King Fahd University of Petroleum & Minerals

DEPARTMENT OF MATHEMATICAL SCIENCES

Technical Report Series

TR293

April 2003

Sample Variance and the First Order Differences

Anwar H. Joarder

DHAHRAN 31261 ● SAUDI ARABIA ● www.kfupm.edu.sa/math/ ● E-mail: mathdept@kfupm.edu.sa

Sample Variance and the First Order Differences

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 sample variance can be calculated by the first order differences of sample observations via a matrix which is constant for any sample of a particular size. The constant matrix itself is open for further study. An alternative method is presented for the calculation of sample variance from a frequency distribution.

Key Words: Teaching; sample variance; difference table; reflection table; first order differences; pattern matrix.

1. Introduction

The variance $(s_n{}^2)$ of n observations in a sample is just the ratio of $TSS(n)$ (Total squared deviations corrected by the mean) to the degrees of freedom where

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

If sample observations are integers but not large in size, the last representation in (1.1) allows you to do the calculation mentally. The usual formula for variance depending on rounding off the sample mean lacks in precision, especially when computer programs are used for the calculation. The problem of calculating sample variance by avoiding the use of sample mean was posed by Ross (1987, 143-144) who offered a recurrence relation of sample variance. In the spirit of Ross (1987), some solutions to the problem were discussed by Joarder (2002).

The quantity $TSS\,(n)$ can also be represented by the following equivalent forms

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

(see e.g. Kotz, Kozubowski and Podgoriski, 2001, 186). Intuitively, the variability of a set of two observations say x_1 and x_2 should be reflected in the difference $|\, x_1 - x_2\,| = d$. Indeed for $n = 2$, it follows from (1.2) that $s_2^2 = (x_1 - x_2)^2/2 = d^2/2$ which is just $1/2$ times the square of the range.

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 $\binom{n}{2}$ distinct pairs and then averaging them. That is for a sample of size $n \geq 2$ the sample variance is given by:

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

This note presents some tabular way for the calculation of sample variance with some mathematical insight. In the spirit of Ross (1987), this note resorts to the first order differences of sample observations to calculate sample variance. The main result is presented in Theorem 3.1. The first order differences of sample observations via a symmetric constant matrix C (depending on the finite sample size n) can also be used for the calculation of variance. Notions are illustrated with hypothetical examples. The pattern matrix C itself is open for further study. An alternative method for the calculation of sample variance from a frequency distribution with equal class widths is presented which is only good if number of classes is small.

2. The Difference Table and the Reflection Table for the Calculation of Sample Variance

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

Example 2.1 To calculate variances of first $n = 2,3,4,5$ ordered observations of the sample (104, 94, 95, 101, 111), we may prepare the following Difference Table:

	94	95	101	104	111
94					
95	$1 = w_{21} = d_1$				
101	$7 = w_{31}$	$6 = w_{32} = d_2$			
104	$10 = w_{41}$	$9 = w_{42}$	$3 = w_{43} = d_3$		
111	$17 = w_{51}$	$16 = w_{52}$	$10 = w_{53}$	$7 = w_{54} = d_4$	

where we have used the notation $w_{i+1,i} = d_i \ \ (i = 1,2,3,4)$ which are in fact the first order differences of the ordered observations of the sample.

A table called **Reflection Table** can be prepared by the use of the ordered observations (in ascending order) in a column followed by columns where each element in a column is obtained by subtracting it from the smallest observation in the previous column. The following reflection table also provides the same set of differences of observations.

$x_{(i)}$	$r_i^{(1)} = x_{(i)} - x_{(1)}$ $(i = 2,3,4,5)$	$r_i^{(2)} = r_i^{(1)} - r_2^{(1)}$ $(i = 3,4,5)$	$r_i^{(3)} = r_i^{(2)} - r_3^{(2)}$ $(i = 4,5)$	$r_i^{(4)} = r_i^{(3)} - r_4^{(3)}$ $(i = 5)$
94				
95	$1 = 95 - 94$			
101	$7 = 101 - 94$	$6 = 7 - 1$		
104	$10 = 104 - 94$	$9 = 10 - 1$	$3 = 9 - 6$	
111	$17 = 111 - 94$	$16 = 17 - 1$	$10 = 16 - 6$	$7 = 10 - 3$

