The Sample Variance and the First Order Differences of Observations

Anwar H. Joarder

Department of Mathematical Sciences, King Fahd University of Petroleum and Minerals
Dhahran 31261, Saudi Arabia, Email: anwarj@ kfupm.edu.sa

Abstract It is proved that the variance can be calculated by the first order differences of sample observations. It is also represented by a quadratic form in the differences of sample observations via a constant matrix (depending on the sample size) which is open for further study. The avoidance of the mean in the calculation of the variance is expected to increase precision especially if computer programs are used. An alternative method is also presented for the calculation of the variance from a frequency distribution.

Key Words: Sample variance; first order differences; pattern matrix.

1. Introduction

Since the sample variance depends on rounding off the sample mean, it lacks in precision especially when computer programs are used for calculation. The problem of calculating the variance by avoiding the use of the sample mean was posed by Ross (1987, 143-144). In the spirit of Ross (1987), some solutions to the problem were discussed by Joarder (2002). The variance $(s_n{}^2)$ of n observations in a sample is just the ratio of TSS (Total squared deviations corrected by the mean) to the degrees of freedom where

$$TSS = (n-1)\, s_n{}^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}x_i^2 - n\bar{x}^2, \quad n \ge 2\,. \tag{1.1}$$

If sample observations are integers but not large in sizes, the last representation in (1.1) allows one to do the calculation mentally. The quantity TSS can also be represented by the following equivalent forms

$$\frac{1}{n}\sum_{1\le i<j\le n}(x_i-x_j)^2 = \frac{1}{2n}\sum_{i==1}^{n}\sum_{j=1}^{n}(x_i-x_j)^2 = \frac{1}{n}\sum_{i=2}^{n}\sum_{j=1}^{i-1}(x_i-x_j)^2 \tag{1.2}$$

(see e.g. Kotz, Kozubowski and Podgorski, 2001, 186). The implication of the result in (1.2) is that the variance of a sample of n observations can be easily calculated by calculating the variances of $\binom{n}{2}$ pairs of observations and then averaging them. That is for a sample of size $n \ge 2$ the variance is given by:

$$s_n^2 = \binom{n}{2}^{-1}\sum_{i=2}^{n}\sum_{j=1}^{i-1}\frac{w_{ij}{}^2}{2} \text{ where } w_{ij} = x_i - x_j \quad (i, j = 1,\ 2,\ \ldots, n)\,. \tag{1.3}$$

This note resorts to the first order differences of sample observations to calculate the variance. It is also represented by a quadratic form in the differences of sample observations via a constant matrix C (depending on the sample size) which is open for further study. An alternative method for the calculation of the variance from a frequency distribution with equal class widths is also presented.

It follows from equation (1.3) that a table showing the differences among observations can be prepared whose entries are $w_{ij} = x_i - x_j$ $(i, j = 1, 2, \ldots, n;\ i > j)$ to calculate the variance.

Example 1.1 To calculate the variance of the sample (104, 94, 95, 101, 111), let us denote the ordered observations by $x_1 = 94,\ x_2 = 95,\ x_3 = 101,\ x_4 = 104,\ x_5 = 111$. Then the values $w_{ij} = x_i - x_j \quad (i, j = 1,\ 2,\ ..., 5;\ \ i > j)$ are given by $w_{21} = 1$, $w_{31} = 7, w_{32} = 6$, $w_{41} = 10,\ w_{42} = 9,\ w_{43} = 3$, ,$w_{51} = 17,\ w_{52} = 16,\ w_{53} = 10,\ w_{54} = 7$. Then by (1.3) we have $\sum_{i=2}^{5}\sum_{j=1}^{4} w_{ij}^2 = 1^2 + 7^2 + 6^2 + ... + 7^2 = 970$ and $s_5^2 = \frac{1}{5(4)}\ (970) = \frac{970}{5(4)} = 48.5$.

2. The Main Results

Consider a lower triangular matrix $W = ((w_{ij}))$, with elements $w_{ij} = x_i - x_j$, $(i = 1,2,...,n;\ j = 1,2,...,n;\ i > j)$. Further consider imaginary right angled triangles with vertices $w_{ij}\,(i > j)'s$ and right angle at the bottom left corner of the lower triangular matrix, and diagonal as the hypotenuse. Then any element, excluding the elements on the hypotenuse, in the right angled vertex of any triangle is the sum of the elements in the corresponding part of the hypotenuse. For example, with a sample of size $n = 5$, the first order differences of the observations are $w_{21} = d_1, w_{32} = d_2,\ w_{43} = d_3$, $w_{54} = d_4$ so that

$$w_{31} = w_{21} + w_{32} = d_1 + d_2$$

$$w_{41} = w_{21} + w_{32} + w_{43} = d_1 + d_2 + d_3,\ w_{42} = w_{32} + w_{43} = d_2 + d_3$$

$$w_{51} = \sum_{i=1}^{4} d_i\ ,\ w_{52} = d_2 + d_3 + d_4,\ w_{53} = d_3 + d_4\,.$$

