Research Methods 2: Week 4

Appendix 1: the form of the Sample Standard Deviation

The idea in computing a measure of spread is first to define a measure of location and then to measure how dispersed the observations are about that point. In this instance the mean is the measure of location, so it is natural that the SD should be based on the differences of the individual points from the sample mean, i.e. the expression is based on $x_1 - m, x_2 - m, \ldots, x_n - m$.

However, one 'obvious' measure of spread, namely to find the average of these deviations, will not work. This is because if you add up $x_1 - m, x_2 - m, \ldots, x_n - m$ the result will always be zero because of the definition of the mean: there will always be the same total deviation on the negative side (below *m*) as there is on the positive side (above *m*). A simple way round this difficulty would be to find the average of these deviations without regard to their sign. This is known as the mean absolute deviation about the mean and is a legitimate measure of spread. However its mathematical properties are poor and the theory based on this measure of spread is not as rich as that using the SD, so it is seldom used.

An alternative way to remove the balance between positive and negative deviations is to square all the deviations, making them all positive and then take the 'mean' of these values. This results in a quantity which is known as the *variance*. It is a very important quantity in theoretical work but is of little direct use in practice because it is on the scale of squared units, i.e. the variance of the heights would be in cm^2. This is readily rectified by taking the square root, and it is this quantity which is the SD, and which has the same units as the original measurements.

The 'mean' of the squares deviations was placed in quotes because, instead of dividing the sum of the *n* squared deviations by *n*, the divisor *n*-1 is used. This is largely a technical matter and is related to the fact that the spread would ideally be measured about the population mean m, rather than the sample mean *m*. The sample mean, being determined as, in some sense, the 'centre' of the sample is actually slightly more central to the sample than the population mean. Consequently the dispersion around the sample mean is slightly less then it would be about the population mean. A correction which slightly inflates the SD is therefore appropriate and is achieved by dividing by *n*-1 rather than *n*.

This description has considered various aspects of the formula for the sample SD. It has not covered why the quantity defined as the sample SD is the appropriate estimator of the parameter which governs the width of the Normal curve. It is, but a demonstration of this is beyond the scope of the present discussion.

Return to Document 2