The variance of $n = 2,3,4,5$ observations is calculated below:

$$s_2^2 = w_{21}^2 / 2 = d_1^2 / 2 = 1/2,$$

$$s_3^2 = \frac{1}{3(2)}\left(w_{21}^2 + w_{31}^2 + w_{32}^2\right) = \frac{1}{3(2)}(1^2 + 7^2 + 6^2) = 43/3 \text{ and}$$

$$s_4^2 = \frac{1}{4(3)}\left(w_{21}^2 + w_{31}^2 + w_{32}^2 + w_{41}^2 + w_{42}^2 + w_{43}^2\right)$$

$$= \frac{1}{4(3)}(1^2 + 7^2 + 6^2 + 10^2 + 9^2 + 3^2) = 276/12 = 23.$$

Since $\sum_{i=2}^{5}\sum_{j=1}^{4} w_{ij}^2 = 1^2 + 7^2 + 10^2 + ... + 7^2 = 970$, it follows that

$$s_5^2 = \frac{1}{5(4)}(970) = \frac{970}{5(4)} = 48.5.$$

An arrangement of observations into ascending order while preparing the above tables produces nonnegative entries.

3. Variance and the First Order Differences of Observations

Consider a triangular matrix $W = ((w_{ij}))$, with elements $w_{ij} = x_i - x_j$, $(i = 1,2,...,n;$ $j = 1,2,...,n;\ i > j)$ as shown in the Difference Table in Section 2. 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 imaginary triangle is the sum of the elements in the corresponding part of the hypotenuse. For example for a sample of size $n = 5$, we have

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

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

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

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

$$w_{i+l,l} = x_{i+l} - x_i = (x_{i+l} - x_{i+l-1}) + (x_{i+l-1} - x_{i+l-2}) + \cdots + (x_{i+2} - x_{i+1}) + (x_{i+1} - x_i)$$
$$= d_{i+l-1} + d_{i+l-2} + \cdots + d_{i+1} + d_i = \sum_{j=i}^{i+l-1} d_j,\ i = 1,2,...,n-l \quad .$$

Then the elements in the $n-1$ diagonal lines (making the lower triangular matrix) can be represented by

$w_{i+1,i} = d_i$ (let), $i = 1,2,...,n-1$ (say, $n-1$ th diagonal line or the hypotenuse line),

$w_{i+2,i} = \sum_{j=i}^{i+1} d_j,\ i = 1,2,...,n-2$ (say, $n-2$ th diagonal line i.e. line below the hypotenuse line),

$w_{i+3,i} = \sum_{j=i}^{i+2} d_j,\ i = 1,2,...,n-3$ (say, $n-3$ th diagonal line),

… … …

$w_{i+n-2,i} = \sum_{j=i}^{i+n-3} d_j,\ i = 1,2$ (say, the second diagonal line),

$w_{i+n-1,i} = \sum_{j=i}^{i+n-2} d_j,\ i = 1$ (say, the first diagonal line).

Theorem 3.1 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 $n \geq 2$ observations is given by

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

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

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

$$n\,TSS(n) = \sum_{i=2}^{n} \sum_{j=1}^{i-1} w_{ij}^{\ 2}$$
$$= w_{21}^2 + (w_{31}^2 + w_{32}^2) + (w_{41}^2 + w_{42}^2 + w_{43}^2) + \cdots$$
$$+ (w_{n-1,1}^2 + w_{n-1,2}^2 + \cdots + w_{n-1,n-3}^2 + w_{n-1,n-2}^2) + \cdots$$
$$+ (w_{n1}^2 + w_{n2}^2 + \cdots + w_{n,n-2}^2 + w_{n,n-1}^2)$$

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

$$
nTSS(n) = \sum_{j=1}^{n-1} d_j^2 + \sum_{i=1}^{n-2} \left(\sum_{j=i}^{i+1} d_j \right)^2 + \sum_{i=1}^{n-3} \left(\sum_{j=i}^{i+2} d_j \right)^2 + \cdots + \sum_{i=1}^{2} \left(\sum_{j=i}^{i+n-3} d_j \right)^2 + \sum_{i=1}^{} \left(\sum_{j=i}^{i+n-2} d_j \right)^2 \quad (3.1)
$$

