The Normal Distribution

The Normal Distribution is the most important of the continuous probability distributions. It arises as the limit of the Binomial Distribution as the number of trials becomes greater and greater. To illustrate this idea, consider tossing a number of coins several times. Suppose we toss four coins a total of 8 times. Pascal's triangle, which gives the values of the Binomial coefficient, indicates the expected frequencies of 0, 1, 2, 3 and 4 heads:

1

1 1

1 2 1

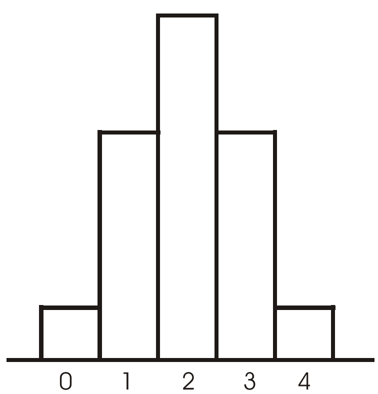
1 3 3 1

1 4 6 4 1

The probabilities are calculated from these. For example, for $X \sim B\left(4, \frac{1}{2}\right)$, the probability distribution is:

x_i	4	3	2	1	0
$P(X = x_i)$	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{16}$

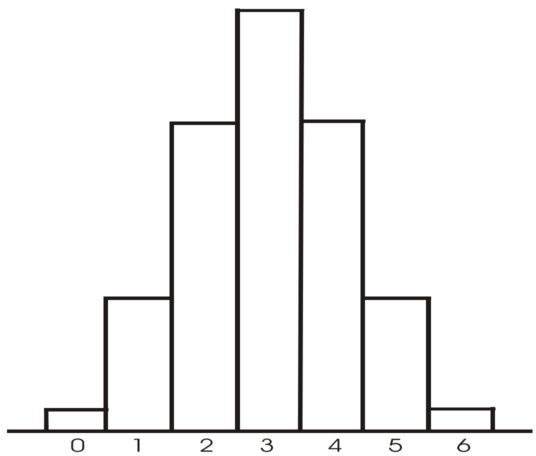
We can represent this discrete probability distribution using a histogram:



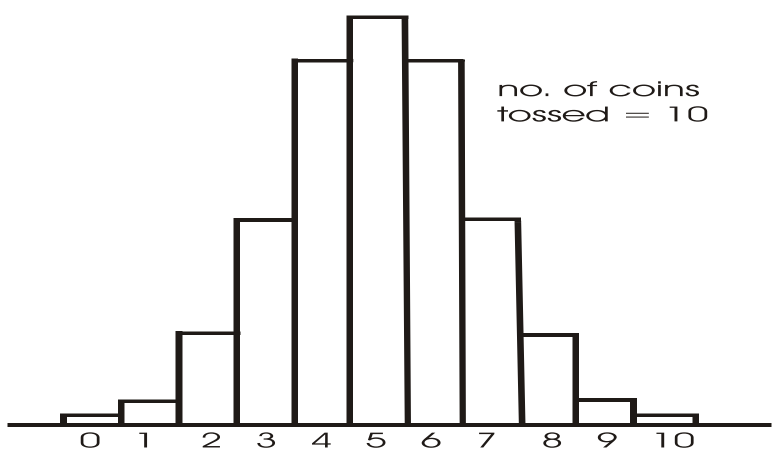
Tossing the coin 64 times would give the following frequencies and probabilities:

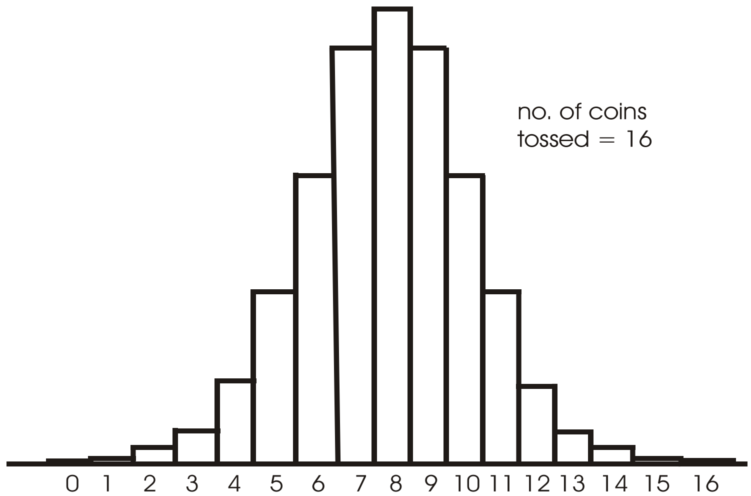
No. of heads	0	1	2	3	4	5	6
Frequency	1	6	15	20	15	6	1
Probability	$\frac{1}{64}$	$\frac{6}{64}$	$\frac{15}{64}$	$\frac{20}{64}$	$\frac{15}{64}$	$\frac{6}{64}$	$\frac{1}{64}$

With histogram:



If we also visualise this process as involving a shrinking of the width of each rectangle, as the number of coins tossed increases, the shape gets closer and closer to being smooth in appearance.





Eventually, the histogram will appear completely smooth, and if we allow the number of coins tossed to become infinite, then the curve will be continuous.

It is this continuous distribution that is called the Normal Distribution.

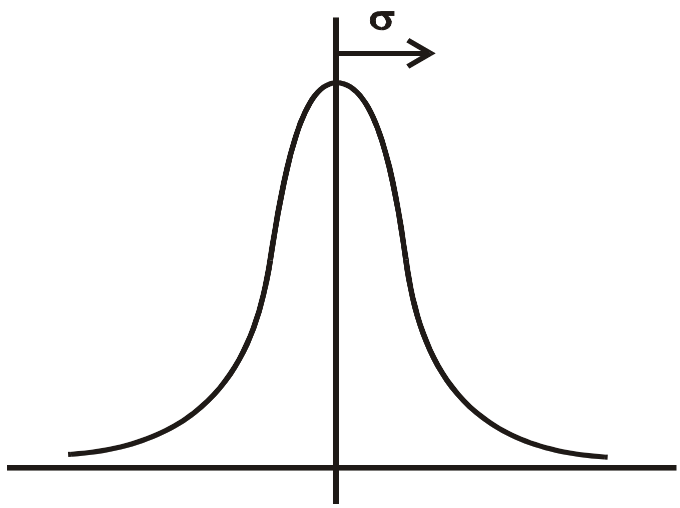
This distribution is of importance in natural science, because it is expected that many properties of nature, if sampled randomly, would give distributions that were "normal" in appearance. For example, the distribution of adult male height would, if sampled, be expected to produce a histogram that corresponded to the normal curve.

This distribution is "bell" shaped, and is symmetric about its middle value – which is the population mean and the expected mean of the sample. Also, the curve is such that 95%

of the area under it lies within 2 standard deviations of the mean. If X is a variable that is believed to be normally distributed, with mean μ and standard deviation σ then we write:

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

The symbol σ^2 denotes the variance of the distribution, which is the square of the standard deviation σ.



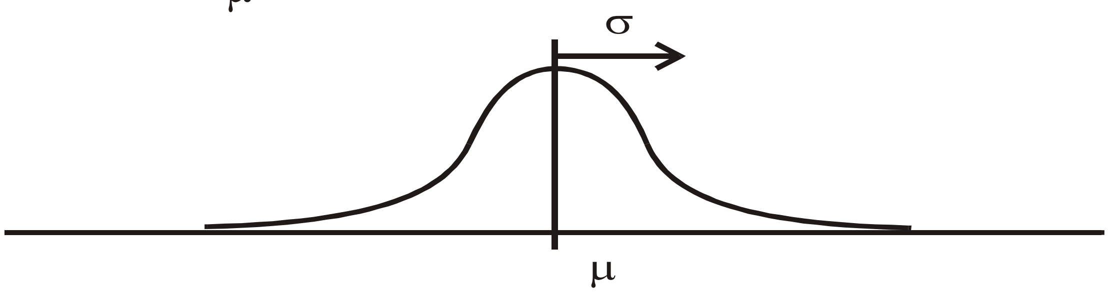
The mathematician Gauss derived an expression for the function that gives the probability, $p(x)$, of a variable X taking a value x, when X is normally distributed with mean μ and variance σ^2.