In general let $w_{ij} = x_i - x_j\ (i, j = 1,2,\cdots,n; i > j)$ and the first order differences be $w_{i+1,i} = d_i\,(i = 1,2,\cdots,n)$. The elements of the $l\ \ (l = 1,2,...,n-1)$ th diagonal line is thus given by

$$w_{i+l,i} = x_{i+l} - x_i = (x_{i+l} - x_{i+l-1}) + (x_{i+l-1} - x_{i+l-2}) + \cdots + (x_{i+2} - x_{i+1}) + (x_{i+1} - x_i)$$

$$= d_{i+l-1} + d_{i+l-2} + \cdots + d_{i+1} + d_i = \sum_{j=i}^{i+l-1} d_j\,,\ \ i = 1, 2, ..., n - l\ \ .$$

Theorem 2.1 Let $d_i\,(i = 1,2,\cdots,n)$ be the first order difference of observations, and $w_{ij} = \sum_{k=j}^{i-1} d_k\ ,\ \ (i > j)$. Then the variance of $n \ge 2$ observations is given by

$$s_n^2 = \frac{TSS}{n-1} = \frac{1}{(n-1)n}\ \ \sum_{i=2}^{n}\sum_{j=1}^{i-1} w_{ij}{}^2 = \frac{1}{n(n-1)}\ \ d'Cd$$

where $d' = (d_1, d_2, \cdots, d_{n-1})$ and $C = ((c_{ij}))$ is a $(n-1)\times(n-1)$ symmetric matrix with $c_{ij} = (n-i)\,j\ \ \text{if}\ i \ge j;\ \ (\,i, j = 1,2,\cdots,n-1)$.

Proof. It follows from (1.2) and the above notations that

$$n\ TSS = \sum_{i=2}^{n}\sum_{j=1}^{i-1} w_{ij}{}^2$$

$$= w_{21}^2 + (w_{31}^2 + w_{32}^2) + (w_{41}^2 + w_{42}^2 + w_{43}^2) + \cdots$$

$$+ (w_{n-1,1}^2 + w_{n-1,2}^2 + \cdots + w_{n-1,n-3}^2 + w_{n-1,n-2}^2) + \cdots$$

$$+ (w_{n1}^2 + w_{n2}^2 + \cdots + w_{n,n-2}^2 + w_{n,n-1}^2)$$

$$= (w_{21}^2 + w_{32}^2 + w_{43}^2 + \cdots + w_{n-1,n-2}^2 + w_{n,n-1}^2) + \cdots$$
$$+ (w_{31}^2 + w_{42}^2 + \cdots + w_{n-1,n-3}^2 + w_{n,n-2}^2) + \cdots$$
$$+ (w_{n-1,1}^2 + w_{n2}^2) + w_{n1}^2$$
$$= \sum_{j=1}^{n-1} d_j^2 + \left[(d_1 + d_2)^2 + (d_2 + d_3)^2 + \cdots + (d_{n-2} + d_{n-1})^2\right]$$
$$+ \left((d_1 + d_2 + d_3)^2 + (d_2 + d_3 + d_4)^2 + \cdots + (d_{n-3} + d_{n-2} + d_{n-1})^2\right)$$
$$+ \cdots$$
$$+ \left(\left(\sum_{i=1}^{n-2} d_i\right)^2 + \left(\sum_{i=2}^{n-1} d_i\right)^2\right) + \left(\sum_{i=1}^{n-1} d_i\right)^2$$

$$n\ TSS = \sum_{j=1}^{n-1} d_j^2 + \sum_{i=1}^{n-2}\left(\sum_{j=i}^{i+1} d_j\right)^2 + \sum_{i=1}^{n-3}\left(\sum_{j=i}^{i+2} d_j\right)^2 + \cdots + \sum_{i=1}^{2}\left(\sum_{j=i}^{i+n-3} d_j\right)^2 + \sum_{i=1}\left(\sum_{j=i}^{i+n-2} d_j\right)^2 \qquad (2.1)$$

Since $n\ TSS = n(n-1)s_n^2$, it follows from (2.1) for $n = 2,3,4,5,\cdots$ that

