Dec 4,2015

# [1] VARIANCE: AN ALTERNATIVE DEFINITION.

The customary definition of the variance of a small sample of N points is

$$s^2 = \frac{\sum_1^N (X_i - Mean)^2}{N-1}$$

where Mean is defined in the usual way as the sum of the $X_i$ divided by N.

This seems simple and straightforward except for the (N − 1) which means that the quantity so defined is not the average of the square of the difference of each data point from the Mean but a somewhat larger value. The following while yielding the same value for s makes the occurrence of (N − 1) obvious.

Let ($X_i$) be a sample of N data points having

$$Mean = \frac{\sum_1^N X_i}{N}$$

Define

$$r_i = X_i - Mean$$

Theorem: The sum of the

$$\sum_1^N r_i = 0$$

Proof trivial.

[1] © William Simon 2015

Now connect each of the data points to each of the others. Thus point one is connected to (N − 1) other points. Point number two is also connected to (N − 1) other points without regard to the fact that it has already been connected from point one in the previous step. (The apparent double counting of the connection between point one and two can be avoided but it complicates the rest of the following.)

This requires N ∗ (N − 1) connections.

Define the length of a connecting line from point i to point j as the absolute value of $\mathrm{X_i} - \mathrm{X_j}$.

We wish to find the sum of the squares of the lengths of these lines given by:

$$SumLengthSqs = \sum_{i=1}^{N} \sum_{j=1}^{N} (X_i - X_j)^2$$

Subtract the Mean from both $\mathrm{X_i}$ and $\mathrm{X_j}$.

$$SumLengthSqs = \sum_{i=1}^{N} \sum_{j=1}^{N} \ ((X_i - Mean) - (X_j - Mean))^2$$

which in terms of the $\mathrm{r_i}$ defined above becomes:

$$SumLengthSqs = \sum_{i=1}^{N} \sum_{j=1}^{N} \quad r_i^2 + r_j^2 - 2 \quad r_i \ r_j \qquad (1)$$

The double sum of product $\mathrm{r_i} * \mathrm{r_j}$ can be factored as follows:

$$r_1 * (r_1 + r_2 + r_3) +$$
$$r_2 * (r_1 + r_2 + r_3) +$$
$$r_3 * (r_1 + r_2 + r_3)$$

From the theorem at the beginning we find that each of the parenthesis sums to zero.

Thus equation 1 is reduced to

$$SumLengthSqs = \sum_{i=1}^{N} \sum_{j=1}^{N} \quad r_i^2 + r_j^2 =$$

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \quad r_i^2 + \sum_{i=1}^{N} \sum_{j=1}^{N} \quad r_j^2$$

Consider the first of the double sums above. Since there are no $\mathrm{r_j}$ in it , it reduces to the sum of N copies of the sum over i.

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \quad r_i^2 = N * \sum_{i=1}^{N} \quad r_i^2 = N * \sum_{i=1}^{N} \, (X_i - Mean)^2$$

The second sum is exactly the same except that the subscripts are reversed. Thus

$$SumLengthSqs = 2 * N * \sum_{i=1}^{N} \, (X_i - Mean)^2$$

Dividing by the number of connecting lines $\mathrm{N} * (\mathrm{N} - 1)$ yields the average squared length.

Average Of the Squares Of The Difference Between Points=

$$\frac{SumLengthSqs}{N*(N-1)} = \frac{2*\sum_{i=1}^{N}(X_i - M)^2}{(N-1)}$$

Thus if we regard Half of the Average Of the Square Of The Difference Between Points as an alternative definition of Variance (AltVar) we find we account for the (N − 1) in the usual definition of variance.

$$AltVar = \frac{\sum_{i=1}^{N}(X_i - M)^2}{(N-1)} =$$

Conventional Variance.

Now the N − 1 has lost its mystery.