11.3.1: Mean and Variance

Suppose you have a set of values a_j. By saying that this is a set, we mean that we have several values a_1, a_2, a_3, and so forth. The notation a_j, in this context, means that j can be replaced by any integer between 1 and the total number of values that you have in order to refer to that specific value. Suppose that we have N total values. The average of all of our values can be written as:

$$\langle a\rangle = \frac{1}{N}\sum_j a_j \tag{11.10}$$

The letter Σ is the capital Greek letter "sigma". This notation means that you sum together all of the values of a_j that you have. For instance, suppose you had just four values, a_1, a_2, a_3, and a_4, then:

$$\sum_j a_j = a_1 + a_2 + a_3 + a_4 \tag{11.11}$$

Therefore, the mean (or average) value of a in this context is:

$$\langle a\rangle = \frac{1}{N}\sum_j a_j = \frac{1}{N}(a_1 + a_2 + a_3 + a_4) \tag{11.12}$$

To quantify the uncertainty on a set of values, we want to say something about how far, on average, a given value is from the mean of all the values. Thus, it's tempting to try to define the uncertainty as follows:

$$\frac{1}{N}\sum_j (a_j - \langle a\rangle) \tag{11.13}$$

Remember that addition is commutative. Realizing that the $\sum$ symbol just indicates a sum, i.e. a whole lot of addition, we can rewrite this as:

$$\frac{1}{N}\left(\sum_j a_j - \sum_j \langle a\rangle\right) \tag{11.14}$$

The second term in the subtraction is a sum over j of the average value. The average value doesn't depend on which a_j we're talking about; it's a constant, it's the same for all of them. Therefore, the sum of that number N times is just going to be equal to $N\langle a\rangle$. Making this substitution and distributing the $1/N$ into the parentheses:

$$\frac{1}{N}\sum_j a_j - \frac{1}{N}N\langle a\rangle \tag{11.15}$$

But we recognize the first term in this subtraction as just $\langle a\rangle$. So, the total result of this is zero. Clearly, this is not a good expression for the uncertainty in a. If you think about it, the average deviation of a_j from $\langle a\rangle$ ought to be zero. If $\langle a\rangle$ is the average value of a, then a_j should be below $\langle a\rangle$ about as often as it is above, so your sum will have a mix of positive and negative terms. The very definition of the average insures that this sum will be zero.

Instead, we shall define the variance as:

$$\Delta a^2 = \frac{1}{N}\sum_j (a_j - \langle a\rangle)^2 \tag{11.16}$$

Here, we're using Δa to indicate the uncertainty in a. The variance is defined as the uncertainty squared.[1] The advantage of this expression is that because we're squaring the difference between each value a_j and the average value, we're always going to be summing together positive terms; there will be no negative terms to cancel out the positive terms. Therefore, this should be a reasonable estimate of how far, typically, the measurements a_j are from their average.

We can unpack this sum a bit, first by multiplying out the squared polynomial:

$$\Delta^2 = \frac{1}{N}\sum_j \left(a_j^2 - 2\langle a\rangle a_j + \langle a\rangle^2\right) \tag{11.17}$$

In order to clean this expression up, inside the parentheses both add and subtract $\langle a\rangle^2$:

$$\begin{aligned}\Delta a^2 &= \frac{1}{N}\sum_j \left(a_j^2 - 2\langle a\rangle a_j + 2\langle a\rangle^2 - \langle a\rangle^2\right) \\ &= \frac{1}{N}\sum_j \left(a_j^2 - \langle a\rangle^2 + 2\langle a\rangle\left(\langle a\rangle - a_j\right)\right) \\ &= \frac{1}{N}\sum_j a_j^2 - \frac{1}{N}\sum_j \langle a\rangle^2 + \frac{1}{N}2\langle a\rangle\sum_j \left(\langle a\rangle - a_j\right)\end{aligned} \tag{11.18}$$

Notice that the last term is going to be zero, as it includes the average difference between the mean and each observation. The second term is just going to be $\langle a\rangle^2$, because once again $\langle a\rangle$ is the same for all terms of the sum; the sum will yield $N\langle a\rangle^2$, canceling the N in the denominator. So, we have:

$$\Delta a^2 = \left\langle a^2\right\rangle - \langle a\rangle^2 \tag{11.19}$$

[1]If you know statistics, you may recognizing this as being very similar to how variance is defined there— only in statistics, we divide by $N-1$ rather than by N. The difference becomes unimportant as N gets large.