The Normal Distribution

Prerequisites

You should be familiar with (1) the binomial distribution and (2) the distinction between discrete and continuous probability distributions.

Example (1)

In a series of experiments a fair coin is spun n times. Let X denote the number of times the coin lands heads up. Find the probability distribution of X for the different values of n given below. In each case make a graph of the probability distribution showing the possible values, x, that X can take on the horizontal axis and the probability that X takes these values, $P(X = x)$, on the vertical axis.

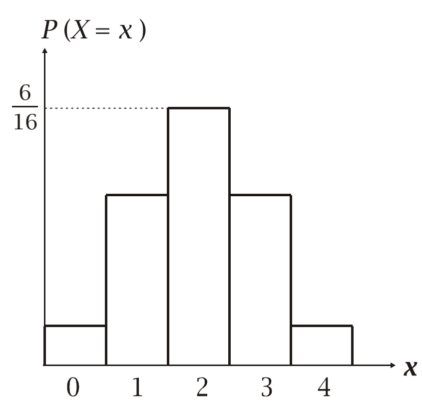
(i) $n = 4$
(ii) $n = 6$
(iii) $n = 10$
(iv) $n = 16$

Suppose n becomes larger and larger. Conjecture what shape the distribution will take.

Solution

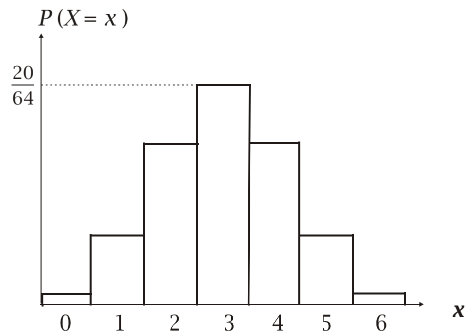
(i) $X \sim B\left(4, \frac{1}{2}\right)$

x	4	3	2	1	0
$P(X = x)$	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$



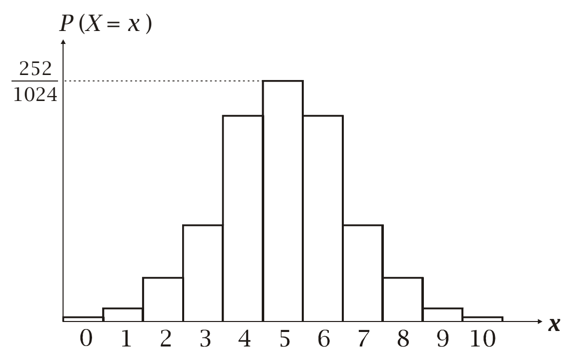
$(ii) \qquad X \sim B\left(6, \frac{1}{2}\right)$

x	0	1	2	3	4	5	6
$P(X=x)$	$\frac{1}{64}$	$\frac{6}{64}$	$\frac{15}{64}$	$\frac{20}{64}$	$\frac{15}{64}$	$\frac{6}{64}$	$\frac{1}{64}$



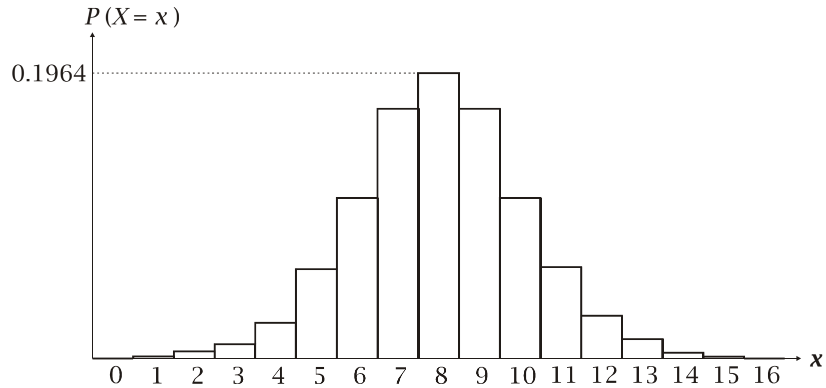
$(iii) \qquad X \sim B\left(10, \frac{1}{2}\right)$

x	0	1	2	3	4	5	6	7	8	9	10
$P(X=x)$	$\frac{1}{1024}$	$\frac{10}{1024}$	$\frac{45}{1024}$	$\frac{120}{1024}$	$\frac{210}{1024}$	$\frac{252}{1024}$	$\frac{210}{1024}$	$\frac{120}{1024}$	$\frac{45}{1024}$	$\frac{10}{1024}$	$\frac{1}{1024}$