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2} / 2\sigma^2 \qquad \text{for } -\infty < x < \infty$$

Values of this function for a standardised variable with $\mu = 0$ and $\sigma = 1$ are given in tables, and the function $p(x)$ is never used directly at this level.

Standardised Normal Variable

One normal distribution differs from the other only in the position of the mean, which acts as an axis of symmetry, and in the size of the standard deviation, which measures the extent to which the data is spread out around the mean. Otherwise, the shape of the curve representing the normal distribution remains the same.



Consequently, it makes sense to represent the normal distribution by a single standardised distribution such that:

$\mu = 0$
$\sigma = 1$

The number of standard deviations a given value is away from the mean is denoted by *z*.

Thus, any normal distribution can be derived from the standard distribution and vice-versa by:

(1) A translation of the mean from its given value to O

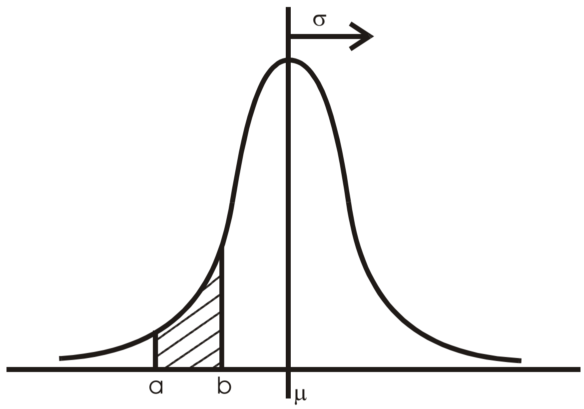
(2) A scaling of the standard deviation from its given value to 1.

Questions on the Normal distribution generally require the calculation of a probability that the value of the variable *X* lies between two values *a* and *b*. That is:

Given $X \sim N\left(\mu, \sigma^2\right)$, find $P\left(a < X < b\right)$.

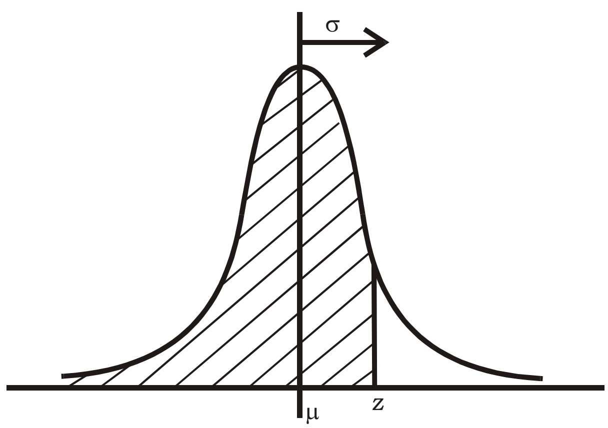
(Given X is a normally distributed variable such that it has mean μ and variance σ^2, find the probability that X takes values between a and b.)

