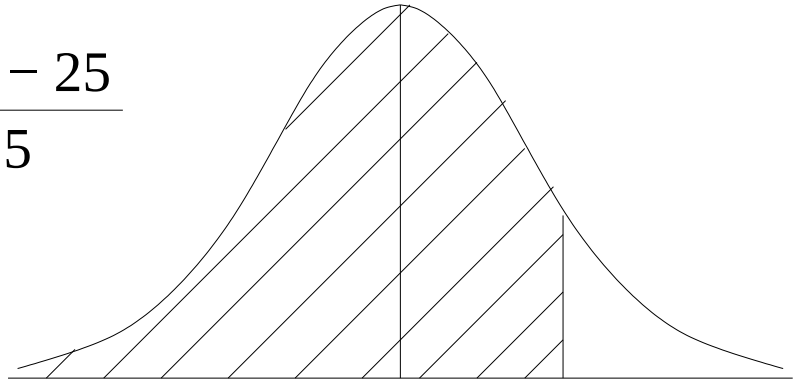
$Z \sim N(0, 1)$

$P(X < 26.5)$



$$Z = \frac{X - 25}{5}$$

$P(Z < 0.3)$



$$
\begin{aligned}
P(X < 26.5) &= P\left(\frac{X-25}{5} < \frac{26.5-25}{5}\right) \\
&= P(Z < 0.3) \qquad \text{where } Z \sim \mathrm{N}(0,1) \\
&= 0.6179
\end{aligned}
$$