

SAMPLE VARIANCE AND FIRST-ORDER DIFFERENCES OF OBSERVATIONS

ANWAR H. JOARDER,* *King Fahd University of Petroleum and Minerals, Dhahran*

Abstract

It is proved that the variance of a sample can be calculated from first-order differences of its observations. It can also be represented by a quadratic form in the differences of the observations via a constant matrix (depending on the sample size) which is open for further study. Avoiding use 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.

Keywords: Sample variance; first-order differences; pattern matrix

AMS 2000 Subject Classification: Primary 62-99

1. Introduction

Since the usual method of calculating sample variance depends on rounding off the sample mean, it may lack precision, especially when computer programs are used for the calculation. The problem of calculating the variance by avoiding the use of the sample mean was posed by Ross (1987, pp. 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 $x_1, \dots, x_n$ is just the total squared deviations of observations from the mean (TSS) divided by the degrees of freedom, where

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

If sample observations are small integers, the last representation in (1.1) allows the calculation to be done mentally. The quantity TSS can also be represented by the following equivalent forms:

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

(see e.g. Kotz *et al.* (2001, p. 186)). The implication of (1.2) is that the variance of a sample of n observations can be easily obtained by calculating the variances of $\binom{n}{2}$ pairs of observations and then averaging them. That is, for a sample of size $n \geq 2$, the variance is given by

$$s_n^2 = \binom{n}{2}^{-1} \sum_{i=2}^n \sum_{j=1}^{i-1} \frac{w_{ij}^2}{2} \quad (1.3)$$

where $w_{ij} = x_i - x_j$ for $i, j = 1, 2, \dots, n$. This note resorts to the first-order differences of sample observations to calculate the variance. This can also be represented by a quadratic form

Received 16 March 2003; revision received 16 May 2003.

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

in the differences of sample observations via a constant matrix $\mathbf{C}$ (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 (1.3) that a table showing the differences between observations can be prepared whose entries, $w_{ij} = x_i - x_j$ for $i, j = 1, 2, \dots, n$ with $i > j$, are to be used 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 required values are $w_{21} = 1, w_{31} = 7, w_{32} = 6, w_{41} = 10, w_{42} = 9, w_{43} = 3, w_{51} = 17, w_{52} = 16, w_{53} = 10, w_{54} = 7$. Then by (1.3) we have

$$\sum_{i=2}^5 \sum_{j=1}^4 w_{ij}^2 = 1^2 + 7^2 + 6^2 + \dots + 7^2 = 970,$$

and so

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

2. Main results

Consider a lower triangular matrix $\mathbf{W} = (w_{ij})$, with nonzero elements $w_{ij} = x_i - x_j$ ($i, j = 1, 2, \dots, n, i > j$). Further, imagine the right-angled triangle with right angle at the bottom-left vertex w_{ij} and hypotenuse along the diagonal. Then the value of the bottom-left vertex w_{ij} is the sum of the elements $w_{(j+1)j}, \dots, w_{i(i-1)}$ in the hypotenuse of this triangle. For example, with a sample of size $n = 5$, if the first-order differences of the observations are $w_{21} = d_1, w_{32} = d_2, w_{43} = d_3, w_{54} = d_4$, then

$$\begin{aligned} w_{31} &= w_{21} + w_{32} = d_1 + d_2, \\ w_{41} &= w_{21} + w_{32} + w_{43} = d_1 + d_2 + d_3, & w_{42} &= w_{32} + w_{43} = d_2 + d_3, \\ w_{51} &= \sum_{i=1}^4 d_i, & w_{52} &= d_2 + d_3 + d_4, & w_{53} &= d_3 + d_4. \end{aligned}$$

In general let $w_{(i+1)i} = d_i$ ($i = 1, 2, \dots, n$). The elements of the l th ($l = 1, 2, \dots, n-1$) diagonal line of $\mathbf{W}$ are thus

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

Theorem 2.1. As above, let $d_i, i = 1, 2, \dots, n$, be the first-order differences of observations, and let $w_{ij} = \sum_{k=j}^{i-1} d_k, i > j$. Then the variance of $n \geq 2$ observations is given by

$$s_n^2 = \frac{1}{n(n-1)} \mathbf{d}^\top \mathbf{C} \mathbf{d},$$

where $\mathbf{d} = (d_1, d_2, \dots, d_{n-1})^\top$ and $\mathbf{C} = (c_{ij})$ is an $(n-1) \times (n-1)$ symmetric matrix with $c_{ij} = c_{ji} = (n-i)j$ for $i, j = 1, 2, \dots, n-1, i \geq j$.