A pictorial representation shows this to be an area under the curve representing the distribution:

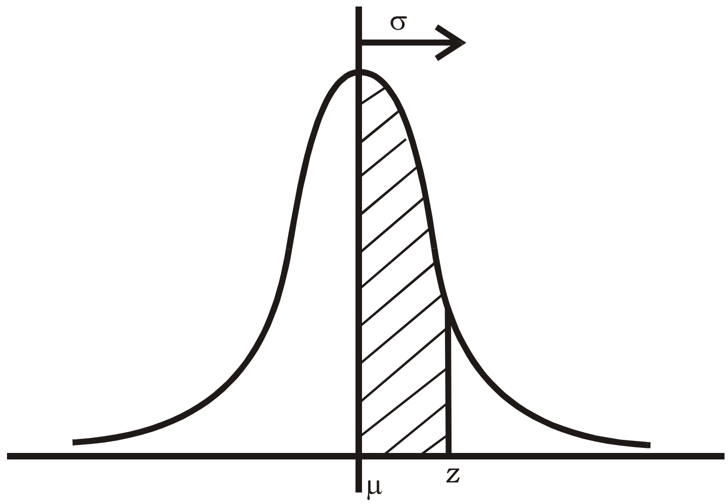


This area represents the probability that X lies between a and b.

Tables give the area under the curve for the Normal Distribution in the standardised form. This can be in several formats. Two of the most popular are



This first format gives the area under the curve up to a given value z, taken from $-\infty$ to z. It is $P(Z < z)$. (In this expression Z (upper case) is the variable, and z (small case) is the value that the variable can take.)



This second format gives the area under the curve up to the given value *z* from *0* to *z*. It is $P(0 < Z < z)$.

It is important to realise that the tables differ in this way and to use the values of *P(z)* accordingly. Here we shall use tables in the second form. Note, this means that *P(z)* is an abbreviation for $P(0 < Z < z)$ - an abbreviation for the probability that the value of *Z* lies between *0* and *z*.

To use the tables for the standardised variable, it is necessary to compute the *z* score corresponding to the values initially given.

That is:

Given $X \sim N(\mu, \sigma^2)$
find $P(X < x)$

We calculate the *z* score corresponding to *x*. The *z* score is the number of standard deviations *x* is away from the mean. It can take a positive or a negative direction.

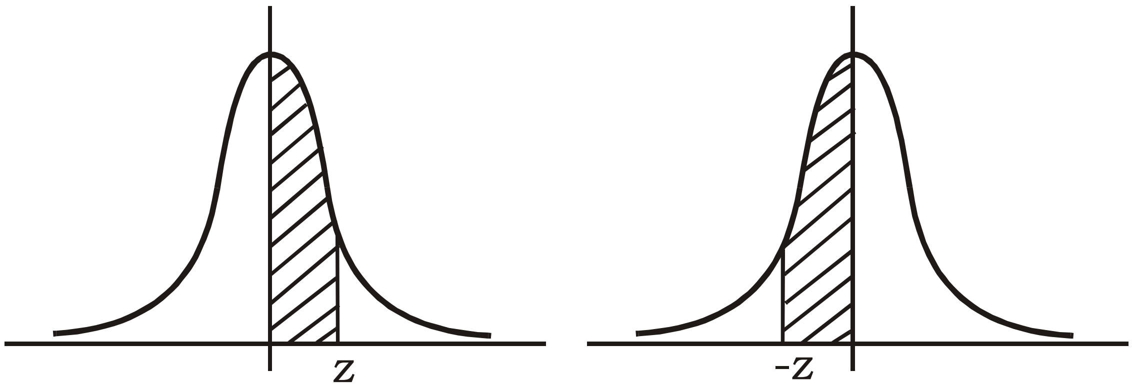
We calculate the *z* score corresponding to *x* by:

