Some Basic Statistical Formulae

Suppose we are given a sample of n observations on some variable of interest. Label the observations $x_1, x_2, \ldots, x_n$. The basic statistical formulae we have considered can be written as follows:

$$\text{Mean} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\text{Variation} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\text{Variance} = s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

$$\text{Standard Deviation} = s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

Sometimes the data may be presented in a different way: we are given m distinct values of the variable of interest, $x_1, x_2, \ldots, x_m$, along with a set of m frequencies, $f_1, f_2, \ldots, f_m$, representing the number of occurrences of each of the values in the sample. In this case the formulae should be adjusted as below. (Note that $\sum_{i=1}^{m} f_i$, the sum of the frequencies, is equivalent to n in the first case above.)

$$\text{Mean} = \frac{\sum_{i=1}^{m} x_i f_i}{\sum_{i=1}^{m} f_i}$$

$$\text{Variation} = \sum_{i=1}^{m} (x_i - \bar{x})^2 f_i$$

$$\text{Variance} = \frac{\sum_{i=1}^{m} (x_i - \bar{x})^2 f_i}{\sum_{i=1}^{m} f_i}$$

$$\text{Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{m} (x_i - \bar{x})^2 f_i}{\sum_{i=1}^{m} f_i}}$$

Besides these basic stats, we have also spoken of the *standard error of the mean*, sometimes written as $s_{\bar{x}}$. This is a standard deviation of sorts, but it should not be confused with the "regular" standard deviation, s.

§ The regular standard deviation is a measure of the variability or dispersion of the data in a given sample.

§ The standard error of the mean is an estimate of the variability of dispersion of the sample mean, $\bar{x}$, if we were to draw repeated random samples of size n from a given population. Its formula is

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

The standard error of the mean gives us a measure of our uncertainty over the true or population mean, when we're using sample information. As a rule of thumb, we can have 95 percent confidence that the true mean lines within two standard errors of the sample mean, that is, in the interval $\bar{x} \pm 2s/\sqrt{n}$.