



King Fahd University of Petroleum & Minerals

**DEPARTMENT OF MATHEMATICAL SCIENCES**

---

Technical Report Series

TR 269

September 2001

**On Some Representations of Sample Variance**

Anwar H. Joarder

# On Some Representations of Sample Variance

Anwar H. Joarder

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

**Abstract.** The usual formula of variance depending on the rounding off the sample mean lacks in precision especially when computer programs are used for the calculation. The simplification of the total sums of squares does not always benefit. Since the variance of two observations is easily calculated without the use of sample mean, and the variance of a sample of  $n$  observations is the average of variances of  $n(n-1)/2$  distinct subsets of size 2 from the sample, it is argued that this sense of pairing may result in precision. Some other forms of variance have been presented which provides some insight into it. Contribution of a new observation to variance is highlighted which is important in sequential sampling. Notions are illustrated with examples.

**Keywords:** sample variance, recurrence relation of variance, distinct pairing, computational efficiency

## 1. Introduction

The variance is a measure of variability that exists in a sample. There are two important reasons for measuring variability. The first reason is how well the average value depicts the data. A second reason is to learn the extent of scatter so that steps may be taken to control the existing variation. For example, while maintaining a long average mileage is the most important objective of the manufacturer of a tire, he tries to improve the uniformity in the mileage of it through better inspection and other quality control procedures; otherwise some customers would be satisfied and some would remain upset. This is desired in many real world situations (Kolman, Anton and Averbach, 1992, 312).

The sample variance is one of the very basic notion a student learn in the beginning week of a statistics course. But to many it is a tongue twister, and most frustratingly, its meaning has nothing to do with the mathematical expression of the definition or the way we calculate it. Can we explain it in esay-to-understand terms? It is based on deviations of observations from the sample mean denoted by  $x_i - \bar{x}$ , ( $i = 1, 2, \dots, n$ ). These do help

understand the variation in the sample observations. For a sample (4,5,11,14) the deviations are given by  $-4.5, -3.5, 2.5, 5.5$ .

What information do these distances or deviations  $x_i - \bar{x}$ , ( $i = 1, 2, \dots, n$ ) contain? If they tend to be large in absolute values, the data are spread out or highly variable. If they are mostly small in the absolute sense, the data are clustered around the sample mean and therefore do not exhibit much variability. A deviation indicates the amount by which an observation is away from the sample mean. Thus the deviation ' $-4.5$ ' indicates that there is an observation in the sample which is 4.5 units below the sample mean. Similarly the deviation ' $5.5$ ' indicates that there is an observation which is 5.5 units above the sample mean.

Now the question is how to condense the information on deviations so as to form a single numerical measure of variability. Note that the deviations always add to zero and as such we do not gain any information with the total given by

$$\sum_{i=1}^n (x_i - \bar{x}) = 0.$$

To gauge the variability of the observations in a sample, we, however, do not care whether an observation is below or above the sample mean, all we care about is whether an observation is away from the sample mean. Thus we may use the absolute deviations or the squared deviations. The measure of variability produced by the absolute deviations did not gain popularity because on the one hand it presents analytic difficulties and on the other hand it does not bring any benefit while compared with its counterpart. If we square the deviation  $-4.5$  it would be 20.25. The latter implies there is an observation which is  $\sqrt{20.25} = 4.5$  units away from the mean. The measure of variability produced by squared deviations, known as variance, indicates the variability of a set of observations around their mean.

The sum of the squared deviations is variously known as Total Sums of Squares (TSS), Corrected Sums of Squares (CSS) or simply as Sums of Squares (SS), and can be mathematically written as

$$TSS = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 = \sum_{i=1}^n (x_i - \bar{x})^2$$

where the sum is over the sample observations. This denotes the total variation among observations in a sample. For the above sample TSS is given by

$$TSS = (-4.5)^2 + (-3.5)^2 + (2.5)^2 + (5.5)^2 = 69$$

The deviations are not always symmetric around zero though they should add to zero. However, because of round-off error, the sum of the deviations may not be exactly zero. It may be remarked here that the fact that deviations add to zero implies that if  $n-1$  of them are known, the other one is automatically determined. This number  $n-1$  is called the degrees of freedom of the sample or of the sample mean or of TSS.