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

We find $P(Z < z)$ using the tables. Then,

$$P(X < x) = P(Z < z)$$

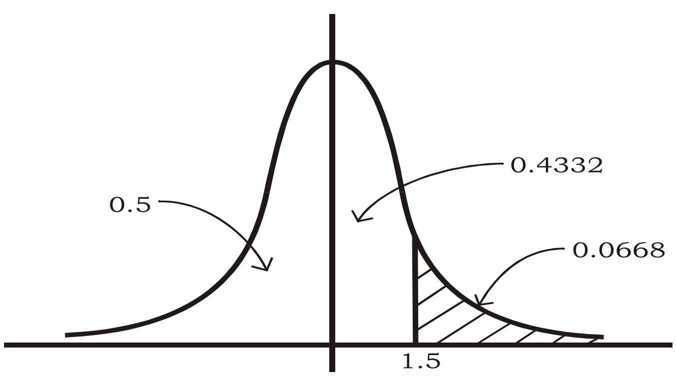
The *z* score can be positive or negative. The sign of *z* gives the direction from the mean, μ, of the value, and the magnitude of *z*, $|z|$ (the modulus of *z*), gives the number of standard deviations away from the mean that the value lies regardless of the direction.



Example (1)

Find $P(Z > 1.5)$

$$\begin{aligned} P(Z > 1.5) &= 1 - P(Z < 1.5) - 0.5 \\ &= 1 - 0.4332 - 0.5 \\ &= 0.0668 \end{aligned}$$

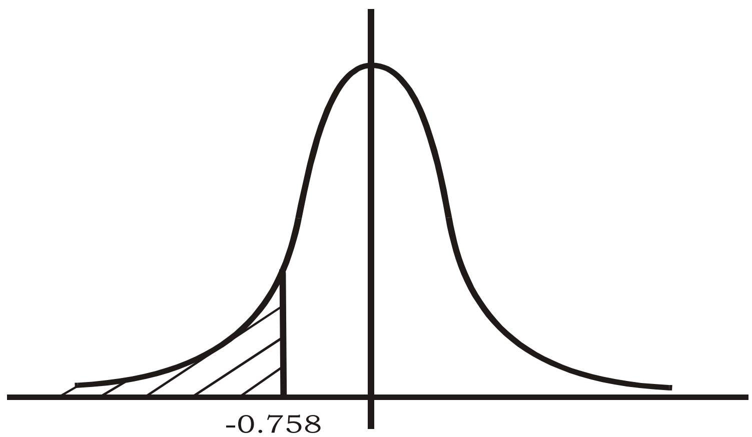


Example (2)

$X \sim N\left(4, 1.32^2\right)$, find $P\left(X < 3\right)$

The z value corresponding to $x = 3$ is:

$$
\begin{aligned}
z &= \frac{x-\mu}{\sigma} \\
&= \frac{3-4}{1.32} \\
&= -0.758 \text{ } (3.S.F.)
\end{aligned}
$$



$$
\begin{aligned}
P\left(X < 3\right) &= P\left(Z < -0.758\right) \\
&= 0.5 - P\left(Z < 0.758\right) \\
&= 0.5 - 0.2758 \\
&= 0.2242 \\
&= 0.224 \text{ (3.S.F.)}
\end{aligned}
$$

These examples illustrate that using tables it is possible:

(1) Given a z value, to find a probability corresponding to that value.
(2) Given an x value, and that $X \sim N\left(\mu, \sigma^2\right)$ to find a probability corresponding to that value.

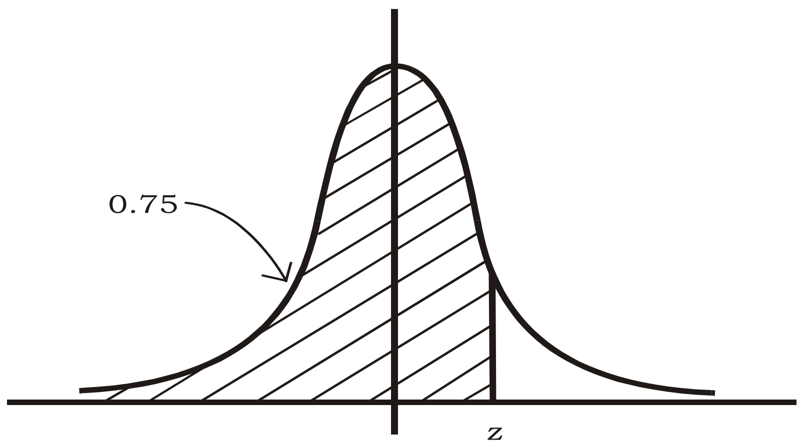
It is also possible:

(3) Given a probability, to find the corresponding z value.

