Find the Arithmetic Mean and Other Statistics Iteratively

Paul Mayer, Aug 19, 2018

Imagine that you need to find the arithmetic mean of many thousands or even millions of data values, let's call that number of values *N*. Also imagine that these values are arriving as a stream and not as a completed list. The first way one could do it is to store all N data values in a table and do a one-time calculation of all the values using the formula:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Another way that is more convenient, and does not require storing all the values, is to update the mean value interatively as the data streams in. This has the added advantage that one can make decisions on the updating average instead of waiting for all the values to arrive. An example of this is statistical process control charts in manufacturing. Typically a group of three to five values are grouped and the mean of those values found and charted. Based on the incoming group of new data, the overall average can be updated. Call the number of data points in this group *m*.

Here is how to iteratively find the mean.

$$\bar{x}_{n+1} = \frac{N_n \bar{x}_n + \sum_{i=1}^{m} x_i}{N_n + m}$$

For the case of m =1, we get

$$\bar{x}_{n+1} = \frac{N_n \bar{x}_n + x_i}{N_n + 1}$$

One way to think of this is the the old mean is weighted by N_n/ (N_n+m) and the newly added values are weighted by m/(N_n+m). So each successive data group affects the overall mean by a smaller and smaller percentage. It can easily be shown that the iterative mean gives the same mean for a given set of data points as all the points calculated at once.

The other statistic that might be of interest is the standard deviation. In this case calculating it iteratively will not give the exact standard deviation, but it will converge to that value as N gets large. First, calculate variance iteratively. The question is, does one use the updated overall mean to calculate the variance, or the mean of the group of points that is coming in. Since the group size can be just 1, it makes sense to use the updated overall mean.

$$\sigma^2{}_{n+1} = \frac{N_n \sigma_n^2 + \sum_{i=1}^{m} (x_i - \bar{x}_{n+1})^2}{N_n + m}$$

$$\sigma_{n+1} = \sqrt{\sigma^2{}_{n+1}}$$

Let's look at the examples using randomly generated numbers.

The following tables have randomly generated X values between 300 and 600 in Excel using the function =RANDBETWEEN(300,600). Each time a new data point is added, the iterative average column (4) is

calculating an updated average value, which matches the average calculated all the X values using the function =AVERAGE($B$2:B10), where the first cell is locked down and the last cell is the row number n plus 1.

N	X values	Average	Iterative Avg
1	318	318.00	318.00
2	581	449.50	449.50
3	451	450.00	450.00
4	310	415.00	415.00
5	358	403.60	403.60
6	423	406.83	406.83
7	312	393.29	393.29
8	507	407.50	407.50
9	380	404.44	404.44
10	359	399.90	399.90
11	418	401.55	401.55
12	541	413.17	413.17
13	452	416.15	416.15
14	554	426.00	426.00
15	507	431.40	431.40
16	446	432.31	432.31
17	404	430.65	430.65
18	553	437.44	437.44
19	313	430.89	430.89
20	550	436.85	436.85

Here is the formula view:

N	X values	Average	Iterative Avg
1	=RANDBETWEEN(300,600)	=AVERAGE($B$2:B2)	=B2
2	=RANDBETWEEN(300,600)	=AVERAGE($B$2:B3)	=(A2*D2+B3)/A3
3	=RANDBETWEEN(300,600)	=AVERAGE($B$2:B4)	=(A3*D3+B4)/A4
4	=RANDBETWEEN(300,600)	=AVERAGE($B$2:B5)	=(A4*D4+B5)/A5
5	=RANDBETWEEN(300,600)	=AVERAGE($B$2:B6)	=(A5*D5+B6)/A6
6	=RANDBETWEEN(300,600)	=AVERAGE($B$2:B7)	=(A6*D6+B7)/A7
7	=RANDBETWEEN(300,600)	=AVERAGE($B$2:B8)	=(A7*D7+B8)/A8
8	=RANDBETWEEN(300,600)	=AVERAGE($B$2:B9)	=(A8*D8+B9)/A9
9	=RANDBETWEEN(300,600)	=AVERAGE($B$2:B10)	=(A9*D9+B10)/A10
10	=RANDBETWEEN(300,600)	=AVERAGE($B$2:B11)	=(A10*D10+B11)/A11
11	=RANDBETWEEN(300,600)	=AVERAGE($B$2:B12)	=(A11*D11+B12)/A12
12	=RANDBETWEEN(300,600)	=AVERAGE($B$2:B13)	=(A12*D12+B13)/A13