The variance ( $s_n^2$ ) of the observations in a sample of size  $n$  is just the ratio of the total squared deviations to the degrees of freedom i.e.

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

Obviously  $0 \leq s_n^2 < \infty$ . In case  $n = 1$ , the variance is usually defined to be 0.

Most statisticians use a simplified form of variance given by (2.1). In this paper some different forms of variance have been represented with the hope of shedding more light. Though most of them are scattered in different text books, neater proofs of related theorems have been presented. Most importantly a new direction is emphasized for calculating variance that avoids using the sample mean and thereby guarantee least rounding off error. Since the variance of two observations is easily calculated without the use of sample mean, and the variance of a sample of  $n$  observations is the average of variances of  $n(n-1)/2$  distinct subsets of size 2 from the sample, it is argued that this sense of pairing may result in precision. The result is implicit in Lindgren (1993, 206). Recurrence relation of variance, which is important in sequential sampling for quality controls in industry, is also emphasized for calculation. A grouping or pairing technique is introduced. Notions are illustrated with hypothetical examples.

## 2. Some Representations of Sample Variance

In this section we present seven different forms of variance. Though they are algebraically they same, they do differ in precision and time it takes to calculate them. Let  $\bar{x}_n$  and  $s_n^2$  be sample mean and variance of  $n$  observations respectively.

### 2.1 A Simplified Formula

The TSS can be simplified as

$$TSS = \sum (x - \bar{x}_n)^2 = \sum x^2 - \frac{(\sum x)^2}{n}. \quad (2.1)$$

If sample observations are integers but not large in size, this representation allows you to do the calculation mentally.

Since  $TSS = (n-1)s^2$  and  $\sum x = n\bar{x}$ , it follows from the above representation that

$$\sum x^2 = (n-1)s^2 + n\bar{x}^2. \quad (2.2)$$

### 2.2 Recurrence Relation Depending on Sample Mean

A representation of variance due to Ross (1987, p 143) is presented below with an elegant proof.

**Theorem 2.1** For  $n \geq 2$  the following recurrence relation holds:

$$s_{n+1}^2 = \left(1 - \frac{1}{n}\right) s_n^2 + \frac{1}{n+1} (x_{n+1} - \bar{x}_n)^2 \quad (2.3)$$

**Proof:** It is easy to check that

$$n s_{n+1}^2 = \sum_{i=1}^{n+1} (x_i - \bar{x}_{n+1})^2 = \sum_{i=1}^n (x_i - \bar{x}_{n+1})^2 + (x_{n+1} - \bar{x}_{n+1})^2 \quad (2.4)$$

Since  $(n+1)\bar{x}_{n+1} = \sum_{i=1}^{n+1} x_i = n\bar{x}_n + x_{n+1}$  we have  $\bar{x}_{n+1} = \bar{x}_n + \frac{x_{n+1} - \bar{x}_n}{n+1}$ .

It then follows that

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x}_{n+1})^2 &= \sum_{i=1}^n \left[ (x_i - \bar{x}_n) - \frac{x_{n+1} - \bar{x}_n}{n+1} \right]^2 \\ &= (n-1) s_n^2 + \frac{n}{(n+1)^2} (x_{n+1} - \bar{x}_n)^2 \end{aligned}$$

$$\text{and } x_{n+1} - \bar{x}_{n+1} = x_{n+1} - \frac{1}{n+1} (n\bar{x}_n + x_{n+1}) = \frac{n}{n+1} (x_{n+1} - \bar{x}_n).$$

The proof is immediate by plugging the above two identities back in (2.4).

