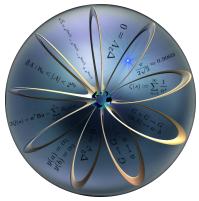

planetmath.org

Math for the people, by the people.

one-pass algorithm to compute sample variance

Canonical name	OnepassAlgorithmToComputeSampleVariance
Date of creation	2013-03-22 16:45:19
Last modified on	2013-03-22 16:45:19
Owner	stevecheng (10074)
Last modified by	stevecheng (10074)
Numerical id	9
Author	stevecheng (10074)
Entry type	Algorithm
Classification	msc 68W01
Classification	msc 65-00
Classification	msc 62-00

In many situations it is desirable to calculate, in one iteration, the sample variance of many numbers, and without having to have every number available in computer memory before beginning processing.

Let $x_1, x_2, \ldots$ denote the data. The naïve formula for calculating the sample variance in one pass,

$$v = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{(\sum_{i=1}^{n} x_i^2) - n\overline{x}^2}{n-1}\,, \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,,$$

suffers from computational round-off error. If the mean $\overline{x}$ is large in absolute value, and $\sum_{i=1}^{n} x_i^2$ is close to $n\overline{x}^2$, then the subtraction at the end will tend to lose significant digits on the result. Also, in rare cases, the sum of squares $\sum_{i=1}^{n} x_i^2$ can overflow on a computer.

A better alternative, though requiring more work per iteration, is to calculate the running sample mean and variance instead, and update these as each datum is processed. Here we give the derivation of the one-pass algorithm — which involves nothing more than simple algebraic manipulations.

Define the running arithmetic mean and the sum of squared residuals:

$$a_n = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \quad s_n = \sum_{i=1}^{n}(x_i - a_n)^2\,.$$

We want to express a_{n+1} and s_{n+1} in terms of the old values a_n and s_n.

For convenience, let $\delta = x_{n+1} - a_n$ and $\gamma = a_{n+1} - a_n$. Then we have

$$a_{n+1} = \frac{na_n + x_{n+1}}{n+1} = \frac{(n+1)a_n + x_{n+1} - a_n}{n+1} = a_n + \frac{\delta}{n+1}\,.$$

For the variance calculation, we have

$$\begin{aligned} s_{n+1} &= \sum_{i=1}^{n}\big((x_i - a_n) - \gamma\big)^2 + (x_{n+1} - a_{n+1})^2 \\ &= \sum_{i=1}^{n}(x_i - a_n)^2 - 2\gamma\sum_{i=1}^{n}(x_i - a_n) + \sum_{i=1}^{n}\gamma^2 + (x_{n+1} - a_{n+1})^2 \\ &= s_n + 0 + n\gamma^2 + (x_{n+1} - a_{n+1})^2\,. \end{aligned}$$

Now observe:

$$\gamma = \frac{\delta}{n+1}\,, \quad x_{n+1} - a_{n+1} = \delta - \gamma = (n+1)\gamma - \gamma = n\gamma\,;$$

hence we obtain:

$$\begin{aligned} s_{n+1} = s_n + n\gamma^2 + n^2\gamma^2 = s_n + n(n+1)\gamma^2 &= s_n + n\gamma\delta \\ &= s_n + (x_{n+1} - a_{n+1})\delta\,. \end{aligned}$$

Note that the number to be added to s_n is never negative, so no cancellation error will occur from this procedure. (However, there can still be computational round-off error if $s_{n+1} - s_n$ happens to be very small compared to s_n.)

The recurrence relation for the sample covariance of two lists of numbers $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$ can be derived similarly. If a_n and b_n denote the arithmetic means of first n numbers of each of the two lists respectively, then the sum of adjusted products

$$c_n = \sum_{i=1}^{n} (x_i - a_n)(y_i - b_n)$$

can be incrementally updated by

$$c_{n+1} = c_n + (y_{n+1} - b_{n+1})\,(x_{n+1} - a_n) = c_n + (x_{n+1} - a_{n+1})\,(y_{n+1} - b_n)\,.$$

References

[1] B. P. Welford. "Note on a Method for Calculating Corrected Sums of Squares and Products". *Technometrics*, Vol. 4, No. 3 (Aug., 1962), p. 419-420.

[2] "`http://en.wikipedia.org/wiki/Algorithms_for_calculating_variance`Algorithms for calculating variance". *Wikipedia, The Free Encyclopedia*. Accessed 25 February 2007.