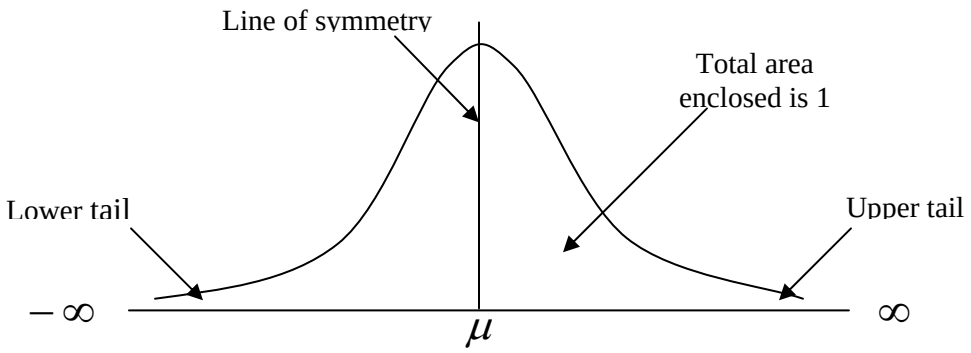
The Normal distribution

The Normal distribution is a continuous theoretical probability distribution and, probably, the most important distribution in Statistics. Its name is justified by the fact that it is suitable to almost any variable in *normal* real-life situations. For example, the distribution of heights of people is normally distributed, that is, there are relatively few extreme values (giants and dwarfs) whilst most people are of medium height, causing a concentration of observations towards the center of the distribution. This phenomenon is best understood when observing the *probability density function* (PDF) of the Normal distribution.

Unlike the *probability mass function* (PMF) for discrete distributions, the probability density function for continuous distributions does not generate probabilities by substitution of values. The PDF is in fact the equation of a curve that represents the *relative frequency* or *frequency density* of occurrence of values of a random variable between given limits. A PDF encloses a total area of 1 so that it can be used to calculate probabilities by mere integration. This is obvious since, in any probability problem, the total probability is 1 and, in this case, analogical to the total area enclosed.

The Normal distribution is completely characterised by its two parameters μ and σ^2, its mean and variance respectively, hence the notation $X \sim N(\mu, \sigma^2)$. The Normal curve is symmetrical about the line $x = \mu$ and bell-shaped. Since a normal variable assumes values ranging from $-\infty$ to ∞, its curve is asymptotic to the *x*-axis, that is, the normal curve keeps on approaching the *x*-axis without touching it at its ends, known as the lower and upper tails. The diagram below is a representation of the normal curve:



As mentioned before, probabilities (areas) are determined by integration. For example, if we were to calculate P[$X < 58$] given that $X \sim N(66, 25)$, the answer would simply be

$$\int_{-\infty}^{58} f(x)\,dx$$

which would represent the area under the curve extending from $-\infty$ to 58, $f(x)$ being naturally the PDF of the Normal distribution.

Note

For continuous distributions, there is no difference between $P[X < a]$ and $P[X \leq a]$ for any a. This is because $P[X \leq a] = P[X < a] + P[X = a]$ and $P[X = a] = 0$ (area of a line). Consequently, we are not expected to calculate probabilities of the type P[X = 45], for example, since the answer would be automatically zero!

The problem arises due to the very complicated nature of the PDF of the Normal distribution, which is given by

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

which is not easily integrable and which would need to be integrated every time that we have to determine a probability.

Fortunately, there exists a Normal table from which we can actually read the areas (probabilities) under curve. It would be extremely ambitious to derive areas for every possible Normal distribution with its own mean and variance. In fact, there is only one such table – the *standard* Normal table where the areas for the Normal variable $Z \sim$ N(0, 1) (mean 0 and variance 1) are given.

Thus, given any Normal distribution $X \sim$ N(μ, σ^2), a transformation, known as *standardisation*, is applied in order to map it onto the standard Normal distribution so that probabilities (areas) can be read. The standardisation formula is given by

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

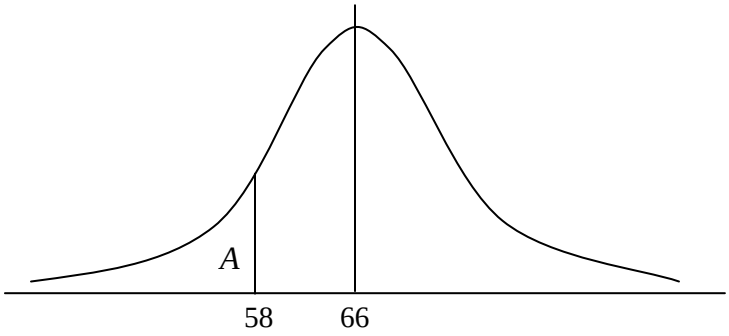
A careful glance at the above tells us that z is just the *number of standard deviations* between the mean and any given x-value.