13	=RANDBETWEEN(300,600)	=AVERAGE($B$2:B14)	=(A13*D13+B14)/A14
14	=RANDBETWEEN(300,600)	=AVERAGE($B$2:B15)	=(A14*D14+B15)/A15
15	=RANDBETWEEN(300,600)	=AVERAGE($B$2:B16)	=(A15*D15+B16)/A16
16	=RANDBETWEEN(300,600)	=AVERAGE($B$2:B17)	=(A16*D16+B17)/A17
17	=RANDBETWEEN(300,600)	=AVERAGE($B$2:B18)	=(A17*D17+B18)/A18
18	=RANDBETWEEN(300,600)	=AVERAGE($B$2:B19)	=(A18*D18+B19)/A19
19	=RANDBETWEEN(300,600)	=AVERAGE($B$2:B20)	=(A19*D19+B20)/A20
20	=RANDBETWEEN(300,600)	=AVERAGE($B$2:B21)	=(A20*D20+B21)/A21

For the standard deviation the exact value does not match the iterative value, but the interative value converges to the exact value.

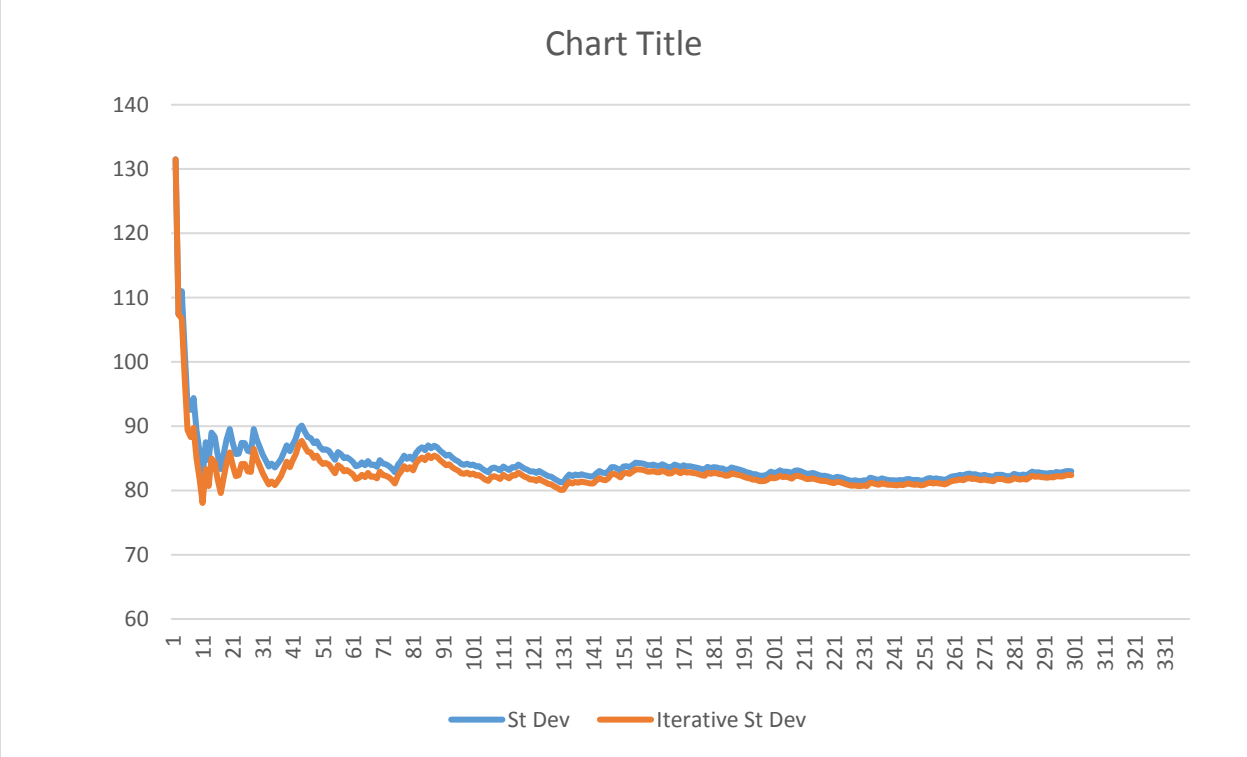
N	X values	Iterative Avg	St Dev	Iterative St Dev
1	318	318.00		
2	581	449.50	131.5	131.5
3	451	450.00	107.3716	107.3708527
4	310	415.00	111.0023	106.7830745
5	358	403.60	101.8678	97.66254144
6	423	406.83	93.27275	89.3972605
7	312	393.29	92.51056	88.28410993
8	507	407.50	94.35439	89.76276961
9	380	404.44	89.37699	85.02050035
10	359	399.90	85.87951	81.68793015
11	418	401.55	82.04806	78.04423652
12	541	413.17	87.50127	83.33731282
13	452	416.15	84.70296	80.68277828
14	554	426.00	89.00803	84.9412567
15	507	431.40	88.3318	84.3506962
16	446	432.31	85.59988	81.74386326
17	404	430.65	83.31086	79.56610742
18	553	437.44	85.67719	81.98106118
19	313	430.89	87.9	84.25377332
20	550	436.85	89.52054	85.9296778

Here are the formulas.

