Online Variance

by Joshua Burkholder

Let n be the number of values, v_n be the biased sample variance of the first n values, v_{n-1} be the biased sample variance for the first $n-1$ values, x_n be the n-th value, $\bar{x}_n$ be the sample mean of the first n values, and $\bar{x}_{n-1}$ be the sample mean of the first $n-1$ values. Then, the recurrence equation for the biased sample variance (a.k.a. online variance) is:

$$v_n = v_{n-1} - \frac{v_{n-1} - \left(x_n - \bar{x}_n\right)\left(x_n - \bar{x}_{n-1}\right)}{n}$$

Proof:

The definition of the biased sample variance of the first n values is defined as:

$$v_n = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^2}{n}$$

If we expand this definition, we have:

$$v_n = \frac{\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x}_n + \bar{x}_n^2\right)}{n}$$

$$v_n = \frac{\sum_{i=1}^{n-1}\left(x_i^2 - 2x_i\bar{x}_n + \bar{x}_n^2\right) + x_n^2 - 2x_n\bar{x}_n + \bar{x}_n^2}{n}$$

Since the recurrence equation for the sample mean is:

$$\bar{x}_n = \bar{x}_{n-1} - \frac{\bar{x}_{n-1} - x_n}{n},$$

then we also have:

$$\bar{x}_n^2 = \left(\bar{x}_{n-1} - \frac{\bar{x}_{n-1} - x_n}{n}\right)^2$$

$$\bar{x}_n^2 = \bar{x}_{n-1}^2 - 2\bar{x}_{n-1}\left(\frac{\bar{x}_{n-1} - x_n}{n}\right) + \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)^2$$

With these, we have:

$$v_n = \frac{\sum_{i=1}^{n-1}\left(x_i^2 - 2x_i\left(\bar{x}_{n-1} - \frac{\bar{x}_{n-1} - x_n}{n}\right) + \left(\bar{x}_{n-1}^2 - 2\bar{x}_{n-1}\left(\frac{\bar{x}_{n-1} - x_n}{n}\right) + \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)^2\right)\right) + x_n^2 - 2x_n\bar{x}_n + \bar{x}_n^2}{n}$$

$$v_n = \frac{\sum_{i=1}^{n-1}\left(x_i^2 - 2x_i\bar{x}_{n-1} + 2x_i\left(\frac{\bar{x}_{n-1} - x_n}{n}\right) + \bar{x}_{n-1}^2 - 2\bar{x}_{n-1}\left(\frac{\bar{x}_{n-1} - x_n}{n}\right) + \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)^2\right) + x_n^2 - 2x_n\bar{x}_n + \bar{x}_n^2}{n}$$

$$v_n = \frac{\sum_{i=1}^{n-1}\left(x_i^2 - 2x_i\bar{x}_{n-1} + \bar{x}_{n-1}^2 + 2x_i\left(\frac{\bar{x}_{n-1} - x_n}{n}\right) - 2\bar{x}_{n-1}\left(\frac{\bar{x}_{n-1} - x_n}{n}\right) + \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)^2 \right) + x_n^2 - 2x_n\bar{x}_n + \bar{x}_n^2}{n}$$

$$v_n = \frac{\sum_{i=1}^{n-1}\left(\left(x_i - \bar{x}_{n-1}\right)^2 + 2x_i\left(\frac{\bar{x}_{n-1} - x_n}{n}\right) - 2\bar{x}_{n-1}\left(\frac{\bar{x}_{n-1} - x_n}{n}\right) + \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)^2 \right) + x_n^2 - 2x_n\bar{x}_n + \bar{x}_n^2}{n}$$

$$v_n = \frac{\sum_{i=1}^{n-1}\left(x_i - \bar{x}_{n-1}\right)^2 + \sum_{i=1}^{n-1}\left(2x_i\left(\frac{\bar{x}_{n-1} - x_n}{n}\right) - 2\bar{x}_{n-1}\left(\frac{\bar{x}_{n-1} - x_n}{n}\right) + \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)^2 \right) + x_n^2 - 2x_n\bar{x}_n + \bar{x}_n^2}{n}$$

