The Variance

The *variance*, or "dispersion" of a set of values is

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$
$$= \text{var}(x).$$

In essence, the variance is the average squared deviation of the x_i's about the mean.

The units of the variance are those of the data squared, e.g. if the data are in degrees Celsius, then the variance has units Celsius-squared, units which may not be intuitively interpretable (or make sense physically).

A related measure, the simple average deviation

$$m = \frac{1}{n}\sum_{i-1}^{n}(x_i - \mu)$$

is always zero, as a consequence of one of the properties of the mean (that the sum of the deviations about the mean is zero). Consequently, the mean deviation is usually expressed as the mean of the absolute values of the observations

$$m = \frac{1}{n}\sum_{i=1}^{n}| x_i - \mu|.$$

For a number of reasons, the variance is preferred as a measure of dispersion over the mean deviation. In particular, the selection of the sum of squared deviations (variance) as a measure of variability is motivated because this choice gives the variance certain desirable properties that the average deviation does not have. In particular, because the squared of the deviation of each observation from the mean is used in calculating the variance, each of those squared deviations can be thought of as the *leverage* a particular observation has on the variance.

The *standard deviation*, is the square root of the variance

$$\sigma = \sqrt{\sigma^2} = (\text{var}(x))^{1/2}$$
$$= (\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2)^{1/2}.$$

The units of the standard deviation are the same as those of x, and this why the standard deviation is used so frequently.

In practice, the true mean, μ, is not known. Substituting the sample mean, $\bar{X}$, into the formulas for the variance and standard deviation yields the sample variance, s^2, and the sample standard deviation, s, where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2, \text{ and}$$

$$s = \sqrt{s^2} = (\text{var}(x))^{1/2}$$
$$= (\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2)^{1/2}.$$

Note that instead of multiplying the term $\sum_{i=1}^{n}(x_i - \bar{X})^2$ by $1/n$, for the sample variance and standard deviation, the value $1/(n\text{-}1)$ is used. This choice "adjusts" the sample variance and standard deviation for their tendencies to underestimate the "true" or population variance and standard deviation.