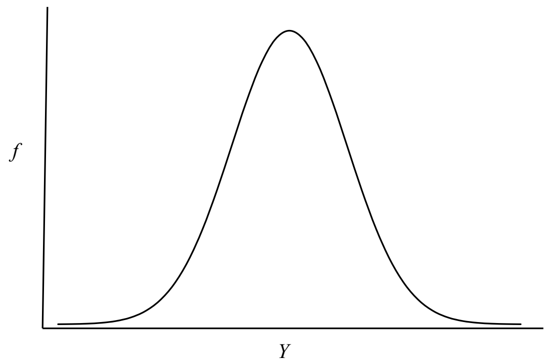
The Normal Distribution and *z*-scores

The normal distribution is the most common reference for distributions in the behavioral and social sciences. It is a theoretical frequency distribution that is symmetric and has particular known mathematical properties that make it valuable for comparing scores. Below is an approximately normal distribution.



The height of the curve represents the frequency of occurrence in the data set. The most frequent scores are in the middle and the least frequent scores appear at the right and left extremes (the tails). In statistical terms, we can think of the height also as a probability of observing a Y score of that value.

Many natural phenomena follow or can be estimated well with a normal or approximately normal probability distribution, including human height, tree failure, velocity of molecules in a gas, atmospheric temperatures, leaf widths, and a variety of social and behavioral variables. The normal distribution is not the only distribution. There are a number of other probability distributions, such as the binomial, Poisson, and so on, some of which we will discuss later.

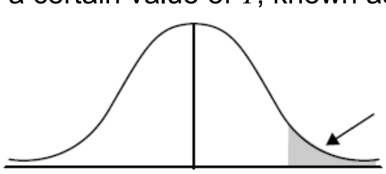
You can generate this shape with the following (terrifying!) formula.

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/2\sigma^2}$$

Myers, J.L., & Well, A.D., & Lorch, R.F.,Jr. (2010). *Research design and statistical analysis (3rd Edition*). Mahwah, NJ: Erlbaum

You won't be computing anything with this, but just so you know, the symbol π is the mathematical constant (approximately 3.14), e is the exponent function, y is an individual score in the data set, μ is the mean, σ is the standard deviation, and σ^2 is the variance.

The mathematical formula for the normal distribution is important, because we can use it to estimate (using integral calculus) the area under the curve and therefore the percentages of cases in the data set at or below a certain value of Y, known as percentiles.



Myers, (2010)

If we can assume that the scores we have are normally distributed or close enough for comfort, the normal distribution is a valuable tool.

Standardized scores (z-scores)
Using a simple conversion of the scores in the data set and what we know about the normal distribution, we can compute percentiles.

$$z_i = \frac{Y_i - \overline{Y}}{s}$$

z_i is the new standard score, Y_i is the original or raw score, $\overline{Y}$, is the mean of the sample, and s is the sample standard deviation. Each raw score can be transformed into an associated z-score. You can then use a spreadsheet, statistical software, or a table (e.g., Table C.2) in the text to find the associated proportion. Multiplying by 100 gives the percent, which is the form percentiles are usually reported in.

A full set of standardized scores has a mean of 0 (remember, we subtracted out the mean for each one) and standard deviation and variance equal to 1. The fact that all data sets have these characteristics when the raw scores have been transformed to standardized scores makes it a convenient metric for comparing scores. This is why many standardized tests, like the SAT, GRE, MMPI, use a transformation of the standardized scores by adding the desired mean (e.g., 500) and multiplying the z-value by the desired standard deviation (e.g., 100).

Note that transforming scores into z-scores does not change the shape of the original distribution. These are linear transformations. If it is not normal to begin with, it will not be normal by transforming the scores in this way. Subtracting all the scores by the same number and dividing them by the same number will not change the shape of the distribution. It is easy to confuse standardized scores with "normalized" scores, which attempt to make a non-normal distribution normal through non-linear transformations.