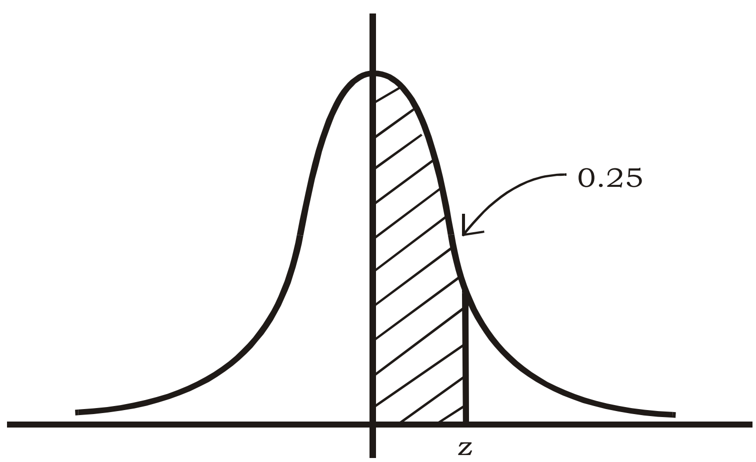
<u>Example (3)</u>

$X \sim N\left(6, 1.54^2\right)$. Find the value of a such that $P\left(X < a\right) = 0.75$.

Since $P\left(X < a\right) = 0.75$, this corresponds to $P\left(z\right) = 0.75$ as indicated in the following diagram:



This corresponds in our form of the tables to $P\left(z\right) = 0.25$



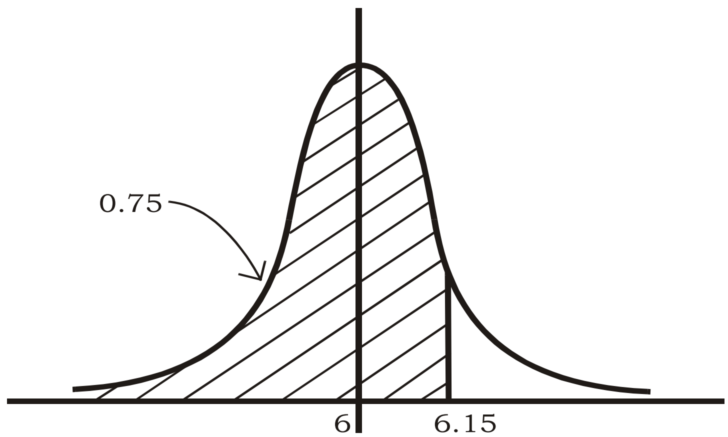
Then z = 0.0987

From the relationship

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

Then,

$$x = 6 + 1.54 \times 0.0987$$
$$= 6.15$$



Thus, in general, it is possible:

Using the relation $z = \frac{x - \mu}{\sigma}$

To find any one of the values, z, x, μ or σ given the other three; or given two simultaneous equations arising from this relation, to find two of the unknown values of z, x, μ or σ, given two others.