Since the biased sample variance for the first $n-1$ values is:

$$v_{n-1} = \frac{\sum_{i=1}^{n-1}\left(x_i - \bar{x}_{n-1}\right)^2}{n-1},$$

then we also have:

$$\sum_{i=1}^{n-1}\left(x_i - \bar{x}_{n-1}\right)^2 = (n-1)v_{n-1}.$$

With this, we have:

$$v_n = \frac{(n-1)v_{n-1} + \sum_{i=1}^{n-1}\left(2x_i\left(\frac{\bar{x}_{n-1} - x_n}{n}\right) - 2\bar{x}_{n-1}\left(\frac{\bar{x}_{n-1} - x_n}{n}\right) + \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)^2 \right) + x_n^2 - 2x_n\bar{x}_n + \bar{x}_n^2}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)\left(\sum_{i=1}^{n-1}\left(2x_i - 2\bar{x}_{n-1} + \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)\right)\right) + x_n^2 - 2x_n\bar{x}_n + \bar{x}_n^2}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)\left(\sum_{i=1}^{n-1}(2x_i) + \sum_{i=1}^{n-1}(-2\bar{x}_{n-1}) + \sum_{i=1}^{n-1}\left(\frac{\bar{x}_{n-1} - x_n}{n}\right)\right) + x_n^2 - 2x_n\bar{x}_n + \bar{x}_n^2}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)\left(2\sum_{i=1}^{n-1}(x_i) - 2\bar{x}_{n-1}\sum_{i=1}^{n-1}(1) + \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)\sum_{i=1}^{n-1}(1)\right) + x_n^2 - 2x_n\bar{x}_n + \bar{x}_n^2}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)\left(2\sum_{i=1}^{n-1}(x_i) - 2\bar{x}_{n-1}(n-1) + \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)(n-1)\right) + x_n^2 - 2x_n\bar{x}_n + \bar{x}_n^2}{n}$$

Since the definition of the sample mean for the first $n-1$ values is:

$$\bar{x}_{n-1} = \frac{\sum_{i=1}^{n-1} x_i}{n-1},$$

then we also have:

$$\sum_{i=1}^{n-1} x_i = (n-1)\bar{x}_{n-1}.$$

With this, we have:

$$v_n = \frac{(n-1)v_{n-1} + \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)\left(2(n-1)\bar{x}_{n-1} - 2\bar{x}_{n-1}(n-1) + \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)(n-1)\right) + x_n^2 - 2x_n\bar{x}_n + \bar{x}_n^2}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)\left(\cancel{2(n-1)\bar{x}_{n-1}} \cancel{-2(n-1)\bar{x}_{n-1}} + \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)(n-1)\right) + x_n^2 - 2x_n\bar{x}_n + \bar{x}_n^2}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)\left(\frac{\bar{x}_{n-1} - x_n}{n}\right)(n-1) + x_n^2 - 2x_n\bar{x}_n + \bar{x}_n^2}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + x_n^2 + (n-1)\left(\frac{\bar{x}_{n-1}^2 - 2x_n\bar{x}_{n-1} + x_n^2}{n^2}\right) - 2x_n\bar{x}_n + \bar{x}_n^2}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + x_n^2 + \frac{(n-1)\bar{x}_{n-1}^2}{n^2} - \frac{2(n-1)x_n\bar{x}_{n-1}}{n^2} + \frac{(n-1)x_n^2}{n^2} - 2x_n\bar{x}_n + \bar{x}_n^2}{n}$$

Since the recurrence equation for the sample mean is:

$$\bar{x}_n = \bar{x}_{n-1} - \frac{\bar{x}_{n-1} - x_n}{n},$$

then we also have:

$$\bar{x}_n = \frac{n\bar{x}_{n-1} - \bar{x}_{n-1} + x_n}{n}$$

$$\bar{x}_n = \frac{(n-1)\bar{x}_{n-1} + x_n}{n}$$

Moreover, we have:

$$\bar{x}_n^2 = \left(\frac{(n-1)\bar{x}_{n-1} + x_n}{n} \right)^2$$

$$\bar{x}_n^2 = \frac{\left((n-1)\bar{x}_{n-1} + x_n\right)^2}{n^2}$$

$$\bar{x}_n^2 = \frac{(n-1)^2 \bar{x}_{n-1}^2 + 2(n-1)x_n\bar{x}_{n-1} + x_n^2}{n^2}$$

$$\bar{x}_n^2 = \frac{(n-1)^2 \bar{x}_{n-1}^2}{n^2} + \frac{2(n-1)x_n\bar{x}_{n-1}}{n^2} + \frac{x_n^2}{n^2}$$