Proof. From (1.1), $s_n^2 = (1/(n-1))TSS$. Starting from (1.2), we have

$$\begin{aligned}
 n \text{ TSS} &= \sum_{i=2}^n \sum_{j=1}^{i-1} w_{ij}^2 \\
 &= \sum_{i=2}^n \sum_{j=1}^{i-1} \sum_{k=j}^{i-1} \sum_{l=j}^{i-1} d_k d_l \\
 &= \sum_{i=2}^n \sum_{k=1}^{i-1} \sum_{j=1}^k \sum_{l=j}^{i-1} d_k d_l \\
 &= \sum_{k=1}^{n-1} \sum_{i=k+1}^n \sum_{l=1}^{i-1} \sum_{j=1}^{\min(k,l)} d_k d_l \\
 &= \sum_{k=1}^{n-1} \sum_{l=1}^{n-1} \sum_{i=\max(k+1,l+1)}^n \sum_{j=1}^{\min(k,l)} d_k d_l.
 \end{aligned}$$

But

$$\sum_{i=\max(k+1,l+1)}^n \sum_{j=1}^{\min(k,l)} 1 = \begin{cases} l(n-k) = c_{kl} & \text{if } k \geq l, \\ k(n-l) = c_{lk} = c_{kl} & \text{if } k \leq l, \end{cases}$$

so

$$\begin{aligned}
 n \text{ TSS} &= \sum_{k=1}^{n-1} \sum_{l=1}^{n-1} d_k c_{kl} d_l \\
 &= \mathbf{d}^\top \mathbf{C} \mathbf{d},
 \end{aligned}$$

as required.

The following corollary is obvious.

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

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

where $TSS(m)$ is the total sum of squares of m observations minus 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 $d_1 = 1, d_2 = 6, d_3 = 3, d_4 = 7$ are used as follows.

(i) Calculation of the variance by the first-order differences and their ‘totals’: from (1.3),

$$\begin{aligned}
 s_5^2 &= \frac{1}{n(n-1)} \sum_{i=2}^n \sum_{j=1}^{i-1} w_{ij}^2 \\
 &= \frac{1}{20} (1^2 + 6^2 + 3^2 + 7^2) + (7^2 + 9^2 + 10^2) + (10^2 + 16^2) + 17^2 = \frac{97}{2}.
 \end{aligned}$$

(ii) Calculation of the variance by the first-order differences and the constant matrix C : from Theorem 2.1 with $d = (1, 6, 3, 7)^\top$ and

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

we obtain $s_5^2 = \frac{1}{20} d^\top C d = \frac{97}{2}$.

(iii) Calculation of the variance by the recurrence relation: the variance can be calculated using Corollary 2.1 as

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

So $s_5^2 = \frac{1}{20} 5TSS(5) = \frac{97}{2}$, once more.

Similarly, other quantities involving sums of squares, such as the correlation coefficient in a bivariate sample, can also be calculated.

3. Variance of observations with frequencies

The variance of a sample classified into k classes in a frequency distribution with mid values $y_1, y_2, \dots, y_k$ and frequencies $f_1, f_2, \dots, f_k$ 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 $f_1, f_2, \dots, f_k$ observations with means $\bar{x}_{(1)}, \bar{x}_{(2)}, \dots, \bar{x}_{(k)}$ and variances $s_1^2, s_2^2, \dots, s_k^2$ respectively. Then*

$$TSS = \sum_{i=1}^k (f_i - 1)s_i^2 + n^{-1} \sum_{i(<l)=1}^k (\bar{x}_i - \bar{x}_l)^2 f_i f_l.$$

In the case where the number of observations in every group is the same, the means of k groups may be denoted by $y_1, y_2, \dots, y_k$. In this case, $s_i^2 = 0$ for $i = 1, \dots, k$, so that the first summand in the above expression for TSS vanishes, and we have the following corollaries.

Corollary 3.1. *Let $y_1, y_2, \dots, y_k$ have frequencies $f_1, f_2, \dots, f_k$ with $n = f_1 + f_2 + \dots + f_k$. Then*

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

TABLE 1: Frequency distribution.

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

Corollary 3.2. *If k observations follow an arithmetic series with common difference w and frequencies f_i , $i = 1, 2, \dots, k$, then*

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

The expression for 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 , the first k positive integers and their corresponding frequencies.

Example 3.1. (Bluman (2001, p. 113).) Thirty automobiles were tested for fuel efficiency (in miles per gallon). Table 1 shows the frequency distribution obtained.

Since $\sum_{i=1}^5 i f_i = 1 \times 3 + 2 \times 5 + 3 \times 15 + 4 \times 5 + 5 \times 2 = 88$ and $\sum_{i=1}^5 i^2 f_i = 288$, it follows from Corollary 3.2 that $TSS = 5^2(288 - 88^2/30) = 746\frac{2}{3}$ so that the variance is approximately 25.747.

Joarder (2003) deduced many special cases from Corollary 3.2 which may be of use in constructing classroom examples quickly. The following corollary follows from Corollary 3.2 or directly from the basic definition of variance.

Corollary 3.3. *If $n = kf$ observations follow an 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. The author also thanks an anonymous referee for constructive suggestions that have made the presentation of Theorem 2.1 and its proof very elegant.

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. *Internat. J. Math. Education Sci. Tech.* **33**, 772–784.
- JOARDER, A. H. (2003). Sample variance and the first-order differences. Tech. Rep. No. 293, Department of Mathematical Sciences, King Fahd University of Petroleum and Minerals.
- KOTZ, S., KOZUBOWSKI, T. AND PODGORSKI, K. (2001). *The Laplace Distribution and Generalizations*. Birkhäuser, Boston, MA.
- ROSS, S. M. (1987). *Introduction to Probability and Statistics for Engineers and Scientists*. John Wiley, New York.