Statistics and Probability Letters 114 (2016) 14–19

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

An accurate updating formula to calculate sample variance from weighted successive differences

Tien-Lung Sun [a], Kuo-Hung Lo [a,b], Juei-Chao Chen [c,*]

[a] Department of Industrial Engineering & Management, Yuan Ze University, Taiwan, ROC

[b] Department of Marketing & Logistics Management, Yu Da University of Science and Technology, Taiwan, ROC

[c] Department of Statistics and Information Science, Fu Jen Catholic University, Taiwan, ROC

ARTICLE INFO

Article history:
Received 3 February 2015
Received in revised form 11 August 2015
Accepted 8 March 2016
Available online 15 March 2016

Keywords:
Successive difference
Variance updating

ABSTRACT

We proposed and proved a new variance updating formula that is based on successive differences. Performance evaluation using StRD and other testing datasets shows that our formula is advantageous in handling integer data and floating point data with larger stiffness.

1. Introduction

Sample variance updating is needed in many data analysis situations, it is defined as:

$$S_n^2 = SS_n / (n-1), \quad n \geq 2, \tag{1}$$

where $SS_n = \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$, $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$. Formula (1) inquires two passes through the data: first to compute $\bar{x}_n$ and then to sum the terms $(x_i - \bar{x}_n)^2$. The two-pass formula prohibits real-time data compilation and makes it infeasible for handling large datasets.

Many one-pass, updating formulas have appeared in the literature. These formulas read the input data only once and update SS_n by adding an updating term to the previously calculated and stored value of SS_{n-1}. Most of the variance updating formulas, however, need to perform mean updating and then use the updated mean to update the variance (e.g. Welford, 1962; Van Reeken, 1968; Hanson, 1975; Cotton, 1975; West, 1979; Knuth, 1998). Mean updating is error-prone as rounding errors are introduced when performing division operations.

Joarder (2003) developed a variance updating formula that avoids mean updating by using weighted successive differences:

Theorem 1.

$$nSS_n = \mathbf{d}_{1\times n}^{\mathrm{T}} \mathbf{c}_{n\times n} \mathbf{d}_{n\times 1} = \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} d_i d_j, \tag{2}$$

* Corresponding author. Tel.: +886 2 2905 2935.
E-mail addresses: tsun@saturn.yzu.edu.tw (T.-L. Sun), s978904@mail.yzu.edu.tw (K.-H. Lo), 006884@mail.fju.edu.tw (J.-C. Chen).

http://dx.doi.org/10.1016/j.spl.2016.03.003
0167-7152/

where $\mathbf{c}_{n\times n} = \left[c_{ij}\right]_{n\times n}$ *is the weight matrix with:*

$$c_{ij} = \begin{cases} (n-i+1)\,(j-1)\,, & i \geq j, \\ (n-j+1)\,(i-1)\,, & i < j, \end{cases} \tag{3}$$

$\mathbf{d}^{\mathrm{T}}_{1\times n} = \begin{bmatrix} d_1 & d_2 & \dots & d_n \end{bmatrix}_{1\times n}$ *is the successive differences vector with* $d_i = x_i - x_{i-1}$, *for* $x_0 = 0$, $i = 1,\ 2, \dots,\ n$.

Although (2) and (3) are easy to calculate by computer with much efficiency, the double summation structure needs additional memory of size n to store d_i during computation. If n is large, these restrictions could become substantial limitation. To avoid this problem, the authors derived an updating formula.

Corollary 1. *Let* $SS_1 = 0$. *Then, for* $n \geq 2$,

$$nSS_n = (n-1)\,SS_{n-1} + U^*_n, \tag{4}$$

where the updating term U^*_n *is*

$$U^*_n = \left(\sum_{i=2}^{n} d_i\right)^2 + \left(\sum_{i=3}^{n} d_i\right)^2 + \cdots + (d_{n-1} + d_n)^2 + d_n^2. \tag{5}$$

Notice that the updating term U^*_n in (5) contains $(n-1)$ updating items. Moreover, the successive difference d_n calculated from the n^{th} input data has to be added to all the $(n-1)$ updating terms in U^*_n. This greatly reduces the computational efficiency. The reason with this problem is because the weight formula is not in a unified format but has to be represented as two formulas. Consequently, the derived updating formula cannot be represented in a compact format. To address this problem, we have derived a unified weight formula (Lo et al., 2014):

Theorem 2.

$$nSS_n = \mathbf{d}^{\mathrm{T}}_{1\times n}\mathbf{w}_{n\times n}\mathbf{d}_{n\times 1} = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij(n)}\,d_i d_j, \tag{6}$$

where $\mathbf{w}_{n\times n} = \left[w_{ij(n)}\right]_{n\times n}$ *is* $n \times n$ *symmetric matrix with:*

$$w_{ij(n)} = (n+1)\,(i+j-1) - ij - \frac{n}{2}\,(i+j+|i-j|)\,, \quad i,\ j = 1,\ 2, \dots, n. \tag{7}$$

With this unified weight formula we developed and proved a variance updating formula in this paper that reduces the number of updating items from $(n-1)$ to only two. The proposed updating formula is also advantageous over the double summation structure proposed in Lo et al. (2014), i.e., Eqs. (6) and (7), as no additional memory is needed during computation.

In the remaining of this paper, we first derived and proved the updating formula in Section 2. Then in Section 3 we compare the computational performance of our formula with other formulas reported in literature using StRD benchmark and other testing datasets.

2. Main results

Theorem 3. *Given a temporally ordered set of observations* $x_1,\ x_2, \dots, x_n, \dots,$

$$nSS_n = (n-1)\,SS_{n-1} + U_n, \tag{8}$$

where

$$U_n = U_{n-1} + 2\left(\sum_{i=1}^{n-1}(i-1)\,d_i\right)d_n + (n-1)\,d_n^2. \tag{9}$$

Proof.

$$\begin{aligned} \mathbf{w}_{n\times n} &= \left[w_{ij(n)}\right]_{n\times n} = \left[w_{ij(n-1)} + \frac{1}{2}\,(i+j-|i-j|-2)\right]_{n\times n} \\ &= \left[w_{ij(n-1)}\right]_{n\times n} + \left[g_{ij(n)}\right]_{n\times n}, \quad \text{with } g_{ij(n)} = \frac{1}{2}\,(i+j-|i-j|-2) \\ &= \begin{bmatrix} \mathbf{w}_{(n-1)\times(n-1)} & \mathbf{0}_{(n-1)\times 1} \\ \mathbf{0}_{1\times(n-1)} & 0 \end{bmatrix}_{n\times n} + \mathbf{g}_{n\times n}, \end{aligned} \tag{10}$$