### 2.3 Distinct Pairing (Variance Without Sample Mean)

Intuitively, the variability of a set of two observations say  $x_1$  and  $x_2$  should be reflected in the difference  $(x_1 - x_2)$ . Indeed for  $n = 2$ , it follows from (1.1) that

$$s_2^2 = \frac{(x_1 - x_2)^2}{2}$$

which is just  $1/2$  times the square of the range. In what follows we prove a theorem implicit in Bernard (1993, 2000) 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.

**Theorem 2.2** For a sample of size  $n \geq 2$  the following result hold:

$$s_n^2 = \frac{1}{\binom{n}{2}} \sum_{i(>j)=1}^n \sum_{j=1}^n \frac{(x_i - x_j)^2}{2} \quad (2.5)$$

**Proof.** It is easily checked that

$$\sum_{i(>j)=1}^n \sum_{j=1}^n (x_i - x_j)^2 = (n-1) \sum_{i=1}^n x_i^2 - 2 \sum_{i(>j)=1}^n \sum_{j=1}^n x_i x_j \quad (2.6)$$

$$\text{and } \left( \sum_{i=1}^n x_i \right)^2 = \sum_{i=1}^n x_i^2 + 2 \sum_{i(>j)=1}^n \sum_{j=1}^n x_i x_j. \quad (2.7)$$

Then it follows that

$$\begin{aligned} \sum_{i(>j)=1}^n \sum_{j=1}^n (x_i - x_j)^2 &= (n-1) \sum_{i=1}^n x_i^2 - \left[ \left( \sum_{i=1}^n x_i \right)^2 - \sum_{i=1}^n x_i^2 \right] \\ &= n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2 = n \sum_{i=1}^n (x_i - \bar{x})^2 \end{aligned} \quad (2.8)$$

The proof is then completed by dividing both sides of (2.8) by  $(n-1)$ .

It follows from Theorem 2.2 that a table showing the differences among observations can be made whose entries are  $w_{ij} = x_i - x_j$  ( $i, j = 1, 2, \dots, n$ ). Then

$$s_n^2 = \frac{1}{\binom{n}{2}} \sum_{i(>j)=1}^n \sum_{j=1}^n \frac{w_{ij}^2}{2} = \frac{1}{n(n-1)} \sum_{i(>j)=1}^n \sum_{j=1}^n w_{ij}^2 = \frac{1}{2n(n-1)} \sum_{i=1}^n \sum_{j=1}^n w_{ij}^2 \quad (2.9)$$

where the factor  $n(n-1) = n^2 - n$  is the number of off-diagonal elements of the matrix with elements  $w_{ij}$  ( $i, j = 1, 2, \dots, n$ ). Note that  $\frac{(n-1)s_n^2}{n} = V$ , the right hand side of (2.9), is the second sample moment as reported by Bernard (1993, 206).

#### 2.4 Variance Depending on Distinct Pairing and Sample Mean

The following theorem is a direct consequence of (2.3) and (2.5).

$$\textbf{Theorem 2.3} \quad s_{n+1}^2 = \frac{1}{n^2} \sum_{i(>j)=1}^n \sum_{j=1}^n (x_i - x_j)^2 + \frac{1}{n+1} (x_{n+1} - \bar{x}_n)^2 \quad (2.10)$$

#### 2.5 Recurrence Relation of Variance Without Sample Mean

The following Recurrence Relation follows from Theorem 2.2.

$$\begin{aligned}
\binom{n+1}{2} s_{n+1}^2 &= \sum_{i(>j)=1}^{n+1} \sum_{j=1}^{n+1} \frac{(x_i - x_j)^2}{2} \\
&= \sum_{i(>j)=1}^n \sum_{j=1}^n \frac{(x_i - x_j)^2}{2} + \sum_{j=1}^n \frac{1}{2} (x_{n+1} - x_j)^2 \\
&= \binom{n}{2} s_n^2 + \frac{1}{2} \sum_{i=1}^n (x_{n+1} - x_i)^2
\end{aligned} \tag{2.11}$$

If the above recurrence relation is used in conjunction with Distinct Pairing (Theorem 2.2), i.e the expression in the middle of (2.11) is used, the sample variance is calculated without the sample mean. Avoidance of sample mean may result in precision since the sample mean usually needs to be rounded off .