$2(1)s_2^2 = d_1^2$,

$3(2)s_3^2 = d_1^2 + d_2^2 + (d_1 + d_2)^2 = 2\left(d_1^2 + d_2^2 + d_1 d_2\right) = d'Cd$ where $d' = (d_1, d_2)$ and $C = ((c_{ij}))$ with $c_{11} = c_{22} = 2,\ \ c_{21} = 1$,

$4(3)s_4^2 = d_1^2 + d_2^2 + d_3^2 + (d_1 + d_2)^2 + (d_2 + d_3)^2 + (d_1 + d_2 + d_3)^2 = d'Cd$ where $d' = (d_1, d_2, d_3)$ and $C = ((c_{ij}))$ with $c_{11} = c_{33} = 3,\ \ c_{21} = c_{32} = 2,\ \ c_{31} = 1$, and

$$5(4)s_5^2 = d_1^2 + d_2^2 + d_3^2 + d_4^2 + (d_1 + d_2)^2 + (d_2 + d_3)^2 + (d_3 + d_4)^2$$
$$+ (d_1 + d_2 + d_3)^2 + (d_2 + d_3 + d_4)^2 + (d_1 + d_2 + d_3 + d_4)^2 = d'Cd$$

where $d' = (d_1, d_2, d_3, d_4)$ and $C = ((c_{ij}))$ with $c_{11} = c_{32} = c_{44} = 4,$
$c_{21} = c_{43} = 3,\quad c_{31} = c_{42} = 2,\ \ c_{41} = 1.$

Proceeding thus we have the general expression as stated in the theorem.

Note that $d' = (w_{21},\ w_{32}\ ,\cdots, w_{i,i-1}, \cdots, w_{n,n-1}) = (d_1, d_2, \cdots, d_{n-1})$. Though the algebra in (2.1) looks a bit complicated, the entire calculation can be facilitated by ordering the sample observations and then preparing a table showing d_i's (the first order differences) in a column, followed by appropriate totals as needed by (2.1) in succeeding columns. The following corollary is obvious.

Corollary 2.1 For a sample of size $n \geq 2$, the following recurrence relation holds:

$$(n+1)\ TSS\,(n+1) - nTSS\,(n)$$
$$= \left(\sum_{i=1}^{n} d_i\right)^2 + \left(\sum_{i=2}^{n} d_i\right)^2 + \cdots + \left(d_{n-2} + d_{n-1} + d_n\right)^2 + \left(d_{n-1} + d_n\right)^2 + d_n^2$$

where $TSS\,(m)$ is the total sums of squares of m observations corrected by the sample mean.

Example 2.1 To calculate the variance of the sample (104, 94, 95, 101, 111), the first order differences of the ordered observations given by $d_1 = 1,\ d_2 = 6,\ d_3 = 3,\ d_4 = 4$ are used as follows:

(i) Calculation of the variance by the first order differences and their "totals"

By (2.1) we have

$$5(4)\, s_5^2 = (1^2 + 6^2 + 3^2 + 7^2) + (7^2 + 9^2 + 10^2) + (10^2 + 16^2) + 17^2 = 970$$

so that the variance is $s_5^2 = 97/2$.

(ii) Calculation of the variance by the first order differences and the constant matrix C

By the use of Theorem 2.1 we have $5(4)\ s_5^2 = d'Cd = 970$ where $d' = (d_1, d_2, d_3, d_4) = (1,6,3,7)$ and

$$C = \begin{pmatrix} 4 & 3 & 2 & 1 \\ 3 & 6 & 4 & 2 \\ 2 & 4 & 6 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix}$$

so that the variance is $s_5^2 = 97/2$.

(iii) *Calculation of the variance by the recurrence relation*

The variance can be calculated by Corollary 2.1 as follows:

$$\begin{aligned}
2TSS(2) &= d_1^2 = 1 \\
3TSS(3) &= 2TSS(2) + (d_1 + d_2)^2 + d_2^2 = 1 + 7^2 + 6^2 = 86 \\
4TSS(4) &= 3TSS(3) + (d_1 + d_2 + d_3)^2 + (d_2 + d_3)^2 + d_3^2 = 86 + 10^2 + 9^2 + 3^2 = 276 \\
5TSS(5) &= 4TSS(4) + (d_1 + d_2 + d_3 + d_4)^2 + (d_2 + d_3 + d_4)^2 + (d_3 + d_4)^2 + d_4^2 \\
&= 276 + 17^2 + 16^2 + 10^2 + 7^2 = 970
\end{aligned}$$

Similarly, other quantities involving sums of squares e.g. correlation coefficient in a bivariate sample can be calculated.