In the previous example, where we had to determine P[X < 58] given that $X \sim$ N(66, 25), we proceed by standardising the distribution as follows:

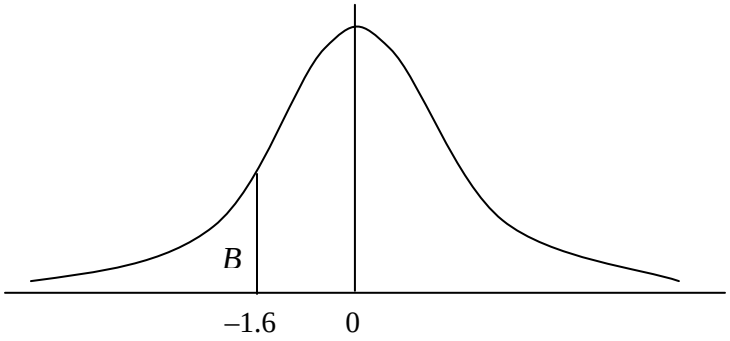
At 58, $z = \frac{58-66}{5} = -1.6$; this means that there are 1.6 standard deviations between 58 and 66, the negative sign indicating that 58 is found to the *left* of 66 on the real number line. We deduce that areas are determined according to the number of standard deviations on the horizontal axis of the normal curve.

It is true that the initial and standard Normal distributions do not have the same shape because of their different standard deviations (or variances) but areas under the Normal curves are only determined by z-values. The diagrams below (not to scale) show the similarity in terms of areas.

X~ N(66, 25)



Z~ N(0, 1)



Areas A and B are both equivalent to 1 – 0.9452 = 0.0548 from the cumulative probability table under the Normal curve. Therefore, P[X < 58] = 0.0548.

Example

Given X~ N(41, 36) and that P[X > b] = 0.05, find the value of b.

Solution

If the area to the right-hand side of b is 0.05, it means that b is to the *right* of the mean 41 (think about it and verify this fact by drawing a normal distribution!) Bear in mind that the location of b is made easier by comparing with an area of 0.5, that is, half of the area under the normal curve.

$$\text{At } b,\ z = \frac{b - 41}{6} = 1.645\text{; therefore, } b = 41 + (6)(1.645) = 50.87.$$

You will probably have to look for an area of 1 – 0.05 = 0.95 from the cumulative normal table in order to obtain the value 1.645. It means that b is 1.645 standard deviations to the right of the mean.

Example

Given X~ N(27, 25) and that P[$X > c$] = 0.90, find the value of c.

Solution

This time we have to be very careful when determining the z-value. Using the same reasoning as that in the previous example, since the area to the right of c is 0.95, it is clear that c is found on the *left* of the mean.

At c, $z = \frac{c - 27}{5} = -1.282$; note the *negative sign* in the z-value. This is the most common mistake made by students – they forget the location of c with respect to the mean. The fact that there are no negative z-values in the normal table does not help either! The value of c is therefore 27 – (5)(1.282) = 20.59.

Note

We should *always* draw a normal distribution (sketch) to verify the validity of our answer. If we made the mistake of using 1.282 instead of –1.282, our answer would have been greater than the mean 27 (which does not make sense since we already said previously that c is found on the *left* of the mean.)

Example

Given X~ N(20, σ^2) and that P[$X > 12$] = 0.75, find the value of σ.

Solution

It is clear that 12 is found on the left of the mean 20.

Thus, at 12, $z = \frac{12 - 20}{\sigma} = -0.674$ and $\sigma = \frac{12 - 20}{-0.674} = 11.87$.

Note

It is assumed that you can easily read the normal table, obtain areas for given z-values and z-values for given areas.

Example

Given X~ N(μ , 16) and that P[X > 53] = 0.005, find the value of μ .

Solution

Using some logical thinking, we can deduce that 53 is found on the right of μ since the area to the right of 53 is 0.005.

Thus, at 53, $z = \dfrac{53 - \mu}{4} = 2.576$ and $\mu = 53 - (4)(2.576) = 42.696$.

Example

Given X~ N(μ , σ^2), P[X < 63] = 0.975 and P[X > 46] = 0.6, find the values of μ and σ .

Solution

By inspection of the probability statements, we deduce that 46 and 63 are to the left and right of μ respectively.

At 63, $z = \dfrac{63 - \mu}{\sigma} = 1.96$ and at 46, $z = \dfrac{46 - \mu}{\sigma} = -0.253$. This gives two equations to be solved simultaneously:

$$\mu + 1.96\sigma = 63 \qquad \mu - 0.253\sigma = 46$$

Verify that the solutions are $\mu = 47.95$ and $\sigma = 7.68$ (correct to 2 decimal places).