## 2.6 Variance by Grouping

The variance or the TSS can be calculated by grouping the sample observations, calculating variance of different groups and finally combining them by the following theorem. However, it is usually proved by labeling the observations with two suffixes which can be avoided since means and variances of groups are all that we need. The proof is made neater further by the use of identity in (2.2).

**Theorem 2.4** Let  $n$  observations be divided into  $k$  groups containing  $n_1, n_2, \dots, n_k$  observations with means  $\bar{x}_{(1)}, \bar{x}_{(2)}, \dots, \bar{x}_{(k)}$  respectively. Then the TSS will be given by

$$TSS = \sum_{i=1}^k (n_i - 1) s_{(i)}^2 + \sum_{i(<l)=1}^k \frac{n_i n_l}{n} (\bar{x}_{(i)} - \bar{x}_{(l)})^2 \tag{2.12}$$

**Proof:** Let the observations be divided into two groups (*i.e.*  $k = 2$ ) containing  $n_1$  and  $n_2$  observations with means  $\bar{x}_{(1)}$  and  $\bar{x}_{(2)}$ , and variances  $s_{(1)}^2$  and  $s_{(2)}^2$  respectively. Then

$$\begin{aligned}
TSS &= \sum_{i=1}^{n_1+n_2} (x_i - \bar{x}_n)^2 = \sum_{i=1}^{n_1+n_2} x_i^2 - (n_1 + n_2) \bar{x}_n^2 = \sum_{i=1}^{n_1} x_i^2 + \sum_{i=n_1+1}^{n_1+n_2} x_i^2 - (n_1 + n_2) \bar{x}_n^2 \\
&= \left\{ (n_1 - 1) s_{(1)}^2 + n_1 \bar{x}_{(1)}^2 \right\} + \left\{ (n_2 - 1) s_{(2)}^2 + n_2 \bar{x}_{(2)}^2 \right\} - \frac{(n_1 \bar{x}_{(1)} + n_2 \bar{x}_{(2)})^2}{n_1 + n_2} \\
&= (n_1 - 1) s_{(1)}^2 + (n_2 - 1) s_{(2)}^2 + \frac{n_1 n_2}{n_1 + n_2} (\bar{x}_{(1)} - \bar{x}_{(2)})^2
\end{aligned}$$

Similarly for 3 groups we have

$$\begin{aligned}
TSS &= \sum_{i=1}^{n_1+n_2+n_3} (x_i - \bar{x}_n)^2 = \sum_{i=1}^{n_1} x_i^2 + \sum_{i=n_1+1}^{n_1+n_2} x_i^2 + \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} x_i^2 - (n_1 + n_2 + n_3) \bar{x}_n^2 \\
&= (n_2 - 1) s_{(1)}^2 + (n_2 - 1) s_{(2)}^2 + (n_3 - 1) s_{(3)}^2 \\
&\quad + \frac{n_1 n_2}{n_1 + n_2} (\bar{x}_{(1)} - \bar{x}_{(2)})^2 + \frac{n_1 n_3}{n_1 + n_3} (\bar{x}_{(1)} - \bar{x}_{(3)})^2 + \frac{n_2 n_3}{n_2 + n_3} (\bar{x}_{(2)} - \bar{x}_{(3)})^2
\end{aligned}$$

The proof for  $k$  groups is thus obvious.