With this, we have:

$$v_n = \frac{(n-1)v_{n-1} + x_n^2 + \frac{(n-1)\bar{x}_{n-1}^2}{n^2} - \frac{2(n-1)x_n\bar{x}_{n-1}}{n^2} + \frac{(n-1)x_n^2}{n^2} - 2x_n\bar{x}_n + \left(\frac{(n-1)^2 \bar{x}_{n-1}^2}{n^2} + \frac{2(n-1)x_n\bar{x}_{n-1}}{n^2} + \frac{x_n^2}{n^2} \right)}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + x_n^2 + \frac{(n-1)\bar{x}_{n-1}^2}{n^2} \cancel{- \frac{2(n-1)x_n\bar{x}_{n-1}}{n^2}} + \frac{(n-1)x_n^2}{n^2} - 2x_n\bar{x}_n + \frac{(n-1)^2 \bar{x}_{n-1}^2}{n^2} \cancel{+ \frac{2(n-1)x_n\bar{x}_{n-1}}{n^2}} + \frac{x_n^2}{n^2}}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + x_n^2 + \frac{(n-1)^2 \bar{x}_{n-1}^2}{n^2} + \frac{(n-1)\bar{x}_{n-1}^2}{n^2} + \frac{(n-1)x_n^2}{n^2} + \frac{x_n^2}{n^2} - 2x_n\bar{x}_n}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + x_n^2 + \left((n-1)^2 + (n-1)\right)\left(\frac{\bar{x}_{n-1}^2}{n^2} \right) + \left((n-1)+1\right)\left(\frac{x_n^2}{n^2} \right) - 2x_n\bar{x}_n}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + x_n^2 + \left((n-1)\left((n-1)+1\right)\right)\left(\frac{\bar{x}_{n-1}^2}{n^2} \right) + (n)\left(\frac{x_n^2}{n^2} \right) - 2x_n\bar{x}_n}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + x_n^2 + \left((n-1)(n)\right)\left(\frac{\bar{x}_{n-1}^2}{n^2} \right) + (n)\left(\frac{x_n^2}{n^2} \right) - 2x_n\bar{x}_n}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + x_n^2 + (n-1)\left(\frac{\bar{x}_{n-1}^2}{n} \right) + \frac{x_n^2}{n} - 2x_n\bar{x}_n}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + x_n^2 - x_n\bar{x}_n + \frac{(n-1)\bar{x}_{n-1}^2}{n} + \frac{x_n^2}{n} - x_n\bar{x}_n}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + x_n^2 - x_n\bar{x}_n + \frac{(n-1)\bar{x}_{n-1}^2}{n} - \frac{nx_n\bar{x}_n}{n} + \frac{x_n^2}{n}}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + x_n^2 - x_n\bar{x}_n + \frac{(n-1)\bar{x}_{n-1}^2}{n} - \frac{nx_n\bar{x}_n - x_n^2}{n}}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + x_n^2 - x_n\bar{x}_n + \frac{(n-1)\bar{x}_{n-1}^2}{n} - \frac{(n\bar{x}_n - x_n)x_n}{n}}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + x_n^2 - x_n\bar{x}_n + \frac{(n-1)\bar{x}_{n-1}^2}{n} - \left(\frac{n-1}{n-1}\right)\left(\frac{(n\bar{x}_n - x_n)x_n}{n}\right)}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + x_n^2 - x_n\bar{x}_n + \frac{(n-1)\bar{x}_{n-1}^2}{n} - \left(\frac{(n-1)x_n}{n}\right)\left(\frac{n\bar{x}_n - x_n}{n-1}\right)}{n}$$

