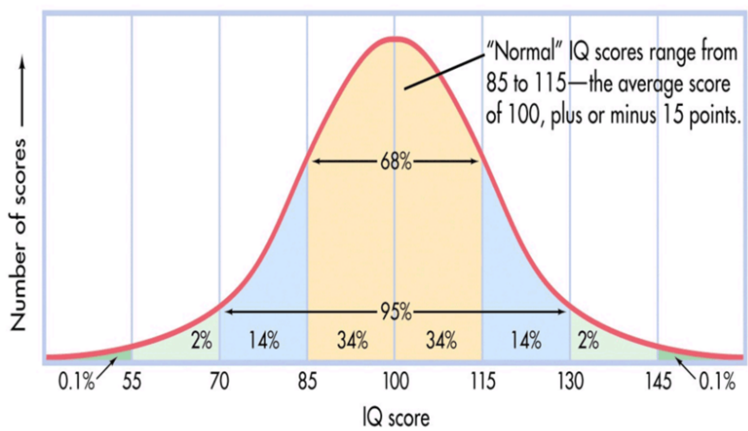
Range, Normal Distribution, and Standard Deviation

When we collect data even for the same thing (e.g., the test scores of a given class) it typically spans a range in any given dimension. **Range** in unidimensional data is nothing more than the difference between maximum and minimum values. In test scores within a class, for instance, the maximum score might be 100% and the minimum score might be 24%. The range of scores is therefore $100\% - 24\% = 76\%.$

When we gather scientific data (say determining the period of a swinging pendulum time and again with a stopwatch in order to get "the best value" for a given set of parameters – such as length, amplitude, and/or mass), we generally find that the data contain a certain amount of random error. The error in the measured period probably results in the slow response of the human reflex, lack of attention, etc. One might measure the same period time and again and get different values. It's true for everyone.

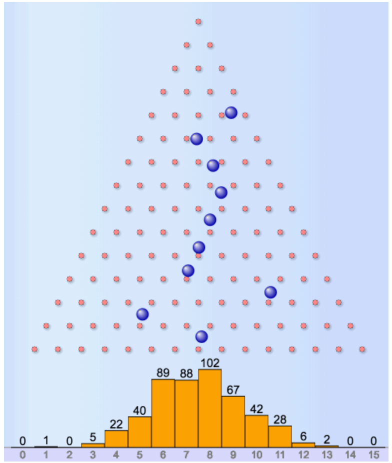
This error is typically distributed in a particular fashion, particularly if it is random error. When this occurs, we typically get a "bell shaped" distribution. A histogram showing IQ scores for a large group of students might look like this:

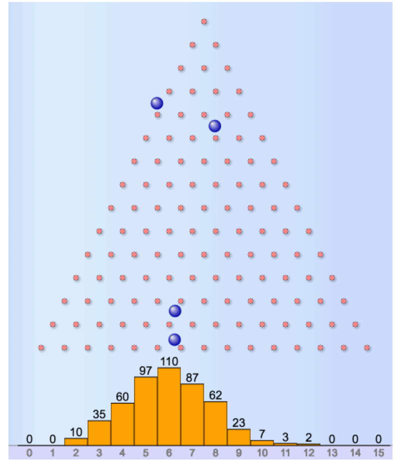


This distribution clusters around a central value (100, the arithmetic mean), and the "wings" of the distribution are symmetrical around this central value. IQ scores of 55 or less are quite rare as are IQ scores of 145.

Such symmetrical distributions of data that typically following a normal distribution are things like IQs, heights of people, errors in measurements, blood pressure, scores on a test and so forth. Of course, this assumes that there is large enough sampling to get the normal distributions. A small sample can result in data that are skewed – being spread out on one end more so than the other – lopsided if you will – or show multiple peak values, or simply jumbled. If fewer than 30 data points are collected, it's unlikely that the distribution of data will reasonably reflect a bell curve.

To see how a normal distribution is produced on the basis if randomness, consider a variation of the ancient quincunx. The left image below shows what happen when 500 marbles are allowed to drop through a field of pins where the odds of moving left or right are 50/50. As the number of marbles approaches infinity, we move toward a perfect normal distribution. A 60/40 bias toward the left produces an asymmetrical distribution such as shown in the second image below. Such a distribution is said to be skewed toward the left. That is, the left wing is much wider than the right wing. This skewness might be real or reflect a bias in data collection. Skewness toward the right is also possible.





There are two additional ways of characterizing a normal distribution. One deals with how wide a distribution is. A standard distribution can be rather short and wide or rather high and narrow. (A normal distribution will always be symmetrical left and right unless skewed at which point it is no longer a normal distribution.) The measurements of these phenomena are known as standard deviations. If the standard deviation from the central value is small, then the base of the distribution is narrow and the peak is high. If the standard deviation of values from the central value is large, then the base of the distribution is wide and the peak low. This of course assumes the same number of data points.

Standard deviation is symbolized by σ (the Greek letter sigma). Standard deviation is the square root of the variance – the average of the squared differences from the mean. Calculating the variance, σ^2, is easy. Consider the situation where you want to measure a physical variable 10 times in order to get the best value assuming that there is random error in your measurement.

- First, collect your date.
- Second, calculate the mean of all numbers in your data (the simple average of the numbers)
- Third, for each number, subtract the mean and square the result (the *squared difference*).
- Fourth, work out the average of those squared differences by dividing by the number of data points.
- Fifth, take the square root of the variance and you have the standard deviation.

$$\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$$

The standard deviation allows us to determine just how far something is from the mean in relation to other data. A detailed study of the normal distribution shows us that (see above drawing):

68% of values are within **1 standard deviation** of the mean ($\pm 1\sigma$)

95% of values are within **2 standard deviations** of the mean ($\pm 2\sigma$)

99.7% of values are within **3 standard deviations** of the mean ($\pm 3\sigma$)

All distributions have means, medians, and modes. These are presented in *Measurements of Central Tendency.*