Since sample mean or group means in this context needs to be rounded, we prefer to use totals of the pairs (groups) to avoid rounding errors as much as possible. Let  $T_{(i)} = n_i \bar{x}_{(i)}$ , the total of the  $i$ th group. Then the following form may be helpful in calculating TSS.

$$\begin{aligned}
TSS &= \sum_{i=1}^k (n_i - 1) s_{(i)}^2 + \sum_{i(>l)=1}^k \frac{n_i n_l}{n} \left( \frac{T_{(i)}}{n_i} - \frac{T_{(l)}}{n_l} \right)^2 \\
&= \sum_{i=1}^k (n_i - 1) s_{(i)}^2 + \sum_{i(>l)=1}^k \frac{(n_i T_{(i)} - n_l T_{(l)})^2}{n n_i n_l} \quad (2.13)
\end{aligned}$$

For ease of calculation by hand the following representation may be better.

$$TSS = \sum_{i=1}^k (n_i - 1) s_{(i)}^2 + \sum_{i(>l)=1}^k \frac{1}{n n_i n_l} \begin{vmatrix} n_i & T_{(i)} \\ n_l & T_{(l)} \end{vmatrix}^2 \quad (2.14)$$

The first  $k$  terms in the Theorem is the contribution of the observations due to variation within groups (VWG), while the next  $\binom{k}{2}$  terms can be attributed to the variation

between groups (VBG). Groups having less variation among the observations within the groups may be used to have smaller contribution by VWG and more by VBG. Groups having more variation among the observations within the groups may be used to have larger contribution by variation within the groups (VWG) and less by VBG. Samples having modal observation with high frequencies may be a good example for the first case (See Section 4.5). The idea of attributing the variation here is much similar to what led Fisher (1947) to discover the analysis of variance .

If variance of every group vanishes, the overall variance will be given by the second summand in (2.14). If , on the other hand, the sample observations are grouped in a way that the group means are the same then

$TSS = \sum_{i=1}^k (n_i - 1) s_{(i)}^2$  so that the variance will be  $s^2 = \frac{1}{n-1} \sum_{i=1}^k (n_i - 1) s_{(i)}^2$  which is a weighted sum of group variances.



## 2.7 Variance by Pairing

If sample size is 2, the sample variance is  $\frac{1}{2}$  times the square of the range. This suggests us that the grouping technique in (2.14) can be used to calculate variance by choosing  $n_i = 2$  ( $i = 1, 2, \dots, k-1$ ) and  $n_k = 2$  or 1 depending on whether sample size is even or odd. Let  $T_{(i)} = n_i \bar{x}_{(i)}$ , the total of the  $i$ th group, and  $s_{(i)}^2 = w_i^2 / 2$ , the variance of the  $i$ th pair. For ease of calculation by hand the following representation that follow from (2.13) and (2.14), may be better.

$$TSS = \sum_{i=1}^k \frac{w_i^2}{2} + \sum_{i(>l)=1}^k \frac{1}{n n_i n_i} \left| \begin{array}{c} n_i \quad T_{(i)} \\ n_i \quad T_{(i)} \end{array} \right|^2 \quad (2.15)$$

If sample observations are paired in a way that the means of pairs are the same then

$$TSS = \sum_{i=1}^k \frac{w_i^2}{2} \text{ so that the variance will be } s^2 = \frac{1}{n-1} \sum_{i=1}^k \frac{w_i^2}{2}. \text{ Note that } w_i = 0 \text{ if } n_i = 1.$$

## 2.8 Geometric Interpretation

The pair  $\bar{x}_n$  and  $s_n^2$  can be derived from the Euclidean minimization problem. Suppose that for observations  $x_i$  ( $i = 1, 2, \dots, n$ ), we want to find the value of  $t$  that minimize  $\sum (x_i - t)^2$ . Note that the sums of squares is the square of the Euclidean distance in  $n$  dimensions between  $n$ -dimensional point  $(t, t, \dots, t)$  and the observations expressed as the point  $(x_1, x_2, \dots, x_n)$ . The minimization problem is amenable to calculus, but algebra is all that is needed here. Since  $x_i$ 's are known sample values, the expression  $\sum (x_i - t)^2 = \sum (x_i^2 - 2x_i t + t^2) = \sum x_i^2 - 2t \sum x_i + nt^2$  is quadratic in  $t$ . Its minimum occurs at its vertex and algebra shows that the minimum is  $\sum (x_i - \bar{x}_n)^2$ . This is the TSS we discussed in Section 1. Interested readers may go through an stimulating paper by Farnsworth (2000).

