Batch Recursive Formula for Variance

Jingxuan Yang*

July 1, 2022

In this note, we will give the recursive formulas for sample mean and sample variance, and their generalized forms for batch updates.

Suppose the data samples are $\{x_i\}_{i=1}^n$. The sample mean is

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1}$$

The sample variance is

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2. \tag{2}$$

To find the recursive formula for sample variance, we note that for a random variable X, the variance is

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - \mathbb{E}^2[X].\end{aligned} \tag{3}$$

Therefore, the sample variance can be expressed in another form as

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2, \tag{4}$$

which can also be obtained by some straightforward manipulations from (2) and is more suitable for the following derivations.

1 Recursive Formula for Variance

The sample means for n and $n-1$ samples are

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{5}$$

and

$$\bar{x}_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n-1} x_i, \tag{6}$$

*Jingxuan Yang is with the Department of Automation, Tsinghua University, Beijing 100084, China (email: yangjx20@mails.tsinghua.edu.cn).

respectively. Therefore, we have

$$n\bar{x}_n = (n-1)\bar{x}_{n-1} + x_n, \tag{7}$$

and hence

$$\bar{x}_n = \bar{x}_{n-1} + \frac{x_n - \bar{x}_{n-1}}{n}, \tag{8}$$

which is the recursive formula for sample mean.

The sample variances for n and $n-1$ samples are

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2, \tag{9}$$

and

$$\sigma_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n-1}(x_i - \bar{x}_{n-1})^2 = \frac{1}{n-1}\sum_{i=1}^{n-1} x_i^2 - \bar{x}_{n-1}^2, \tag{10}$$

respectively. Therefore, we have

$$n\sigma_n^2 - (n-1)\sigma_{n-1}^2 = x_n^2 - n\bar{x}_n^2 + (n-1)\bar{x}_{n-1}^2, \tag{11}$$

and hence

$$\sigma_n^2 = \frac{n-1}{n}(\sigma_{n-1}^2 + \bar{x}_{n-1}^2) + \frac{x_n^2}{n} - \bar{x}_n^2, \tag{12}$$

which is already the recursive formula for sample variance. Another form is to replace the $\bar{x}_n$ with its recursive formula (8). Multiply (7) and (8) to get

$$\begin{aligned} n\bar{x}_n^2 &= [(n-1)\bar{x}_{n-1} + x_n]\left(\bar{x}_{n-1} + \frac{x_n - \bar{x}_{n-1}}{n}\right) \\ &= (n-1)\bar{x}_{n-1}^2 + \frac{n-1}{n}\bar{x}_{n-1}(x_n - \bar{x}_{n-1}) + x_n\bar{x}_{n-1} + \frac{x_n(x_n - \bar{x}_{n-1})}{n}. \end{aligned} \tag{13}$$

Therefore, we have

$$(n-1)\bar{x}_{n-1}^2 - n\bar{x}_n^2 = \frac{n-1}{n}\bar{x}_{n-1}^2 - 2\frac{n-1}{n}x_n\bar{x}_{n-1} - \frac{1}{n}x_n^2. \tag{14}$$

Substituting (14) into (11), we obtain

$$\begin{aligned} n\sigma_n^2 - (n-1)\sigma_{n-1}^2 &= x_n^2 + \frac{n-1}{n}\bar{x}_{n-1}^2 - 2\frac{n-1}{n}x_n\bar{x}_{n-1} - \frac{1}{n}x_n^2 \\ &= \frac{n-1}{n}(x_n^2 - 2x_n\bar{x}_{n-1} + \bar{x}_{n-1}^2) \\ &= \frac{n-1}{n}(x_n - \bar{x}_{n-1})^2, \end{aligned} \tag{15}$$

and

$$\begin{aligned} \sigma_n^2 &= \frac{n-1}{n}\sigma_{n-1}^2 + \frac{n-1}{n^2}(x_n - \bar{x}_{n-1})^2 \\ &= \frac{n-1}{n}\left[\sigma_{n-1}^2 + \frac{1}{n}(x_n - \bar{x}_{n-1})^2\right], \end{aligned} \tag{16}$$

which is the second form of the recursive formula for sample variance and is more frequently adopted in practice. It expresses the variance of n samples with the variance of $n-1$ samples and the square distance between the sample x_n and sample mean $\bar{x}_{n-1}$.

2 Batch Recursive Formula for Variance

Sometimes we want to update the sample mean and sample variance with a batch of $m \geqslant 1$ samples $\{x_i\}_{i=n-m+1}^{n}$. Denote the batch mean and batch variance as

$$\bar{s}_m = \frac{1}{m}\sum_{i=1}^{m} x_{n-m+i}, \tag{17}$$

and

$$s_m^2 = \frac{1}{m}\sum_{i=1}^{m} x_{n-m+i}^2 - \bar{s}_m^2, \tag{18}$$

respectively.

