The Normal Distribution

There's a model that shows up over and over again when we look at data from the real world. Whether in the heights of people, the weights of elephants, even IQ scores. All of these have distributions approximating the famous bell curve.

In the 17th century, Galileo noted that errors made around astronomical observations were symmetrically distributed, and that smaller errors tended to occur more frequently than larger errors. Then in the early 18th century, DeMoivre started investigating the probability of binomial distributions. This is the distribution of overall frequency, or predicted occurrence, when we look at success and failure probabilities.

For example, recall that when we toss two coins, we have a one in four chance of obtaining zero heads, a one in two chance of obtaining one head, and a one in four chance of obtaining two heads. When we toss three, we have a one in eight chance of no heads, a three in eight chance of one head, or two heads, and a one in eight chance of three heads.

It turns out that as we increase the number of coin tosses, the frequency of heads in the outcomes becomes closer and closer to that normal distribution bell curve.

The bell curve is described in terms of two parameters: the mean, which is the centre of the distribution, or the value that all the observations are usually clustered around, and the standard deviation, which describes how wide the distribution is.

For data that is normally distributed we expect 68% of the observations to fall within one standard deviation of the mean, we expect 95% of data to fall within two standard deviations of the mean, and we expect 99.7% of data to fall within three standard deviations of the mean.

This means that, for example, if we know that the weights of male Asian elephants are normally distributed with a mean of 3,300 kilograms and a standard deviation of 400 kilograms, we expect that 68% will be between 2,900 and 3,700 kilograms, 95% will be between 2,500 and 4,100 kilograms, and almost all Asian elephants will weigh between 2,100 and four 4,500. Of course, in nature, the distributions aren't always perfectly normal, but we do often see observations distributed in a way that resembles this bell curve.

An explanation for why this happens comes back to the coin tosses example. Something like height comes down to a number of genetic and environmental factors, and each one may contribute a little towards whether or not someone grows taller or shorter. Putting all these random processes together, and observing the effect on height, ends up being like flipping a huge number of coins and counting the number that came up tall.