Since $nTSS(n) = n(n-1)s_n^2$, it follows from (3.1) 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 $c_{11} = c_{22} = 2,\ c_{12} = 1$,

$4(3)s_4^2 = d_1^2 + d_2^2 + d_3^2 + (d_1 + d_2)^2 + (d_2 + d_3)^2 + (d_1 + d_2 + d_3)^2 = d'Cd$ where $d' = (d_1, d_2, d_3)$ and $C = ((c_{ij}))$ with $c_{11} = c_{33} = 3,\ c_{12} = c_{23} = 2,\ c_{13} = 1$, and

$$
\begin{aligned}
5(4)s_5^2 &= d_1^2 + d_2^2 + d_3^2 + d_4^2 + (d_1 + d_2)^2 + (d_2 + d_3)^2 + (d_3 + d_4)^2 \\
&\quad + (d_1 + d_2 + d_3)^2 + (d_2 + d_3 + d_4)^2 + (d_1 + d_2 + d_3 + d_4)^2 = d'Cd
\end{aligned}
$$

where $d' = (d_1, d_2, d_3, d_4)$ and $C = ((c_{ij}))$ with $c_{11} = c_{23} = c_{44} = 4$,
$c_{12} = c_{34} = 3,\quad c_{13} = c_{24} = 2,\ c_{14} = 1$.

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 (3.1) looks complicated, the entire calculation can be made simple by ordering the sample observations and then preparing a table showing d_i's (the first order differences), in a column, followed by totals of consecutive pairs of $d_i's$ in the next column followed by totals of consecutive triplets of differences $d_i's$ in the next column and so on.

Corollary 3.1 For a sample of size $n \geq 2$ the following recurrence relation holds:

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

Proof. It follows from (3.1) that

$$(n+1)TSS$$
$$= \sum_{j=1}^{n} d_j^2 + \sum_{i=1}^{n-1}\left(\sum_{j=i}^{i+1} d_j\right)^2 + \sum_{i=1}^{n-2}\left(\sum_{j=i}^{i+2} d_j\right)^2 + \cdots + \sum_{i=1}^{2}\left(\sum_{j=i}^{i+n-2} d_j\right)^2 + \sum_{i=1}^{}\left(\sum_{j=i}^{i+n-1} d_j\right)^2$$
$$= \left[\sum_{j=1}^{n-1} d_j^2 + d_n^2\right] + \left[\sum_{i=1}^{n-2}\left(\sum_{j=i}^{i+1} d_j\right)^2 + (d_{n-1} + d_n)^2\right] + \left[\sum_{i=1}^{n-3}\left(\sum_{j=i}^{i+1} d_j\right)^2 + (d_{n-2} + d_{n-1} + d_n)^2\right] + \cdots$$
$$+ \left[\left(\sum_{i=1}^{n-1} d_i\right)^2 + \left(\sum_{i=2}^{n} d_i\right)^2\right] + \left(\sum_{i=1}^{n} d_i\right)^2$$
$$= nTSS(n) + d_n^2 + (d_{n-1} + d_n)^2 + (d_{n-2} + d_{n-1} + d_n)^2 + \cdots + \left(\sum_{i=2}^{n} d_i\right)^2 + \left(\sum_{i=1}^{n} d_i\right)^2$$

Example 3.1 To calculate variances of first $n = 2,3,4,5$ ordered observations of the sample (104, 94, 95, 101, 111), we may prepare the following table:

$x_{(i)}$	d_i $(i = 1,2,3,4)$	$d_i^{(2)} = d_i + d_{i+1}$ $(i = 1,2,3)$	$d_i^{(3)}$ $= d_i + d_{i+1} + d_{i+2}\,(i = 1,2)$	$d_i^{(4)} = \sum_{i=1}^{4} d_i$
94				
95	$1 = 95 - 94$			
101	$6 = 101 - 95$	$7 = 1 + 6$		
104	$3 = 104 - 103$	$9 = 6 + 3$	$10 = 1 + 6 + 3$	
111	$7 = 111 - 104$	$10 = 3 + 7$	$16 = 6 + 3 + 7$	$17 = 1 + 6 + 3 + 7$