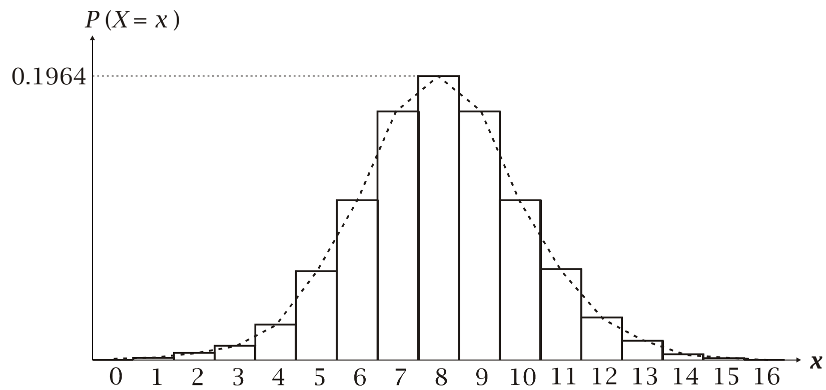


$(iv) \qquad X \sim B\left(16, \frac{1}{2}\right)$

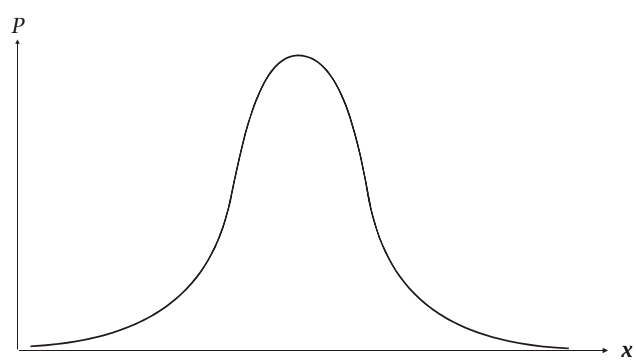
x	0	1	2	3	4	5	6	7	8
$P(X=x)$	.0000	.0002	.0018	.0085	.0278	.0667	.1222	.1746	.1963
x		9	10	11	12	13	14	15	16
$P(X=x)$		.1746	.1222	.0667	.0278	.0085	.0018	.0002	.0000



Let us join the midpoints of the rectangles in this graph to form a frequency polygon.



We see that as *n* gets larger and larger the shape of the graph is getting closer and closer to a smooth continuous bell-shaped curve. We call this the continuous smooth curve that arises from the binomial distribution this way as $n \to \infty$ the *normal distribution*. We describe this curve as "bell-shaped".

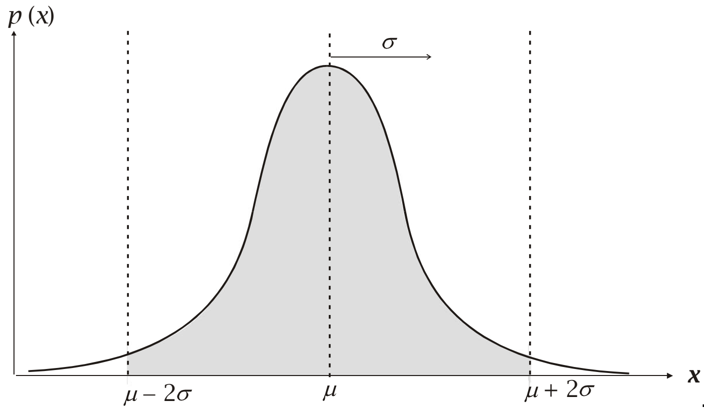


This continuous bell-shaped curve is the graph of a probability density function.