## 3. Contribution of a New Observation

3.1 The contribution of a new observation  $x_{n+1}$  in the variance formula of Theorem 2.1 is given by

$$c_{n+1} = s_{n+1}^2 - s_n^2 = -\frac{1}{n} s_n^2 + \frac{1}{n+1} (x_{n+1} - \bar{x}_n)^2. \quad (3.1)$$

If the new observation  $x_{n+1} = 0$ , then it follows from (3.1) that the contribution of the new observation to the variance is given by

$$c_{n+1} = \frac{\bar{x}_n^2}{n+1} - \frac{s_n^2}{n}. \quad (3.2)$$

If the new observation  $x_{n+1} = \bar{x}$ , then it follows from (3.1) that the contribution of the new observation to the variance is given by

$$c_{n+1} = -\frac{s_n^2}{n}. \quad (3.3)$$

Thus the quantity  $c_{n+1}$  satisfies the following

$$-\frac{s_n^2}{n} \leq c_{n+1} \leq \frac{\bar{x}^2}{n+1} - \frac{s_n^2}{n} \quad (3.4)$$

3.2 The contribution of a new observation  $x_{n+1}$  in the variance formula of Theorem 2.2 can be calculated easily from (2.11) as follows:

$$\begin{aligned} c_{n+1} &= s_{n+1}^2 - s_n^2 = \frac{1}{\binom{n+1}{2}} \sum_{i=1}^n \frac{(x_{n+1} - x_i)^2}{2} - \frac{2}{n+1} s_n^2 \\ &= \frac{1}{(n+1)n} \sum_{i=1}^n (x_{n+1} - x_i)^2 - \frac{2}{n+1} s_n^2 \end{aligned} \quad (3.5)$$

If the new observation  $x_{n+1} = 0$ , then it follows from (3.5) and  $\sum_{i=1}^n x_i^2 = n\bar{x}^2 + (n-1)s_n^2$  that the contribution of the new observation to the variance is also given by (3.2). If the new observation  $x_{n+1} = \bar{x}$ , i.e. it is the same as the original sample mean, then it follows from (3.4) that its contribution to the variance is given by  $c_{n+1} = -\frac{s_n^2}{n}$ . Thus the quantity  $c_{n+1}$  satisfies the inequality in (3.4).

## 4. Some Illustrations

Consider calculating the variance of the sample (4, 5, 11, 14) sequentially by the use of different representations we have in Section 2.

### 4.1 Variance by Recurrence Relation Depending on Sample Mean

By the use of Theorem 2.1

$$s_2^2 = \frac{1}{2}(x_2 - x_1)^2 = \frac{1}{2}(5 - 4)^2 = \frac{1}{2}, \quad \bar{x}_2 = \frac{9}{2}$$

$$s_{2+1}^2 = \left(1 - \frac{1}{2}\right) s_2^2 + \frac{1}{2+1} (x_{2+1} - \bar{x}_2)^2 = \frac{1}{2} \frac{1}{2} + \frac{1}{3} \left(11 - \frac{9}{2}\right)^2 = \frac{43}{3}, \quad \bar{x}_3 = \frac{20}{3}$$

$$s_{3+1}^2 = \left(1 - \frac{1}{3}\right) s_3^2 + \frac{1}{3+1} (x_{3+1} - \bar{x}_3)^2 = \frac{2}{3} \frac{43}{3} + \frac{1}{4} \left(14 - \frac{20}{3}\right)^2 = 23.$$

