Online Variance

by Joshua Burkholder

Let  be the number of values,  be the biased sample variance of the first  values,  be the biased sample variance for the first  values,  be the -th value,  be the sample mean of the first  values, and  be the sample mean of the first  values. Then, the recurrence equation for the biased sample variance (a.k.a. online variance) is:



Proof:

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



If we expand this definition, we have:



Since the recurrence equation for the sample mean is:

,

then we also have:



With these, we have:





Since the biased sample variance for the first  values is:

,

then we also have:

.

With this, we have:



Since the definition of the sample mean for the first  values is:

,

then we also have:

.

With this, we have:



Since the recurrence equation for the sample mean is:

,

then we also have:



Moreover, we have:



With this, we have:





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

,

then we have:



With this, we have:



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

,

then we have:



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



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 ]