Welford's Equations

September 25, 2025

Computational Considerations

- Consider the sample standard deviation equation

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Requires two passes through the data
  1. Compute the mean $\bar{x}$
  2. Compute the squared differences about $\bar{x}$
- Must store or re-create the entire sample — bad when n is large

The Conventional One-Pass Algorithm

- A mathematically equivalent, one-pass equation for s^2:

$$
\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2\bar{x}x_i + \bar{x}^2) \\
&= \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - \left(\frac{2}{n}\bar{x}\sum_{i=1}^{n}x_i\right) + \left(\frac{1}{n}\sum_{i=1}^{n}\bar{x}^2\right) \\
&= \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - 2\bar{x}^2 + \bar{x}^2 \\
&= \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - \bar{x}^2
\end{aligned}
$$

- Round-off error is problematic

Welford's One Pass $\bar{x}$, s^2
(derivation)

Welford's One-Pass Algorithm

- Running sample mean:

$$\bar{x}_i = \frac{1}{i}(x_1 + x_2 + \cdots + x_i)$$

- Running sample sum of squared deviations:

$$v_i = (x_1 - \bar{x}_i)^2 + (x_2 - \bar{x}_i)^2 + \cdots + (x_i - \bar{x}_i)^2$$

- $\bar{x}_i$ and v_i can be computed recursively ($\bar{x}_0 = 0, v_0 = 0$) (Theorem 4.1.2):

$$\begin{aligned} \bar{x}_i &= \bar{x}_{i-1} + \frac{1}{i}(x_i - \bar{x}_{i-1}) \\ v_i &= v_{i-1} + \left(\frac{i-1}{i}\right)(x_i - \bar{x}_{i-1})^2 \end{aligned}$$

- Then $\bar{x}_n$ is the sample mean, v_n/n is the variance

Welford Advantages

- Don't need to know n ahead of time
- Better numerical performance: we are adding up squares of differences that should be of approximately the same magnitude.

Welford Warnings (against a common misconception)

Welford's equations are iterative, they are not convergent or asymptotic toward the true $\bar{x}$ and s^2

Which means: if the true mean of $n = 1000$ data points is $\hat{x}$,
$\bar{x}_{950}$, $\bar{x}_{975}$ or even $\bar{x}_{999}$ is not guaranteed to be particularly close to, consistently larger, or consistently smaller than $\hat{x}$ (they are simply the true mean of the first 950, 975 and 999 data points).

But $\bar{x}_{1000} = \hat{x}$ and $v_{1000} = 1000 \cdot s^2$ is guaranteed.

So: don't consider "recent" intermediate values better estimates for statistics of the full data set.

(Not a real loss compared to the other techniques: you need data point x_n to determine $\hat{x}$ and s^2 for all the methods.)

Time-Averaged Sample Statistics

- Let $x(t)$ be the sample path of a stochastic process for $0 < t < \tau$
- *Sample-path mean*:

$$\bar{x} = \frac{1}{\tau} \int_0^\tau x(t)\, dt$$

- *Sample-path variance*:

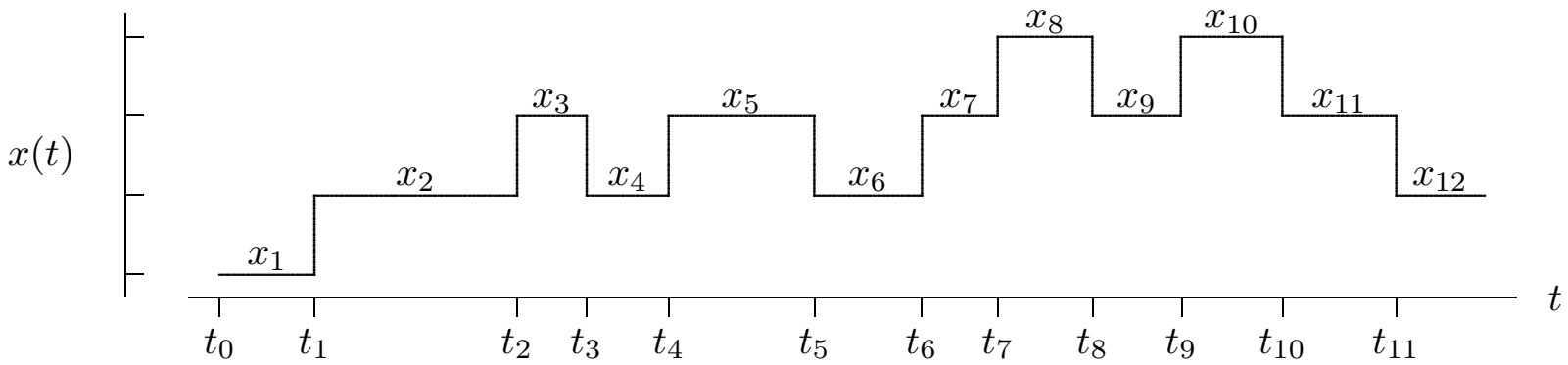
$$s^2 = \frac{1}{\tau} \int_0^\tau \left(x(t) - \bar{x}\right)^2 dt$$