Example (4)

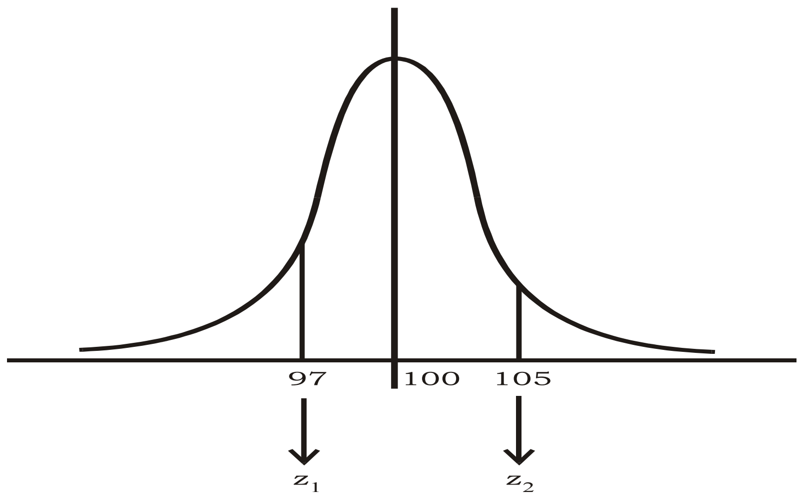
Two companies, Alpha and Beta, produce cricket balls. The diameters of the cricket balls are measured in millimetres.

(i) The diameters of the output of Alpha may be modelled by a normal distribution with mean 100*mm* and standard deviation 4*mm*. What is the probability that the diameter of a ball selected at random from Alpha's production line is between 97 and 105*mm*?