The sample means for n and $n-m$ samples are

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{19}$$

and

$$\bar{x}_{n-m} = \frac{1}{n-m}\sum_{i=1}^{n-m} x_i, \tag{20}$$

respectively. Therefore, we have

$$\begin{aligned} n\bar{x}_n &= (n-m)\bar{x}_{n-1} + \sum_{i=1}^{m} x_{n-m+i} \\ &= (n-m)\bar{x}_{n-1} + m\bar{s}_m, \end{aligned} \tag{21}$$

where the second equality is obtained from (17), and hence

$$\bar{x}_n = \bar{x}_{n-m} + \frac{m}{n}(\bar{s}_m - \bar{x}_{n-m}), \tag{22}$$

which is the batch recursive formula for sample mean.

The sample variances for n and $n-m$ samples are

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2, \tag{23}$$

and

$$\sigma_{n-m}^2 = \frac{1}{n-m}\sum_{i=1}^{n-m} x_i^2 - \bar{x}_{n-m}^2, \tag{24}$$

respectively. Therefore, we have

$$\begin{aligned} n\sigma_n^2 - (n-m)\sigma_{n-m}^2 &= \sum_{i=1}^{m} x_{n-m+i}^2 - n\bar{x}_n^2 + (n-m)\bar{x}_{n-m}^2 \\ &= m(s_m^2 + \bar{s}_m^2) - n\bar{x}_n^2 + (n-m)\bar{x}_{n-m}^2, \end{aligned} \tag{25}$$

where the second equality is obtained from (18), and hence

$$\sigma_n^2 = \frac{n-m}{n}(\sigma_{n-m}^2 + \bar{x}_{n-m}^2) + \frac{m}{n}(s_m^2 + \bar{s}_m^2) - \bar{x}_n^2, \tag{26}$$

which is already the batch recursive formula for sample variance. Another form is to replace the $\bar{x}_n$ with its batch recursive formula (22). Multiply (21) and (22) to get

$$\begin{aligned} n\bar{x}_n^2 &= [(n-m)\bar{x}_{n-m} + m\bar{s}_m]\left(\bar{x}_{n-m} + \frac{m}{n}(\bar{s}_m - \bar{x}_{n-m})\right) \\ &= (n-m)\bar{x}_{n-m}^2 + \frac{m(n-m)}{n}\bar{x}_{n-m}(\bar{s}_m - \bar{x}_{n-m}) + m\bar{s}_m\bar{x}_{n-m} \\ &\quad + \frac{m^2}{n}\bar{s}_m(\bar{s}_m - \bar{x}_{n-m}). \end{aligned} \tag{27}$$

Therefore, we have

$$(n-m)\bar{x}_{n-m}^2 - n\bar{x}_n^2 = \frac{m(n-m)}{n}\bar{x}_{n-m}^2 - 2\frac{m(n-m)}{n}\bar{s}_m\bar{x}_{n-m} - \frac{m^2}{n}\bar{s}_m^2. \tag{28}$$

Substituting (28) into (25), we obtain

$$\begin{aligned} n\sigma_n^2 - (n-m)\sigma_{n-m}^2 &= m(s_m^2 + \bar{s}_m^2) + \frac{m(n-m)}{n}\bar{x}_{n-m}^2 - \frac{m^2}{n}\bar{s}_m^2 \\ &\quad - 2\frac{m(n-m)}{n}\bar{s}_m\bar{x}_{n-m} \\ &= ms_m^2 + \frac{m(n-m)}{n}(\bar{s}_m^2 - 2\bar{s}_m\bar{x}_{n-m} + \bar{x}_{n-m}^2) \\ &= ms_m^2 + \frac{m(n-m)}{n}(\bar{s}_m - \bar{x}_{n-m})^2, \end{aligned} \tag{29}$$

and

$$\begin{aligned} \sigma_n^2 &= \frac{n-m}{n}\sigma_{n-m}^2 + \frac{m(n-m)}{n^2}(\bar{s}_m - \bar{x}_{n-m})^2 + \frac{m}{n}s_m^2 \\ &= \frac{n-m}{n}\left[\sigma_{n-m}^2 + \frac{m}{n}(\bar{s}_m - \bar{x}_{n-m})^2 + \frac{m}{n-m}s_m^2\right], \end{aligned} \tag{30}$$

which is the second form of the batch recursive formula for sample variance and is more frequently adopted in practice. It expresses the variance of n samples with the variance of $n-m$ samples, the square distance between the batch mean $\bar{s}_m$ and sample mean $\bar{x}_{n-m}$ and the batch variance s_m^2.