- *Sample-path standard deviation*: $s = \sqrt{s^2}$
- One-pass equation for variance:

$$s^2 = \left(\frac{1}{\tau} \int_0^\tau x^2(t)\, dt\right) - \bar{x}^2$$

Computational Considerations

- For DES, a sample path is *piecewise constant*
- Changes in the sample path occur at *event times* $t_0, t_1, \ldots$



- For computing statistics, integrals reduce to summations

Computational Sample-Path Formulas

Theorem (4.1.3)

Consider a piecewise constant sample path

$$x(t) = \begin{cases} x_1 & t_0 < t \leq t_1 \\ x_2 & t_1 < t \leq t_2 \\ \vdots & \vdots \\ x_n & t_{n-1} < t \leq t_n \end{cases}$$

- *Sample-path mean:*

$$\bar{x} = \frac{1}{\tau} \int_0^\tau x(t)\, dt = \frac{1}{t_n} \sum_{i=1}^{n} x_i\, \delta_i$$

- *Sample-path variance:*

$$s^2 = \frac{1}{\tau} \int_0^\tau \left(x(t) - \bar{x}\right)^2 dt = \frac{1}{t_n} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 \delta_i = \left(\frac{1}{t_n} \sum_{i=1}^{n} x_i^2\, \delta_i\right) - \bar{x}^2$$

Welford's Sample Path Algorithm

- Based on the definitions

$$\begin{aligned} \bar{x}_i &= \frac{1}{t_i}(x_1\delta_1 + x_2\delta_2 + \cdots + x_i\delta_i) \\ v_i &= (x_1 - \bar{x}_i)^2\delta_1 + (x_2 - \bar{x}_i)^2\delta_2 + \cdots + (x_i - \bar{x}_i)^2\delta_i \end{aligned}$$

- $\bar{x}_i$ is the sample-path mean of $x(t)$ for $t_0 \le t \le t_i$
- v_i/t_i is the sample-path variance
- $\bar{x}_i$ and v_i can be computed recursively ($\bar{x}_0 = 0, v_0 = 0$) (Theorem 4.1.4):

$$\begin{aligned} \bar{x}_i &= \bar{x}_{i-1} + \frac{\delta_i}{t_i}(x_i - \bar{x}_{i-1}) \\ v_i &= v_{i-1} + \frac{\delta_i t_{i-1}}{t_i}(x_i - \bar{x}_{i-1})^2 \end{aligned}$$

Outliers : 0

No such thing! Quoth the text (page 152, emphasis by book author):

> *Generally, with simulation generated data, there should not be any outliers. An outlier is a data point that some omniscient being considers to be so different from the rest that it should be excluded from any statistical analysis. Although outliers are common with some kinds of experimentally measured data, it is difficult to argue that* ***any*** *data generated by a* ***valid*** *disrete-event simulation program is an outlier, no matter how unlikely the value may appear to be.*

I see no reason this rule does not apply to **any** type of simulation generated data.

Unfortunately, the author also uses the term **outlier** in algorithms to mean "out of bounds value" — as in array index out of bounds having to do with memory allocation issues. Keep this in mind in your reading.