The normal distribution

The normal distribution is the most important of the continuous probability distributions. It arises as the limit of the binomial distribution $X \sim B(n,p)$ as the number of trials *n*, becomes greater and greater. This distribution is of importance in natural science, because it is expected that many properties of nature, if sampled randomly, follow the normal distribution. For example, the distribution of a sample of adult male heights would, if sampled, be expected to produce a histogram that corresponded more and more to the probability density function that gives the normal curve as the sample size increased. The graph of the normal distribution is "bell" shaped, and is symmetric about its middle value – which is the mean, μ, of the sample. The curve is such that 95% of the area under it lies within 2 standard deviations, σ, of the mean. When X is a continuous random variable that follows a normal distribution 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.



Historical note

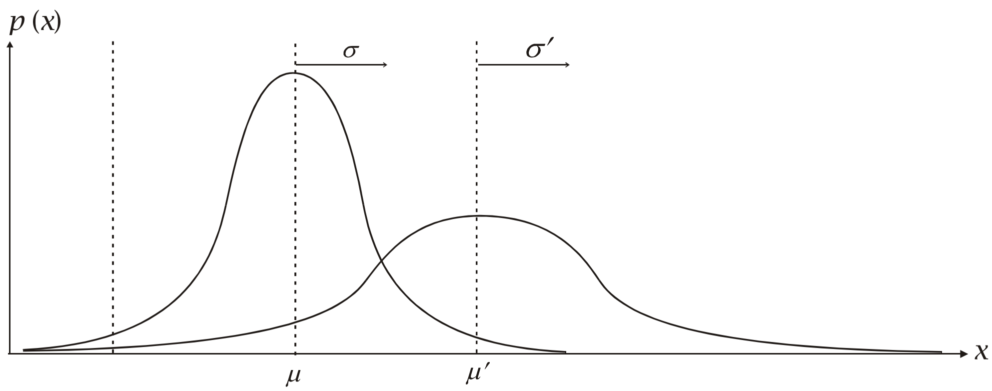
The mathematician Gauss derived an expression for the probability density function $p(x)$ for a normally distributed variable $X \sim N\left(\mu,\sigma^2\right)$ as

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)}$$

where x is a value of X lying in the interval $(-\infty,\infty)$. However, we do not use this rather nasty looking function directly but derive probabilities from tables of values.

Standard 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.



The probability density function of one normal variable $X \sim N\left(\mu,\sigma^2\right)$ may be derived from that of another $Y \sim N\left(\mu',(\sigma')^2\right)$ by means of a single translation followed by a single horizontal scaling. Therefore, it makes sense to represent the normal distribution by a single standardised distribution.

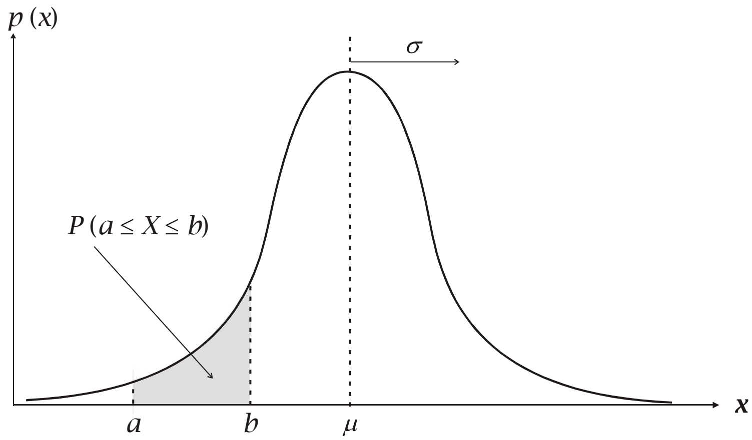
Standard normal variable

The *standard normal variable Z* is the normal variable with mean $\mu = 0$ and standard deviation $\sigma = 1$.

$Z \sim N(0,1)$

For a continuous probability distribution probabilities are not directly assigned to values *x* that the variable *X* can take. Instead to such a value, *x*, there is assigned a probability density $p(x)$. Probabilities are assigned to intervals $a \le X \le b$ and are given as the integrals of the probability density function over this integral.