The variance for sample sizes $n = 2,3,4,5$ are given below:

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

$$2\, s_2^2 = d_1^2 = 1,$$
$$3(2)\, s_3^2 = (1^2 + 6^2) + 7^2 = 86$$
$$4(3)\, s_4^2 = (1^2 + 6^2 + 3^2) + (7^2 + 9^2) + 10^2 = 276$$
$$5(4)\, s_5^2 = (1^2 + 6^2 + 3^2 + 7^2) + (7^2 + 9^2 + 10^2) + (10^2 + 16^2) + 17^2 = 970$$

so that variances are $s_2^2 = 1/2,\ s_3^2 = 43/3,\ s_4^2 = 23,\ s_5^2 = 97/2$.

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

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

3(2) $s_3^2 = d'Cd = 86$ where $d' = (d_1, d_2) = (1,6)$ and $C = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$,

4(3) $s_4^2 = d'Cd = 276$ where $d' = (d_1, d_2, d_3) = (1,6,3)$ and

$C = \begin{bmatrix} 3 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \end{bmatrix}$, and

5(4) $s_5^2 = d'Cd = 970$ where $d' = (d_1, d_2, d_3, d_4) = (1,6,3,7)$ and

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

(iii) *Calculation of variance by the Recurrence Relation*

$$2TSS(2) = d_1^2 = 1$$
$$3TSS(3) = 2TSS(2) + (d_1 + d_2)^2 + d_2^2 = 1 + 7^2 + 6^2 = 86$$
$$4TSS(4) = 3TSS(3) + (d_1 + d_2 + d_3)^2 + (d_2 + d_3)^2 + d_3^2 = 86 + 10^2 + 9^2 + 3^2 = 276$$
$$5TSS(5) = 4TSS(4) + (d_1 + d_2 + d_3 + d_4)^2 + (d_2 + d_3 + d_4)^2 + (d_3 + d_4)^2 + d_4^2$$
$$= 276 + 17^2 + 16^2 + 10^2 + 7^2 = 970$$

The variance can then be calculated as before.

4. Correlation Coefficient

The sum of products between two variables x and y can be variously written as

$$s_{xy} = \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)(y_i - y_j)$$
$$= \frac{1}{n}\sum_{1 \le i < j \le n}(x_i - x_j)(y_i - y_j) = \frac{1}{n}\sum_{i=2}^{n}\sum_{j=1}^{i-1}(x_i - x_j)(y_i - y_j)$$

(see e.g. Kotz, Kozubowski and Podgoriski, 2001, 186). The calculation can be done by difference table or preferably reflection table described Section 2. Then the correlation coefficient r can be calculated by $r\sqrt{s_{xx}}\sqrt{s_{ss}} = s_{xy}$.

Example 4.1 Consider the following grades of 5 students in Midterm and final exam:

$$(50, 60),\ (60, 70),\ (75, 90),\ (80,\ 80),\ (85,\ 90)$$

To calculate the correlation coefficient we prepare the following Reflection Table for x and y values:

	(50, 60)	(60, 70)	(75, 90)	(80, 80)	(85, 90)
(50, 60)					
(60, 70)	(10, 10)				
(75, 90)	(25, 30)	(15, 20)			
(80, 80)	(30, 20)	(20, 10)	(5, −10)		
(85, 90)	(35, 30)	(25, 20)	(10, 0)	(5, 10)	

Then $s_{xx} = \left(10^2 + 25^2 + \cdots + 5^2\right)/5 = 4250/5$, $s_{yy} = \left(10^2 + 30^2 + \cdots + 10^2\right)/5 = 3400/5$ and $s_{xy} = \left[10(10) + 25(30) + \cdots + 5(10)\right]/5 = 3500/5$. The correlation coefficient is then given by $r = \dfrac{3500}{\sqrt{(4250)(3400)}} \approx 0.92$.

