Standard Deviation

Imagine that we collect a large population of measurements of, say, the length of a given Cambrian trilobite species. The mean or average of those measurements is given by

$$\textbf{average} = (\text{value}_1 + \text{value}_2 + \ldots + \text{value}_N) / N$$

where N is the number of observations. This is not the only statistic that we might find useful. Was there a wide range of lengths, or were they all pretty much the same? What is the distribution of measurements on either side of the average? Actually, there are many questions one could ask about the data set.

One particularly useful and common statistic is the ***standard deviation***. In a normally distributed population of data (as the measurements of trilobite lengths might well be), approximately 68% of the measurements will be within one standard deviation larger or smaller than the average. The sample standard deviation (s) is given by:

$$\text{sample standard deviation} = s = \sqrt{\frac{\sum_{i=1}^{N}\left[(x_i - \bar{x})^2\right]}{N-1}}$$

For many, mathematical notation is approximately as clear as mud, so let me provide an *algorithm* (*i.e.*, a recipe) for computing the sample standard deviation based on the expression above.

1) Count the number of measurements you have, and call that number "N."
2) Calculate the average of those measurements, and call that number "$\bar{x}$."
3) Subtract $\bar{x}$ from each of the individual measurements (x_i) and square the result each time: $(x_i - \bar{x})^2$.
4) Add all of those squared results together: $\Sigma[(x_i - \bar{x})^2]$
5) Divide the result of step 4 by the quantity (N – 1).
6) Find the square root of the result of step 5.

The result of step 6 is the sample standard deviation.

Computing the standard deviation allows us to make an estimate of the likely error in the estimate of the average (mean) of a set of values, such as measurements of the length of a Cambrian trilobite species. That error estimate is made at a given confidence interval (CI), typically at a 95% CI indicating a high degree of confidence. To compute the 95% confidence interval (CI_{95}), we need to know how many observations have been averaged (N) and what their standard deviation is (s), and we need to read a number called ***Student's t factor*** from a table like the following (after Fisher, 1958):

No. of observations (N):	2	3	4	5	6
Student's t factor (t):	12.706	4.303	3.182	2.776	2.571

The 95% confidence interval (CI_{95}) around the average is:

$$95\% \text{ confidence interval} = CI_{95} = \frac{t * s}{\sqrt{N}}$$

From Cronin, 2004, Geology Laboratory Projects for Group Learning: Dubuque, Iowa, Primis McGraw-Hill, 186 p.