N	X values	Iterative Avg	St Dev	Iterative St Dev
1	=RANDBETWEEN(300,600)	=B2		
2	=RANDBETWEEN(300,600)	=(A2*D2+B3)/A3	=STDEV.P($B$2:B3)	=E3
3	=RANDBETWEEN(300,600)	=(A3*D3+B4)/A4	=STDEV.P($B$2:B4)	=((A3*F3^2+(B4-D4)^2)/A4)^0.5
4	=RANDBETWEEN(300,600)	=(A4*D4+B5)/A5	=STDEV.P($B$2:B5)	=((A4*F4^2+(B5-D5)^2)/A5)^0.5
5	=RANDBETWEEN(300,600)	=(A5*D5+B6)/A6	=STDEV.P($B$2:B6)	=((A5*F5^2+(B6-D6)^2)/A6)^0.5

6	=RANDBETWEEN(300,600)	=(A6*D6+B7)/A7	=STDEV.P($B$2:B7)	=((A6*F6^2+(B7-D7)^2)/A7)^0.5
7	=RANDBETWEEN(300,600)	=(A7*D7+B8)/A8	=STDEV.P($B$2:B8)	=((A7*F7^2+(B8-D8)^2)/A8)^0.5
8	=RANDBETWEEN(300,600)	=(A8*D8+B9)/A9	=STDEV.P($B$2:B9)	=((A8*F8^2+(B9-D9)^2)/A9)^0.5
9	=RANDBETWEEN(300,600)	=(A9*D9+B10)/A10	=STDEV.P($B$2:B10)	=((A9*F9^2+(B10-D10)^2)/A10)^0.5
10	=RANDBETWEEN(300,600)	=(A10*D10+B11)/A11	=STDEV.P($B$2:B11)	=((A10*F10^2+(B11-D11)^2)/A11)^0.5
11	=RANDBETWEEN(300,600)	=(A11*D11+B12)/A12	=STDEV.P($B$2:B12)	=((A11*F11^2+(B12-D12)^2)/A12)^0.5
12	=RANDBETWEEN(300,600)	=(A12*D12+B13)/A13	=STDEV.P($B$2:B13)	=((A12*F12^2+(B13-D13)^2)/A13)^0.5
13	=RANDBETWEEN(300,600)	=(A13*D13+B14)/A14	=STDEV.P($B$2:B14)	=((A13*F13^2+(B14-D14)^2)/A14)^0.5
14	=RANDBETWEEN(300,600)	=(A14*D14+B15)/A15	=STDEV.P($B$2:B15)	=((A14*F14^2+(B15-D15)^2)/A15)^0.5
15	=RANDBETWEEN(300,600)	=(A15*D15+B16)/A16	=STDEV.P($B$2:B16)	=((A15*F15^2+(B16-D16)^2)/A16)^0.5
16	=RANDBETWEEN(300,600)	=(A16*D16+B17)/A17	=STDEV.P($B$2:B17)	=((A16*F16^2+(B17-D17)^2)/A17)^0.5
17	=RANDBETWEEN(300,600)	=(A17*D17+B18)/A18	=STDEV.P($B$2:B18)	=((A17*F17^2+(B18-D18)^2)/A18)^0.5
18	=RANDBETWEEN(300,600)	=(A18*D18+B19)/A19	=STDEV.P($B$2:B19)	=((A18*F18^2+(B19-D19)^2)/A19)^0.5
19	=RANDBETWEEN(300,600)	=(A19*D19+B20)/A20	=STDEV.P($B$2:B20)	=((A19*F19^2+(B20-D20)^2)/A20)^0.5
20	=RANDBETWEEN(300,600)	=(A20*D20+B21)/A21	=STDEV.P($B$2:B21)	=((A20*F20^2+(B21-D21)^2)/A21)^0.5

To show that it converges, here is a chart with more data points.



Other ways to estimate the standard deviation is to calculate ranges of a group of data, but that assumes a population distribution, usually the normal distribution, so I won't cover that.