3. The Variance of Observations with Frequencies

The variance of a sample classified into k classes in a frequency distribution with mid values $y_1, y_2, \cdots, y_k$ and frequencies $f_1, f_2, \cdots, f_k$ is given by $(n-1)s^2 = \sum_{i=1}^{k} (y_i - \bar{y})^2 f_i$. The following theorem is well known (Joarder, 2002).

Theorem 3.1 Let $n = \sum_{i=1}^{k} f_i$ observations be divided into k groups containing $f_1, f_2, \cdots, f_k$ observations with means $\bar{x}_{(1)}, \bar{x}_{(2)}, \cdots, \bar{x}_{(k)}$ and variances $s_i^2\ (i = 1,2,\cdots,k)$ respectively. Then , $TSS = \sum_{i=1}^{k} (f_i - 1)^2\ s_i^2 + n^{-1} \sum_{i\,(<l)=1}^{k} (\bar{x}_i - \bar{x}_l)^2\ f_i\ f_l$.

In case observations in every group are the same, the means of k groups may be denoted by $y_1, y_2, \cdots, y_k$ and $s_i^2 = 0$, $(i = 1,2,\cdots,k)$, the variance of the ith group, so that the first summand in the TSS in the above expression vanishes, and we have the following corollaries.

Corollary 3.1 Let $y_1, y_2, \cdots, y_k$ have frequencies $f_1, f_2, \cdots, f_k$ with $n = f_1 + f_2 + \cdots + f_k$ then, $-TSS = n^{-1} \sum_{i\,(<l)=1}^{k} (y_i - y_l)^2 f_i\, f_l$.

Corollary 3.2 If k observations follow arithmetic series with common difference w and frequencies f_i $(i = 1,2,\cdots,k)$, then TSS is given by

$$TSS = n^{-1}w^2 \sum_{i\,(<l)=1}^{k} (i-l)^2 \; f_i\; f_l \; = w^2 \left(\sum_{i=1}^{k} i^2 f_i - n^{-1} \left(\sum_{i=1}^{k} i f_i \right)^2 \right).$$

The TSS in the above corollary follows from the basic definition of the variance by noting that the variance does not depend on the origin of transformation. Note that in this case the variance depends on the observations through the common difference w , first k positive integers and corresponding frequencies.

Example 3.1 (Bluman, 2001, 113) Thirty automobiles were tested for fuel efficiency (in miles per gallon). The following frequency distribution was obtained.

y_i	$7.5-12.5$	$12.5-17.5$	$17.5-22.5$	$22.5-27.5$	$27.5-32.5$
f_i	3	5	15	5	2

Since $\sum_{i=1}^{5} i f_i = 1(3) + 2(5) + 3(15) + 4(5) + 5(2) = 88$ and $\sum_{i=1}^{5} i^2 f_i = 288$, it follows from Corollary 3.2 that $TSS = 5^2 \left(288 - (88)^2 / 30 \right) = 746\tfrac{2}{3}$ so that the variance is 25.747 approximately. Joarder (2003) deduced many special cases from Corollary 3.2 which may be of much use in constructing examples quickly in classrooms. The following corollary follows from Corollary 3.2 or directly by the basic definition of variance.

Corollary 3.3 If $n = kf$ observations follow arithmetic series with common difference w and common frequency f , then the variance is given by

$$s_n^2 = \frac{k(k+1)}{12}\;\frac{n-f}{n-1}\; w^2$$

where $k(k+1)/12$ is the variance of k natural integers.

Acknowledgements

The author gratefully acknowledges the excellent research facilities available at King Fahd University of Petroleum and Minerals, Saudi Arabia. The author also thanks an anonymous referee for constructive suggestions which have improved the readability of the paper.

References

Bluman, A.G. (2001). *Elementary Statistics: A Step by Step Approach*. McGraw Hill, New York.

Joarder, A.H. (2002). On some representations of sample variance. *International Journal of Mathematical Education in Science and Technology*. 33(5), 772-784.

Joarder, A.H. (2003). Sample Variance and the first order differences. Technical Report No. 293. Dept of Mathematical Sciences, King Fahd University of Petroleum and Minerals, Saudi Arabia.

Kotz, S.; Kozubowski, T and Podgorski, K. (2001). *The Laplace Distribution and Generalizations*. Birkhauser, Boston, USA.

Ross, S.M. (1987). *Introduction to Probability and Statistics for Engineers and Scientists*. Wiley, New York.

File: p34msa.doc