

**Research  
1.  Find the simplest and most elegant way to show the Welford recursion.  
(Winner will have prize 😊 )**

**Technical Explanation**

In 1962, Welford (later presented by Donald Knuth) introduced in his paper an approach based on the idea of iteratively calculating the running variance and running mean, updating the result at each step and adding a new observation each time (hence the term "running"). This algorithm is particularly useful when data is collected without enough space to store all the values or when memory access costs outweigh the computational costs.

* : the current observation;
* n : the mean relative to the first observations;
* 2n : the variance relative to the first observations;
* 2n : the variance relative to the reference population.

The recursive formulas referenced by Welford for updating the parameters are as follows:

Immagine che contiene Carattere, linea, calligrafia, diagramma

Descrizione generata automaticamente

where we denote by:

Immagine che contiene testo, Carattere, schermata, linea

Descrizione generata automaticamente

We also define 2n as the sum of the squares of deviations from the current mean. In this case, we obtain the following recursive formulas:

(N.B per me: M2,n è la Media Ricorsiva, S2n è la varianza campionaria, σ2N è la varianza popolazionale.)

Immagine che contiene testo, Carattere, calligrafia, linea

Descrizione generata automaticamente

These formulas can be prone to numerical instability, which occurs when a very small number is subtracted recursively from a very large number.

The algorithmic structure of the method, in which the running mean is calculated, is given by the following pseudo-code:

**def** online\_variance(data):   
 n = 0   
 mean = 0   
 **for** x **in** data:   
 n = n + 1   
 delta = x - mean   
 mean = mean + delta/n   
 **return** mean

“Running” Algorithm for Variance Calculation:

**def** online\_variance(data):   
 n = 0   
 mean = 0   
 M2 = 0   
 **for** x **in** data:   
 n = n + 1   
 delta = x - mean   
 mean = mean + delta/n   
 M2 = M2 + delta\*(x - mean)   
 variance = M2/(n - 1)   
 **return** variance

This approach helps resolve the problem of precision loss, caused by the so-called **Catastrophic Cancellation**: this occurs, for example, when subtracting two numbers that are very close to each other. The operation essentially returns an unacceptable relative error that is much larger than the absolute error of the input values. This algorithm allows for the real-time calculation and updating of the mean and variance as new values are added to the data set, enabling us to avoid this type of computational issue that leads to information loss related to significant data.

**Explanation for a Child**

Imagine you have a box where you put your toys. Every time you add a new toy, you want to know what the average weight of your toys is and how different their weights are from each other.

Here’s how Welford’s method works:

1. **Start with zero**: Before you have any toys, both the average weight and how different the weights are (we call this "diversity") are zero.
2. **Add a toy**: Each time you put a new toy in the box:
   * **Count the toys**: You add one to the total number of toys.
   * **Calculate the new average**: You take the weight of the new toy and compare it to the old average. Then you adjust the average based on this new toy’s weight.
   * **Calculate the diversity**: You figure out how much the new toy’s weight is different from the new average and add that to your "diversity."
3. **Repeat**: You keep doing this every time you add a toy.

In the end, you’ll know the average weight of all your toys and how different their weights are from each other!

So, Welford’s method is a clever way to keep track of the average and the diversity of your toys’ weights each time you add a new one!

**Research  
Make your personal notes about the behavior of mean and variance wrt to time.  
What did you observe in all the 4 different cases (relative/abs freq & Bernoulli/random walk) ?**

**Introduction**

Understanding the behavior of mean and variance over time is fundamental to many areas of statistical analysis, especially in real-time data processing and time series analysis. This study focuses on how these two key statistics evolve in four different cases:

1. **Absolute Frequency**: Counting how often events occur.
2. **Relative Frequency**: The proportion of occurrences of an event over time.
3. **Bernoulli Trials**: A sequence of binary (success/failure) random events.
4. **Random Walk**: A stochastic process where each step depends on the previous one and a random factor.

**Problem Statement**

How do mean and variance behave over time in these cases? Understanding this can help in fields like finance (random walk in stock prices), quality control (Bernoulli trials in defect detection), and machine learning (frequency of event occurrence).

**Methodology**

**1. Welford's Algorithm for Running Statistics**

To track mean and variance efficiently, we use Welford's algorithm. This recursive algorithm allows us to update both mean and variance incrementally as new data points are added, avoiding the need to store all data points.