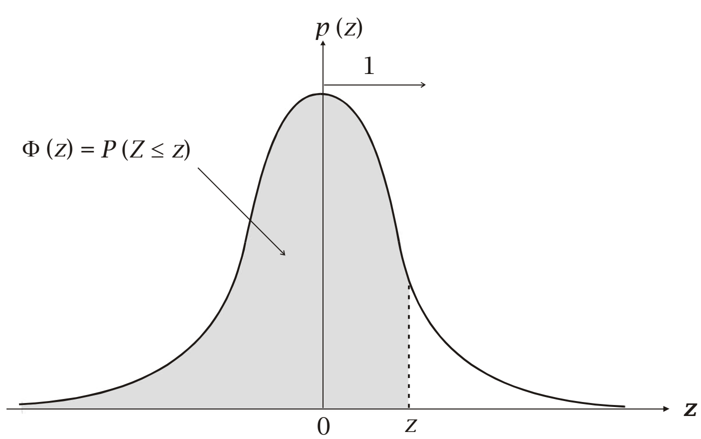
$P(a \le X \le b) = \int_a^b p(x)\,dx$



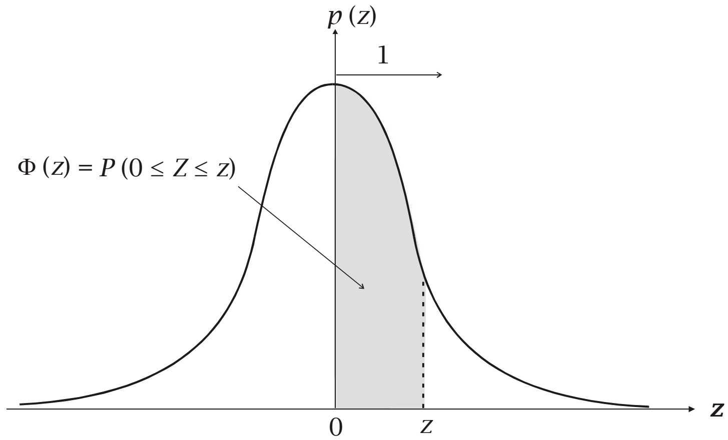
The probability density function of a normal distribution is defined on the whole real line; that is, for the interval $(-\infty,\infty)$. This means that it is theoretically possible for a value of a normal variable to lie anywhere, though the probability of doing so diminishes to practically zero the farther and farther away from the mean the value lies.

Let *z* denote a value of the standard normal variable $Z \sim N(0,1)$. Since the standard deviation is $\sigma = 1$ then *z* represents the number of standard deviations of *z* from the mean $\mu = 0$. The values of *z* may be positive or negative depending on whether *z* lies below the mean or above it. Tables of probabilities map a given value that *Z* takes $Z = z$ to the probability $\Phi(z)$ of an interval associated with this value. Different tables associate different regions of the distribution with the value *z*.

(**1**) In one format the table gives the integral of the probability density function over the region $(-\infty, z]$ so that $\Phi(z) = P(Z \le z) = \int_{-\infty}^{z} p(z)dz$.



(2) In another format the table gives the integral of the probability density function over the region $[0, z]$ so that $\Phi(z) = P(0 \le Z \le z) = \int_0^z p(z)\,dz$.



There are other possible formats for tables of values. Here we shall use tables in the first form where the probability associated with a value z of the standard normal variable $Z \sim N(0,1)$ is

$\Phi(z) = P(Z \le z) = \int_{-\infty}^{z} p(z)\,dz$.

Example (2)

Find $P(Z > 1.236)$.

Solution

The tables we shall use will give us the value $\Phi(1.236) = P(Z < 1.236)$.

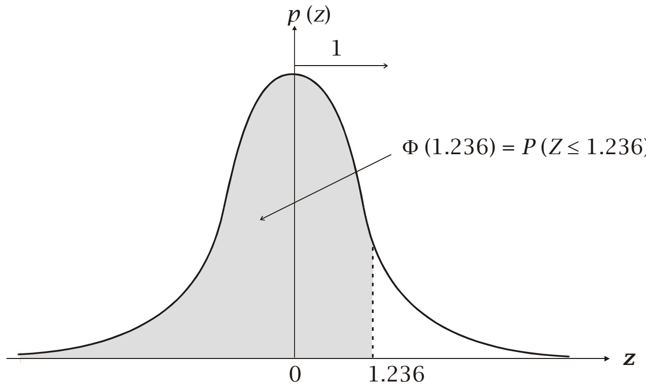
Below shows part of the table with the entries relevant to $z = 1.236$ highlighted.