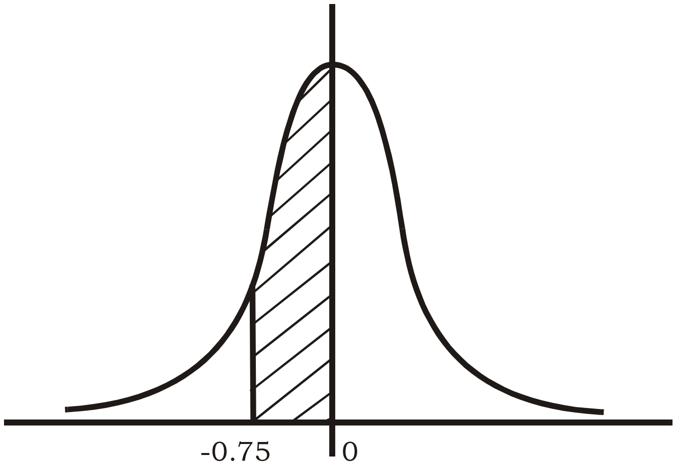
Solution

$X \sim N\left(100, 4^2\right)$

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

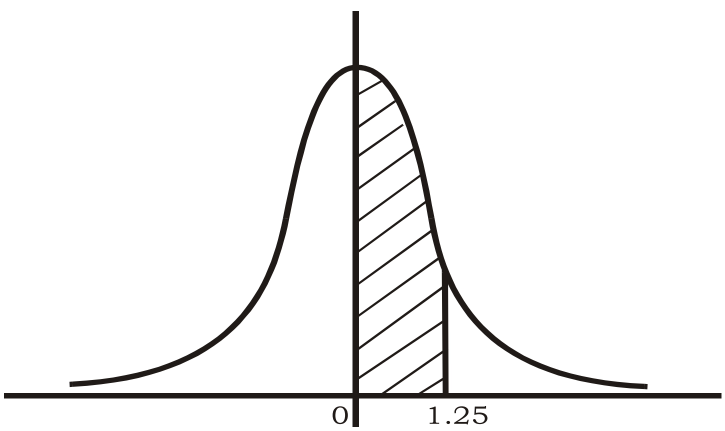


$$z_1 = \frac{97 - 100}{4} = -0.75$$



$$P(97 < X < 100) = P(-0.75 < Z < 0)$$
$$= 0.2734$$

$$z_2 = \frac{105-100}{4} = +1.25$$



$$P(100 < X < 105) = P(0 < Z < 1.25)$$
$$= 0.3944$$

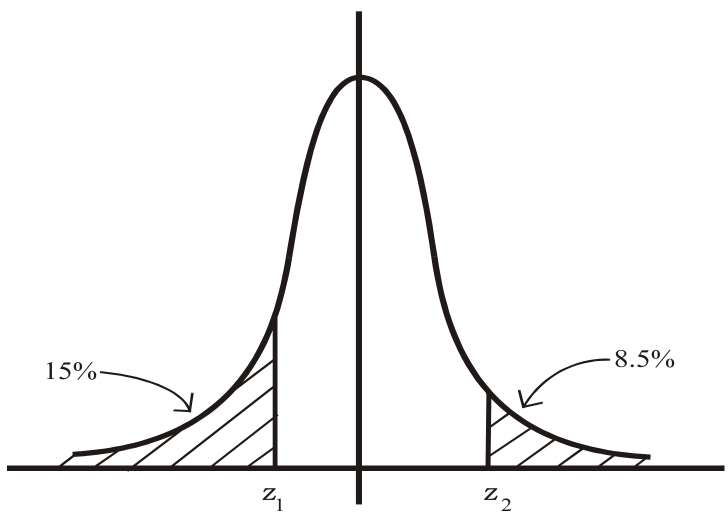
Therefore,

$$P(97 < X < 105) = 0.2734 + 0.3944$$
$$= 0.6678$$
$$= 67\% \text{ (2.S.F.)}$$

(ii) It is found that, for the output of Beta, 15% of the cricket balls are less than 96.5*mm*, and 8.5% are more than 103.4*mm*. Assuming a normal distribution, calculate the mean and standard deviation of the diameters of the cricket balls for Beta.

Solution

Firstly, we find the *z* scores corresponding to the probabilities of 15% and 8.5%.



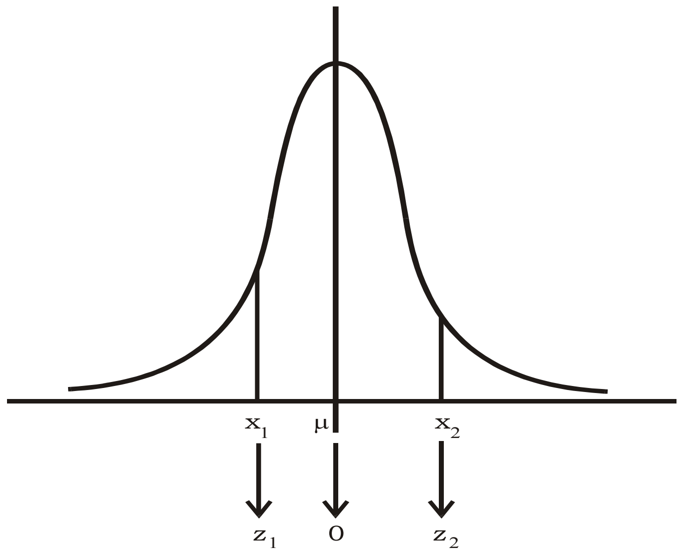
$P(z_1) = 50 - 15 = 35\% = 0.35$

$z_1 = 1.036$ (From tables)

$P(z_2) = 50 - 8.5 = 41.5\% = 0.415$

$z_2 = 1.372$

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



Then, $z_1 = \dfrac{x - \mu}{\sigma}$

Therefore, $-1.036 = \dfrac{96.5 - \mu}{\sigma}$

$1.036\sigma = \mu - 96.5 \qquad (1)$

and, $z_2 = \dfrac{x - \mu}{\sigma}$

Therefore, $1.372 = \dfrac{103.4 - \mu}{\sigma}$

$1.372\sigma = 103.4 - \mu$

$\mu = 103.4 - 1.372\sigma \qquad (2)$

On substituting (2) in (1):

$1.036\sigma = 103.4 - 1.372\sigma - 96.5$

$1.036\sigma + 1.372\sigma = 103.4 - 96.5$

$2.408\sigma = 6.9$

$\sigma = 2.87$ (3.S.F.)

$\mu = 103.4 - 1.372 \times 2.87$

$\mu = 99.5$ (3.S.F.)