To see the contribution by a new observation let us assume that we already have 3 observations with variance  $s_3^2 = \frac{43}{3}$  and a new observation say  $x_{3+1} = 14$ . It follows from (3.1) that

$$c_{3+1} = s_{3+1}^2 - s_3^2 = -\frac{1}{3} s_3^2 + \frac{1}{3+1} (x_{3+1} - \bar{x}_3)^2 = -\frac{1}{3} \frac{43}{3} + \frac{1}{4} \left(14 - \frac{20}{3}\right)^2 = \frac{26}{3}$$

$$\text{so that } s_{3+1}^2 = s_3^2 + c_{3+1} = \frac{43}{3} + \frac{26}{3} = 23.$$

#### 4.2 Variance by Distinct Pairing (Variance Without Sample Mean)

Consider calculating the variance of the sample (4, 5, 11) first. The distinct pairs would result in the subsets: (4,5), (4,11) and (5,11). The ranges are: 1, 7 and 6. The variance of the sample is

$$s_3^2 = \frac{1}{3} \left[ \frac{(1)^2}{2} + \frac{(7)^2}{2} + \frac{(6)^2}{2} \right] = \frac{1}{3} \left( \frac{86}{2} \right) = \frac{43}{3}$$

Similarly for the sample (4,5,11,14), the variance is given by

$$s_4^2 = \frac{1}{6} \left[ \frac{(1)^2}{2} + \frac{(7)^2}{2} + \frac{(6)^2}{2} + \frac{(10)^2}{2} + \frac{(9)^2}{2} + \frac{(3)^2}{2} \right] = \frac{1}{6} \left( \frac{276}{2} \right) = 23$$

The above can better be calculated by preparing the following difference table:

$x$	Difference	Difference	Difference	$n$	$\binom{n}{2}$	$s_n^2$
4				1		
5	1			2	1	$\frac{1}{1} \left( \frac{1}{2} \right)$
11	7	6		3	3	$\frac{1}{3} \left( \frac{86}{2} \right)$
14	10	9	3	4	6	$\frac{1}{6} \left( \frac{276}{2} \right)$

The arrangement of the sample observations in ascending order results in nonnegative differences in the table.

### 4.3 Variance Depending on Distinct Pairing and Sample Mean

Consider calculating the variance of the sample (4, 5, 11) by Theorem 2.3.

$$s_2^2 = \frac{1}{2}(5-4)^2 = \frac{1}{2}$$

$$\begin{aligned} s_{2+1}^2 &= \frac{1}{2^2}(x_2 - x_1)^2 + \frac{1}{2+1}(x_3 - \bar{x}_2)^2 \\ &= \frac{1}{4}(5-4)^2 + \frac{1}{3}\left(11 - \frac{9}{2}\right)^2 = \frac{43}{3} \end{aligned}$$

$$\begin{aligned} s_{3+1}^2 &= \frac{1}{3^2}[(x_2 - x_1)^2 + (x_3 - x_1)^2 + (x_3 - x_2)^2] + \frac{1}{3+1}(x_4 - \bar{x}_3)^2 \\ &= \frac{1}{4}[1^2 + 7^2 + 6^2] + \frac{1}{4}\left(14 - \frac{20}{3}\right)^2 = 23 \end{aligned}$$

### 4.4 Variance by Recurrence Relation Without Sample Mean (see 2.11)

By the use of (2.11), we have

$$s_2^2 = \frac{(5-4)^2}{2} = \frac{1}{2},$$

$$\begin{aligned} \binom{2+1}{2} s_{2+1}^2 &= \binom{2}{2} s_2^2 + \frac{1}{2} \sum_{i=1}^3 (x_{2+1} - x_i)^2 \\ &= \frac{1}{2} + \frac{1}{2} [(x_3 - x_1)^2 + (x_3 - x_2)^2] = \frac{1}{2} + \frac{1}{2} (7^2 + 6^2) = 43 \end{aligned}$$

so that  $s_3^2 = \frac{43}{3}$ .