5. Variance of Observations with frequencies

It is well known that variance of n consecutive integers is given by $n(n+1)/12$. Interestingly, the variance of any three consecutive integers is 1, and that of any 11 consecutive numbers is 11. This kind of result is sometimes important to construct examples quickly in classrooms. In this section we present an alternative method for calculating sample variance (s^2) from a frequency distribution. Note that 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 5.1 Let n 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

$$TSS(n) = \sum_{i=1}^{k} (f_i - 1)^2 \; s_i^2 + \sum_{i(<l)=1}^{k} (\bar{x}_i - \bar{x}_l)^2 \; \frac{f_i f_l}{n} \text{ where } \quad n = f_1 + f_2 + \cdots + f_k .$$

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 the variance of the ith group $s_i^2 = 0, (i = 1,2,\cdots,k)$ so that the first summand in the $TSS(n)$ in the above expression vanishes, and we have the following corollaries.

Corollary 5.1 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

$$TSS(n) = \sum_{i(<l)=1}^{k} (y_i - y_l)^2 \; \frac{f_i f_l}{n} .$$

Corollary 5.2 If k observations follow arithmetic series with common difference w and frequencies $f_i \; (i = 1,2,\cdots,k)$ then $TSS(n)$ is given by

$$TSS(n) = \frac{w^2}{n} \sum_{i(<l)=1}^{k} (i - l)^2 \; f_i f_l \tag{5.1}$$

$$= w^2 \left[\sum_{i=1}^{k} i^2 f_i - \frac{1}{n} \left(\sum_{i=1}^{k} i f_i \right)^2 \right] \tag{5.2}$$

Proof. It follows from Corollary 5.1 that

$$\begin{aligned} n \; TSS(n) &= (y_1 - y_2)^2 f_1 f_2 + (y_1 - y_3)^2 f_1 f_3 + \cdots + (y_1 - y_k)^2 f_1 f_k \\ &\quad + (y_2 - y_3)^2 f_2 f_3 + (y_2 - y_4)^2 f_2 f_4 + \cdots + (y_2 - y_k)^2 f_2 f_k \\ &\quad + \cdots + (y_{k-1} - y_k)^2 f_{k-1} f_k \\ &= [w^2 f_1 f_2 + (2w)^2 f_1 f_3 + \cdots + (k-1)^2 w^2 f_1 f_k] \\ &\quad + [w^2 f_2 f_3 + (2w)^2 f_2 f_4 + \cdots + (k-2)^2 w^2 f_2 f_k] + \cdots + [w^2 f_{k-1} f_k] \\ &= w^2 \sum_{i(<l)=1}^{k} (i - l)^2 \; f_i f_l \end{aligned}$$

The expression in (5.2) follows from the basic definition of sample variance by noting that the variance does not depend on the origin of transformation. Note that in this case variance depends on the observations only through the common difference, first k positive integers and corresponding frequencies.

Example 5.1 (Bluman, 2001, 113) Thirty automobiles were tested for fuel efficiency (in miles per gallon). The following frequency distribution was obtained.

Class boundaries	frequencies
$7.5-12.5$	3
$12.5-17.5$	5
$17.5-22.5$	15
$22.5-27.5$	5
$27.5-32.5$	2

It follows from (5.1) that

$$\begin{aligned} n\ TSS(n)/w^2 &= [f_1\ \{(1)^2 f_2 + (2)^2 f_3 + (3)^2 f_4 + (4)^2\ f_5\}] + [f_2\ \{(1)^2 f_3 + (2)^2\ f_4 + (3)^2 f_5\}] \\ &\quad + [f_3\ \{(1)^2\ f_4 + (2)^2 f_5\}] + [f_4\ \{(1)^2\ f_5\}] \\ &= [3\{1(5) + 4(15) + 9(5) + 16(2)\}] + [5\{1(15) + 4(5) + 9(2)\}] \\ &\quad + [15\{1(5) + 4(2)\}]\ + [5\{1(2)\}] = 896 \end{aligned}$$

so that $TSS(n) = 5^2(896)/30 = 746\frac{2}{3}$ and variance $s^2 = TSS(n)/(n-1) \approx 25.747$.

Alternatively, Since $\sum_{i=1}^{5} if_i = 1(3) + 2(5) + 3(15) + 4(5) + 5(2) = 88$ and

$\sum_{i=1}^{5} i^2 f_i = (1)^2(3) + (2)^2(5) + (3)^2(15) + (4)^2(5) + (5)^2(2) = 288$

it follows from (5.2) that $TSS(30) = 5^2\left[288 - (88)^2/30\right] = 746\frac{2}{3}$ so that the variance is 25.747 approximately.