z	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
											ADD								
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621	2	5	7	9	12	14	16	19	21
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830	2	4	6	8	10	12	14	16	18
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015	2	4	5	7	9	11	13	15	17
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177	2	3	5	6	8	10	11	13	14
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319	1	3	4	6	7	8	10	11	13

The box shaded on the left gives the value of $\Phi(1.23)$, which is $\Phi(1.23) = 0.8907$. The box shaded on the right gives the value that we must *add* to this to accommodate the third decimal place and bring $\Phi(1.23)$ up to $\Phi(1.236)$.

$$\Phi(1.236) = \Phi(1.23) + 0.0011 = 0.8907 + 0.0011 = 0.8918$$

The following diagram shows the region to which this probability is assigned.



The region we seek is in fact the area under the curve that is *not* shaded.

$$\begin{aligned} P(Z > 1.236) &= 1 - P(Z \le 1.236) \\ &= 1 - 0.8918 \\ &= 0.1082 \end{aligned}$$

Remark

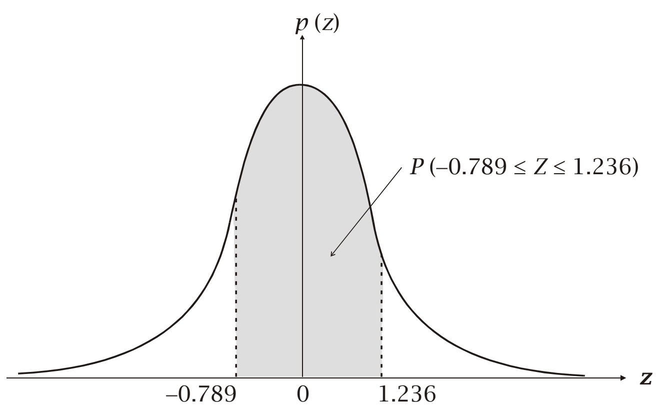
Because of the way probability is defined for a continuous distribution as the integral of the area under the probability density function, then the probability of any *point value* is $P(Z = z) = 0$. Therefore, there is no actual distinction in this context between an exact $(\le)$ and an inexact inequality $(<)$ and $P(Z \le z) = P(Z < z)$.

Example (3)

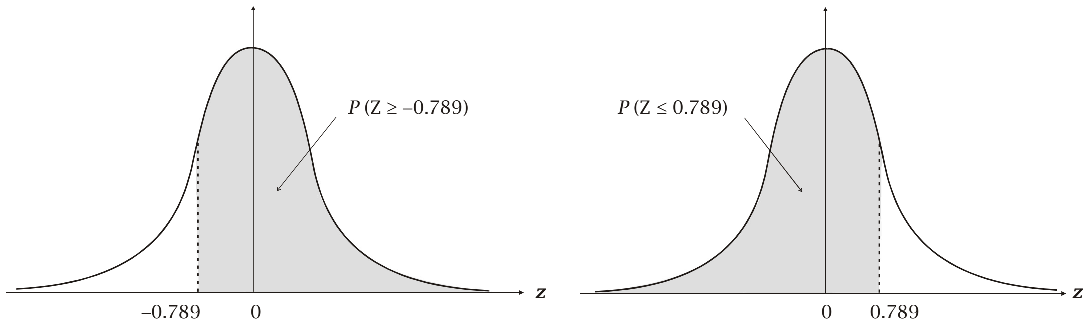
Find $P(-0.789 < Z < 1.236)$.

Solution

The following diagram shows the region we are seeking.



To find the probability associated with $z = -0.789$ we argue by symmetry as the next diagram illustrates.



Therefore

$$P(Z > -0.789) = P(Z < 0.789) = 0.7850$$

Finally

$$\begin{aligned} P(-0.789 < Z < 1.236) &= P(Z < 1.236) - (1 - P(Z < -0.789)) \\ &= 0.8918 - (1 - 0.7850) \\ &= 0.6768 \end{aligned}$$

Standardising any normal variable

We have said that any normal distribution $X \sim N(\mu, \sigma^2)$ can be transformed into the standard distribution $Z \sim N(0,1)$ by

(1) A translation of the mean from its given value to 0.

