Normal Distribution

In the first week, we constructed histograms of sample data of continuous variables, such as height. As we collect more and more data, we can make our histogram more detailed and the intervals more narrow. When we have sampled an infinite amount of data, the histogram becomes perfectly smooth—a continuous probability distribution. The area under the probability distribution between two values of x, x_1 and x_2, corresponds to the probability that a randomly-selected x will have a value between x_1 and x_2: $P(x_1 < x < x_2)$.

One very popular probability distribution is called the "normal" distribution. Although you don't have to know it, the equation for the distribution is:

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

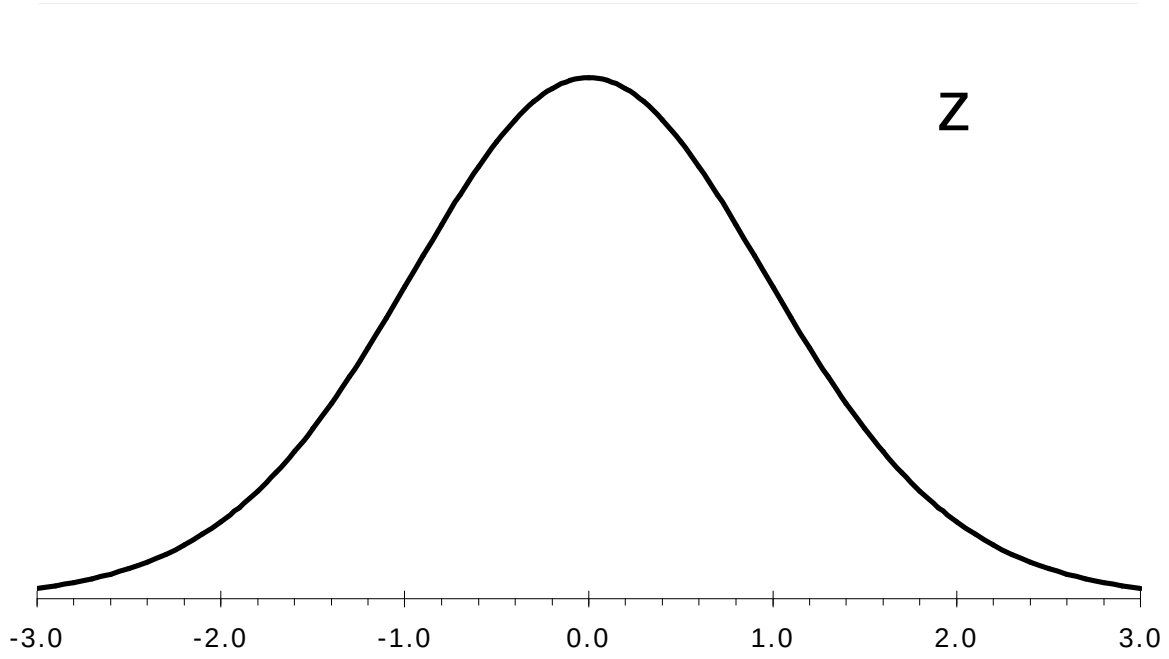
Only two parameters are required to completely describe a normal distribution: the mean and the standard deviation. In this case, we use μ to denote the mean of the actual population, and σ to denote its standard deviation. Earlier, we computed the mean ($\bar{x}$) and standard deviation (s) of a sample; these are approximations to the real mean and standard deviation of the population. As the sample grows larger and larger, $\bar{x}$ and s grow closer and closer to μ and σ.

As noted above, the probability is given by the area under the probability distribution. Unfortunately, there is no analytical solution to the integral of the normal distribution—the area has to be computed numerically. For this reason, we usually transform normal distributions into the “standard” normal distribution by defining a new variable, z, as follows:

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

Z is a measure of the number standard deviations that x is from the mean. This allows us to transformation allows us to use one table to look up the areas (probabilities) between any two values of z.

The standard normal distribution looks like this:



Why is the normal distribution so popular? It turns out that data generated by many phenomena can be approximated by a normal distribution. To get this formula, you only have to assume three things: (1) that the total probability of finding an event somewhere between $-\infty$ and ∞ is 1; (2) that the maximum of the distribution is at the mean (i.e., that the distribution is symmetrical); and (3) that the probability of finding an event in one interval and then in another is equal to the product of the probabilities of finding the event in each interval separately.

The mathematical theorem that underlies the normal distribution is called the "central limit theorem." It states that the average of independent random values tends toward the normal distribution no matter how the random variables are distributed.

There a many levels to understanding the central limit theorem. Think about the height of a person. People of a given age and gender have a well-defined mean height, but many genetic and environmental variables go into determining the height of a particular person. We can think of these as random variables (although they are not). These result in the spread.

The normal distribution is so important that you're going to get to know it inside out. Remember, probabilities are represented by the area under the curve, not by the height of the curve. The area under the normal distribution (or any probability distribution, for that matter) is 1 or 100%. Since curve is perfectly symmetrical, the median equals the mean, and 50% of the area is above the mean and 50% is below the mean.

Learn how to find the area between two values of x. You can do this using the following function in Excel, which gives the area between $-\infty$ and x:

$$P(x < x_0) = \text{NORMDIST}(x_0,\mu,\sigma,1)$$

If you want to know the area between x and $+\infty$, use

$$P(x > x_0) = 1 - \text{NORMDIST}(x_0,\mu,\sigma,1)$$

If you want to know the area between x_1 and x_2, use

$$P(x_1 < x < x_2) = \text{NORMDIST}(x_2,\mu,\sigma,1) - \text{NORMDIST}(x_1,\mu,\sigma,1)$$

You can also convert x to z:

$$P(x < x_0) = P(z < z_0) = \text{NORMSDIST}(z_0)$$

In addition, you should learn how to find the area under the normal distribution using tables. The table in Wonnacott and Wonnacott gives the area to the right of z (i.e., $P(z > z_0)$).