Many special cases that can be deduced from (5.1) may be of much use in constructing examples quickly in classrooms. For example,

(i) If there are two observations with frequencies f_1 and f_2, and the observations are apart by w, then $TSS(n) = \dfrac{w^2}{n}\ f_1 f_2$. (5.3)

(ii) If there are three observations that follows arithmetic series with the common difference w, and they have frequencies f_1, f_2 and f_3, then

$$TSS(n) = \frac{w^2}{n}\ \left[f_1 f_2 + (2)^2 f_1 f_3 + f_2 f_3\right]\ .$$ (5.4)

(iii) If the observations follow arithmetic series with common difference w, and the frequencies of 5 classes are denoted by $a, 2a, 3a, 2a, a$ respectively then

$$TSS(n) = 12\ w^2\ a\ .$$ (5.5)

(iv) If the observations follow arithmetic series with common difference w, and the frequencies of 5 classes are denoted by $5a, 4a, 3a, 2a, a$ respectively then $TSS(n) = (70/3)\, w^2\, a$. (5.6)

Example 5.2 Let the grades of 5 students be given by 60, 70, 70, 60, 70. Since 60 (x_1) occurs twice $(f_1 = 2)$, 70 (x_2) occurs thrice $(f_2 = 3)$ and $w = 70 - 60 = 10$, it follows from (5.3) that $TSS(5) = \frac{10^2}{5}\ (2\times 3) = 120$ so that the variance is $s^2 = TSS/(n-1) = 30$.

Example 5.3 Let the grades of 27 students be given by

x	f
[50, 60)	3
[60, 70)	6
[70, 80)	9
[80, 90)	6
[90, 100)	3

Since the mid values of the classes follow an arithmetic series with common difference 5, and $a = 3$, it follows from (5.4) that $TSS(n) = 12w^2a = 12(10)^2(3) = 3600$ so that the variance is approximately 138.46.

The following corollary follows from (5.2) or directly by the basic definition of variance.

Corollary 5.3 If $n = kf$ observations follow arithmetic series with common difference w and common frequency f, then

$$TSS(n) = \frac{k(k+1)}{12}\ (n - f)w^2 \quad . \tag{5.6}$$

Proof. It is easy to check that

$$\begin{aligned}
&\sum_{i\,(<l)=1}^{k} (i-l)^2\ f_i\, f_l \\
&= \left[\left\{1^2 + 2^2 + \cdots + (1-k)^2\right\} + \{1^2 + 2^2 + \cdots + (2-k)^2\} + \cdots + \{1^2 + 2^2\} + 1^2\}\right] \\
&= \left[(k-1)1 + (k-2)2^2 + (k-3)3^2 + \cdots + \{k-(k-2)\}(k-2)^2 + \{k-(k-1)\}(k-1)^2\right] \\
&= k\sum_{i=1}^{k-1} i^2 - \sum_{i=1}^{k-1} i^3 = k\{(k-1)k(2k-1)/6\} - (k-1)^2k^2/4 = k^2(k^2-1)/12
\end{aligned}$$

since $\sum_{i=1}^{k} i^2 = k(k+1)(2k+1)/6$ and $\sum_{i=1}^{k} i^3 = k^2(k+1)^2/4$. The expression for $TSS(n)$ in (5.6) then follows by virtue of (5.1).

The corresponding variance is then given by

$$s_n^2 = \frac{k(k+1)}{12}\ \frac{n-f}{n-1}\ w^2 \qquad (5.7)$$

where $k(k+1)/12$, the first factor in (5.7), is the variance of k natural integers. The attachment of common frequency to each of the k integers is contributing to the second factor in the above expression. The variance of the first k positive integers each with common frequency f is given by (5.7) with $w = 1$

Acknowledgement

The author gratefully acknowledges the excellent research facilities available at King Fahd University of Petroleum and Minerals, Saudi Arabia.

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 Mathematics Education in Science and Technology*. 33(5), 772-784.

Kotz, S.; Kozubowski, T and Podgoriski, 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.