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

Hence any value *x* that the variable $X \sim N(\mu, \sigma^2)$ takes can be mapped to a standard *z*-value (also called a *z-score*) taken by the standard normal variable $Z \sim N(0,1)$. The mapping is

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

This is intuitively obvious. The expression $x - \mu$ subtracts the value of the mean from the given *x* value; this is then divided by the standard deviation of $X \sim N(\mu, \sigma^2)$ to find the number of standard deviations of *x* from the mean μ, which is precisely what the *z*-value gives. Therefore, any problem requiring one to find the probability of a region of $X \sim N(\mu, \sigma^2)$ can be translated into a problem involving a z-value and tables can be used.

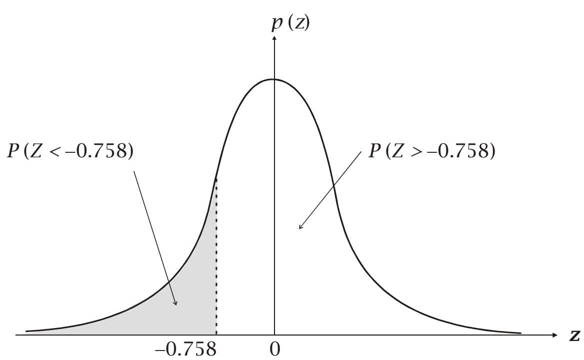
Example (4)

$X \sim N(4, 1.32^2)$. Find $P(X < 3)$.

Solution

The z value corresponding to $x = 3$ is

$$z = \frac{x - \mu}{\sigma} = \frac{3 - 4}{1.32} = -0.758 \text{ (3 s.f.)}$$



$$\begin{aligned} P(X < 3) &= P(Z < -0.758) \\ &= 1 - P(Z < 0.758) \\ &= 1 - 0.7758 \\ &= 0.2242 \\ &= 0.224 \text{ (3 s.f.)} \end{aligned}$$

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 $X \sim N(\mu, \sigma^2)$ to find a probability corresponding to that value.

It is also possible

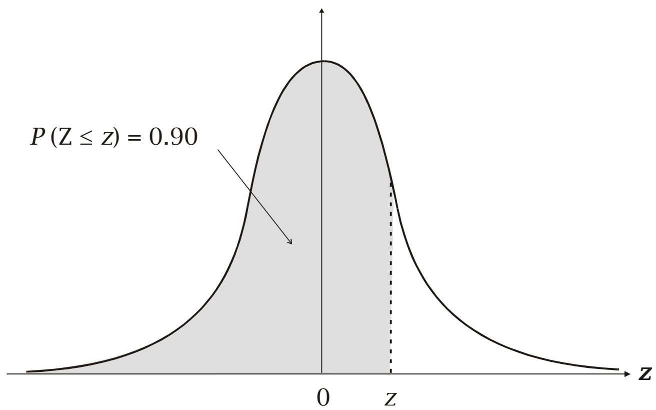
(3) Given a probability, to find the corresponding z value.

Example (5)

$X \sim N\left(6, 1.54^2\right)$. Find the value of a such that $P(X < a) = 0.90$.

Solution

$P(X < a) = 0.90$ corresponds to $\Phi(z) = 0.90$ as indicated in the following diagram.



To find this z value we work "backwards" from the tables.

z	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
											ADD								
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621	2	5	7	9	12	14	16	19	21
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830	2	4	6	8	10	12	14	16	18
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015	2	4	5	7	9	11	13	15	17
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177	2	3	5	6	8	10	11	13	14
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319	1	3	4	6	7	8	10	11	13

The tables show that $\Phi(1.281) = 0.8999$ which is the closest we can get to the value 0.90.