As previously noted, the recurrence equation for the sample mean can be rewritten as:

$$\bar{x}_n = \frac{(n-1)\bar{x}_{n-1} + x_n}{n},$$

then we have:

$$\frac{(n-1)\bar{x}_{n-1} + x_n}{n} = \bar{x}_n$$

$$(n-1)\bar{x}_{n-1} + x_n = n\bar{x}_n$$

$$(n-1)\bar{x}_{n-1} = n\bar{x}_n - x_n$$

$$\bar{x}_{n-1} = \frac{n\bar{x}_n - x_n}{n-1}$$

With this, we have:

$$v_n = \frac{(n-1)v_{n-1} + x_n^2 - x_n\bar{x}_n + \frac{(n-1)\bar{x}_{n-1}^2}{n} - \left(\frac{(n-1)x_n}{n}\right)(\bar{x}_{n-1})}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + x_n^2 - x_n\bar{x}_n + \left(\frac{(n-1)\bar{x}_{n-1}}{n} - \left(\frac{nx_n - x_n}{n}\right)\right)(\bar{x}_{n-1})}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + x_n^2 - x_n\bar{x}_n + \left(\frac{(n-1)\bar{x}_{n-1}}{n} - \frac{\cancel{n}x_n}{\cancel{n}} + \frac{x_n}{n}\right)(\bar{x}_{n-1})}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + x_n^2 - x_n\bar{x}_n + \left(\left(\frac{(n-1)\bar{x}_{n-1} + x_n}{n}\right) - x_n\right)(\bar{x}_{n-1})}{n}$$

Since the recurrence equation of the sample mean can be rewritten as:

$$\bar{x}_n = \frac{(n-1)\bar{x}_{n-1} + x_n}{n},$$

then we have:

$$v_n = \frac{(n-1)v_{n-1} + x_n^2 - x_n\bar{x}_n + (\bar{x}_n - x_n)(\bar{x}_{n-1})}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + x_n^2 - x_n\bar{x}_n + \bar{x}_n\bar{x}_{n-1} - x_n\bar{x}_{n-1}}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + x_n^2 - x_n\bar{x}_n - x_n\bar{x}_{n-1} + \bar{x}_n\bar{x}_{n-1}}{n}$$

$$v_n = \frac{(n-1)v_{n-1} + (x_n - \bar{x}_n)(x_n - \bar{x}_{n-1})}{n}$$

$$v_n = \frac{nv_{n-1} - v_{n-1} + (x_n - \bar{x}_n)(x_n - \bar{x}_{n-1})}{n}$$

$$v_n = \frac{\cancel{n}v_{n-1}}{\cancel{n}} + \frac{-v_{n-1} + (x_n - \bar{x}_n)(x_n - \bar{x}_{n-1})}{n}$$

$$v_n = v_{n-1} - \frac{v_{n-1} - (x_n - \bar{x}_n)(x_n - \bar{x}_{n-1})}{n}$$

Therefore, the recurrence equation for the biased sample variance (a.k.a. online variance) is:

$$v_n = v_{n-1} - \frac{v_{n-1} - (x_n - \bar{x}_n)(x_n - \bar{x}_{n-1})}{n}$$

Reference:

http://en.wikipedia.org/wiki/Algorithms_for_calculating_variance

Example C++ code that computes the online variance:

```
// Filename: main.cpp
#include <iostream>
#include <iomanip>

int main () {

    double x;
    double n = 0;
    double mean = 0;
    double variance = 0;
    double prev_mean; // previous mean
    double prev_variance; // previous variance

    if ( std::cin >> x ) {
        ++n;
        mean = x;
        variance = 0;
        while ( std::cin >> x ) {
            prev_mean = mean;
            prev_variance = variance;
            ++n;
            mean = prev_mean - ( prev_mean - x ) / n;
            variance = prev_variance - ( prev_variance - ( x - mean ) * ( x - prev_mean ) ) / n;
        }
    }

    std::cout << "n:        " << n << '\n';
    std::cout << "mean:     " << std::setprecision( 17 ) << mean     << '\n';
    std::cout << "variance: " << std::setprecision( 17 ) << variance << '\n';

}
```

Example of `data.txt`:

```
6867.55961097
32890.8902819
18178.8157597
.
.
.
```

Command Line:

```
g++ -o main.exe main.cpp -std=c++11 -march=native -O3 -Wall -Wextra -Werror –static
./main.exe < data.txt
```

Note: Mathematica's `Variance[]` function computes the ***unbiased*** sample variance, not the ***biased*** sample variance; therefore, the biased sample variance is computed in Mathematica as:

```
( ( Length[ list ] – 1 ) / Length[ list ] ) * Variance[ list ]
```