Indices of Dispersion

We already know about *quartiles*, and there is not much to say about the *range* (the difference between the highest and the lowest value in a data set). The more interesting indices are the *variance* (and its square root, the *standard deviation*), and the *average deviation*:

Average Deviation

This index chooses an *index of position*, let's call it m, and computes the *mean of the absolute values of the differences between each data point and* m:

$$d = \frac{1}{n}(|x_1 - m| + |x_2 - m| + ... + |x_n - m|) = \frac{1}{n}\sum_{k=1}^{n} |x_k - m|$$

This is a fairly intuitive choice: we are averaging the distances between each data point and our chosen central position for our data set.

It turns out (as discussed in the file on connection between position and dispersion indices) that he "natural" choice for the position index, if you are bent on using this measure of dispersion, is the *median*.

Variance and Standard Deviation

This index chooses an *index of position*, let's call it m, and computes the *mean of the squares of the differences between each data point and* m:

$$v = \frac{1}{n}\Big[(x_1 - m)^2 + (x_2 - m)^2 + ... + (x_n - m)^2\Big] = \frac{1}{n}\sum_{k=1}^{n} (x_k - m)^2$$

The *standard deviation* is imply the square root of the variance: $s = \sqrt{v}$. Note that, if, say, the data represents distances measured in miles, the variance is measured in miles squared, while the standard deviation is measured in miles.

The logic behind averaging the *square* of the differences is manifold: on the one hand, handling squares in mathematical manipulations is much easier than handling absolute values, on the other, this choice enhances the effect of data far from the center, and depresses the effect of values close to the center, since, as is well known,

$$|x| < 1 \Longrightarrow x^2 < |x| \text{ while } |x| > 1 \Longrightarrow x^2 > |x|$$

This differentiation treats small deviations as not very important, while giving more prominence to large deviations.

However, the real interest in this index is, again, its prominent role in the mathematical theory of statistics.

It turns out (as discussed in the file about the connection between position and dispersion indices) that the "natural" choice for the position index, if you are bent on using this measure of dispersion, is the *mean*.

"Population" Variance and "Sample" Variance

The index defined given above is sometimes called "population variance", and is definitely the one relevant whenever you are summarizing a complete observation. As we will see later, **if your data set is to be used for further deductions about a larger population**, it turns out that there are some (weak) theoretical reasons, and some historical reasons, to choose to divide the sum of squares by $n-1$, instead of n. This index is then called the *sample variance*, defined as

$$\frac{1}{n-1}\sum_{k=1}^{n}(x_k-m)^2$$

Correspondingly, there is a *sample standard deviation*, of course. As we will see later, the gain in using this measure is minimal, but is justified, mostly, by the fact that many standard formulas assume you are using it (note that the sample variance is always larger than the population variance, indicating a more "pessimistic" – or more cautious, if you prefer – evaluation of how much your data is scattered). On the other hand, it is fairly obvious that for large data sets (the ones for which statistical analysis is most effective), the two indices are hardly different. In fact, in most cases, the precision with which your data is known will be such that the difference between these two "variances" is irrelevant.