Then $z = 1.281$ and on substituting into $z = \dfrac{x - \mu}{\sigma}$ we have $1.281 = \dfrac{x - \mu}{\sigma}$. We are given $X \sim N\left(6, 1.54^2\right)$, so $\mu = 6$ and $\sigma = 1.54$; hence

$$1.281 = \frac{x - 6}{1.54}$$

$$\begin{aligned} x &= 1.281 \times 1.54 + 6 \\ &= 7.972\ldots \\ &= 7.97 \text{ (3 s.f.)} \end{aligned}$$

This example illustrates that we can rearrange $z = \frac{x - \mu}{\sigma}$ to obtain $x = \mu + \sigma z$. In practice it is easier to substitute the values into $z = \frac{x - \mu}{\sigma}$ and find the corresponding *x* value. 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 from the set $\{z, x, \mu, \sigma\}$ given two others.

Example (6)

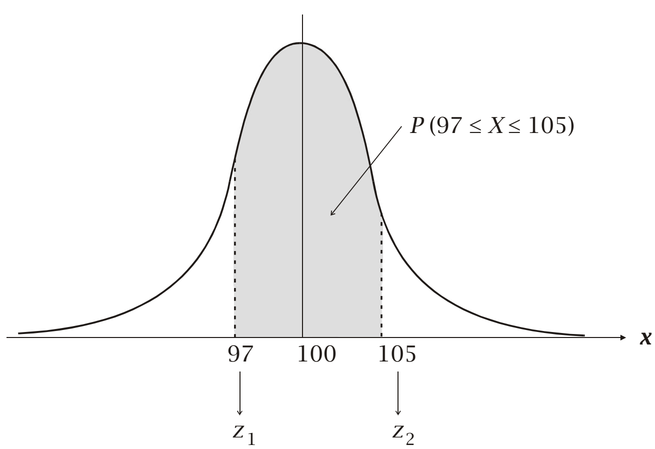
Two companies, Alpha and Beta, produce cricket balls. The diameters of the cricket balls are measured in millimetres.

(*a*) The diameters of the output of Alpha, *X*, 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? Give your answer to 3 significant figures.

(*b*) It is found that, for the output of Beta, the diameters of 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. Give your answer to 3 significant figures.

Solution

(*a*) $X \sim N(100, 4^2)$



We seek the *z*-values corresponding to the *x*-values $x_1 = 97$ and $x_2 = 105$.

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

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

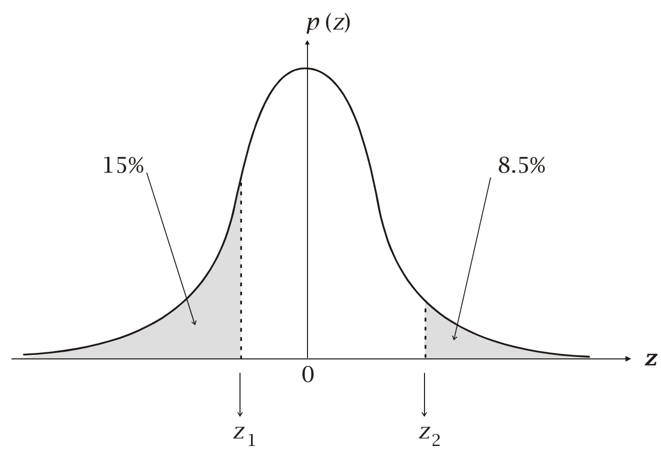
$$z_2 = \frac{105-100}{4} = +1.25$$

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

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

$$P(97 < X < 105) = 0.2734 + 0.3944$$
$$= 0.6678$$
$$= 66.8\% \text{ (3 s.f.)}$$

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



$$\Phi(z_1) = 50 - 15 = 35\% = 0.35 \quad \Rightarrow \quad z_1 = 1.036$$

$$\Phi(z_2) = 50 - 8.5 = 41.5\% = 0.415 \quad \Rightarrow \quad z_2 = 1.372$$

Let Y denote the diameter of cricket balls of Company Beta. Then $Y \sim N(\mu, \sigma^2)$.

Then

$$z_1 = \frac{x-\mu}{\sigma}$$

$$-1.036 = \frac{96.5-\mu}{\sigma}$$

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

$$z_2 = \frac{x-\mu}{\sigma}$$

$$1.372 = \frac{103.4-\mu}{\sigma}$$

$$1.372\sigma = 103.4 - \mu$$

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

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.)