The running mean n after data points is updated as:

The running mean xˉn\bar{x}\_nxˉn​ after nnn data points is updated as:

xˉn=xˉn−1+xn−xˉn−1n\bar{x}\_n = \bar{x}\_{n-1} + \frac{x\_n - \bar{x}\_{n-1}}{n}xˉn​=xˉn−1​+nxn​−xˉn−1​​

The running variance is updated using:

M2,n=M2,n−1+(xn−xˉn−1)(xn−xˉn)M\_{2,n} = M\_{2,n-1} + (x\_n - \bar{x}\_{n-1})(x\_n - \bar{x}\_n)M2,n​=M2,n−1​+(xn​−xˉn−1​)(xn​−xˉn​)

where M2,nM\_{2,n}M2,n​ is the sum of squares of differences from the mean. The variance is then:

σn2=M2,nn\sigma\_n^2 = \frac{M\_{2,n}}{n}σn2​=nM2,n​​

**2. Cases Considered**

We consider four types of processes:

* **Absolute Frequency**: Data is generated by counting how often an event happens in a given time frame.
* **Relative Frequency**: Here we track the proportion of successes (e.g., a coin flip landing heads) over time.
* **Bernoulli Trials**: Binary outcomes (e.g., heads or tails, success or failure) where each trial is independent.
* **Random Walk**: A path in which each step is determined by a random variable, often modeled by adding random noise to a previous value.

For each process, we simulate data and observe how mean and variance behave as more data is added.

**3. Simulation Setup**

We conducted simulations for all four cases using the following parameters:

* **Data Points**: Each simulation ran for 1000 iterations.
* **Initial Conditions**: For random walk and Bernoulli processes, initial values were zero.

**Results and Observations**

**1. Absolute Frequency**

* **Mean**: The mean in the absolute frequency case is expected to increase linearly with the number of events. As we tally up events, the mean tends to grow without bound since we're simply counting occurrences.
* **Variance**: The variance grows as well, reflecting the increasing spread of the count as more events are added. However, due to the deterministic nature of this process (pure counting), the growth in variance is more predictable.

**2. Relative Frequency**

* **Mean**: The mean of relative frequency tends to stabilize over time. For example, if we're flipping a fair coin, the proportion of heads approaches 0.5 as the number of trials increases. Early fluctuations occur due to small sample sizes, but with time, the mean converges toward the theoretical probability.
* **Variance**: The variance in relative frequency decreases over time as we gather more data. This is because the relative frequency stabilizes as the sample size grows larger, reducing the variability in the proportion of successes or failures.

**3. Bernoulli Trials**

* **Mean**: In Bernoulli trials, the mean represents the probability of success over time. With many trials, the mean approaches the true probability ppp, assuming ppp is fixed. Early data might show more variability, but eventually, the mean converges.
* **Variance**: The variance in Bernoulli trials initially fluctuates as we accumulate successes and failures. However, over time, the variance converges to a fixed value based on the formula σ2=p(1−p)\sigma^2 = p(1-p)σ2=p(1−p). With enough trials, the variability in the outcome distribution stabilizes.

**4. Random Walk**

* **Mean**: The mean in a random walk can drift significantly over time, especially if there is a bias in the random steps. However, with unbiased random walks (equal probability of stepping up or down), the mean hovers around zero, reflecting the random nature of the process.
* **Variance**: The variance in a random walk grows without bound as time progresses. This happens because each step adds more uncertainty to the position. For unbiased random walks, the variance grows linearly with time, reflecting the increasing uncertainty about the walker’s position.

**Discussion**

**Mean Behavior**

In all four cases, the behavior of the mean with respect to time showed distinct patterns:

* In **absolute frequency**, the mean increased linearly as more events occurred.
* For **relative frequency**, **Bernoulli trials**, and **random walks**, the mean exhibited convergence behavior. In **relative frequency** and **Bernoulli trials**, the mean approached a fixed value (e.g., the true probability). In **random walks**, the mean remained close to zero, especially for unbiased walks, though it fluctuated more than in other cases.

**Variance Behavior**

Variance showed more varied behavior across the cases:

* In **absolute frequency**, variance grew steadily over time as counts accumulated.
* In **relative frequency** and **Bernoulli trials**, variance decreased or stabilized as the sample size increased, reflecting the decreasing uncertainty in the estimate of the mean.
* For **random walks**, variance grew without limit, reflecting the increasing spread of positions over time. This is characteristic of stochastic processes where variability accumulates with each step.

**Implications**

Understanding the differences in mean and variance behavior across these processes has practical applications:

* In **relative frequency** and **Bernoulli trials**, a larger sample size leads to more reliable estimates of the underlying probability.
* In **random walks**, increasing variance over time signals the need for more advanced prediction methods to manage uncertainty, such as in financial markets where stock prices often follow random walk-like behavior.

**Conclusion**

The study of mean and variance over time reveals key insights into the nature of different statistical processes. While the mean tends to stabilize or converge in relative frequency and Bernoulli trials, it drifts or fluctuates in random walks. Variance shows a stabilizing effect in frequency-based processes but grows unbounded in random walks. These findings are crucial for modeling, predicting, and understanding stochastic behaviors in various domains such as finance, quality control, and machine learning.