$$\begin{aligned} \binom{3+1}{2} s_{3+1}^2 &= \binom{3}{2} s_3^2 + \frac{1}{2} \sum_{i=1}^3 (x_{n+1} - x_i)^2 \\ &= 3 \left(\frac{43}{3}\right) + \left[\frac{(10)^2}{2} + \frac{(9)^2}{2} + \frac{(3)^2}{2}\right] = 43 + 95 \end{aligned}$$

so that  $s_{3+1}^2 = \frac{1}{6}(43+95) = 23$ . To see the contribution by a new observation let us

assume that we already have 3 observations with variance  $s_3^2 = \frac{43}{3}$  and a new

observation say  $x_{3+1} = 14$ . It follows from (3.5) that

$$c_{3+1}^2 = \frac{1}{4(3)} \left[ \frac{(10)^2}{2} + \frac{(9)^2}{2} + \frac{(3)^2}{2} \right] - \frac{2}{3+1} s_3^2 = \frac{104}{12} = \frac{26}{3} \text{ so that } s_{3+1}^2 = \frac{43}{3} + \frac{26}{3} = 23.$$

#### 4.5 Variance by Grouping

Thus to calculate the variance of grades of 9 students (40, 70, 95, 70, 50, 70, 90, 70, 70) may be grouped as (40, 50), (90, 95) and (70, 70, 70, 70, 70) for smaller VWG.

$s_{(i)}^2$ (=VWG)	$i$	pairs	$n_i$	$T_{(i)}$	VBG
$50 = (50 - 40)^2 / 2$	1	(40, 50)	2	90	$\frac{1}{9(2)(2)} \left  \begin{array}{c} 2 \quad 90 \\ 2 \quad 185 \end{array} \right ^2 = 1002.77..$
$12.5 = (95 - 90)^2 / 2$	2	(90, 95)	2	185	$\frac{1}{9(2)(1)} \left  \begin{array}{c} 2 \quad 90 \\ 5 \quad 350 \end{array} \right ^2 = 694.44...$
0	3	70, 70, 70, 70, 70	1	350	$\frac{1}{9(2)(1)} \left  \begin{array}{c} 2 \quad 185 \\ 5 \quad 350 \end{array} \right ^2 = 0562.5$
62.5					2259.722...

The variance is given by  $s^2 = \frac{1}{8}(62.5 + 2259.722...) = \frac{2322.222...}{8} = 4290.273$ .

#### 4.6 Variance by Pairing

To calculate the variance of (4,5,11,14,20) we group them as (4, 20), (5, 14), (11) for larger contribution by variation within groups (VWG). The following table is prepared to apply the formula in (2.15).

$s_{(i)}^2$ (=VWG)	$i$	Pairs	$n_i$	$T_{(i)}$	VBG
$128 = (4 - 20)^2 / 2$	1	(4, 20)	2	24	$\frac{1}{5(2)(2)} \left  \begin{array}{cc} 2 & 24 \\ 2 & 19 \end{array} \right ^2 = 5$
$40.5 = (5 - 14)^2 / 2$	2	(5, 14)	2	19	$\frac{1}{5(2)(1)} \left  \begin{array}{cc} 2 & 24 \\ 2 & 11 \end{array} \right ^2 = 0.4$
0	3	(11)	1	11	$\frac{1}{5(2)(1)} \left  \begin{array}{cc} 2 & 19 \\ 1 & 11 \end{array} \right ^2 = 0.9$
168.5					6.3

The variance is given by  $s^2 = \frac{1}{4}(168.5 + 6.3) = \frac{174.8}{4} = 43.7$ .

## 5. Conclusion

When calculating variance by hand some representations may prove to be much efficient. However if the sample size is large and the computation is performed on a computer, then because of 'round-off error' some methods will be more efficient than the others. It is not surprising if the representation ( 2.2) provides negative value for sample variance (Ross, 1987, 143). Methods that avoid using sample mean ( say equation 2.5 or 2.11) to the extent possible may result in much precision. Different grouping or pairing techniques (2.14 or 2.15) may also be devised for the same. It remains open to check the precision of various methods by computer programs.

## References

- Fansworth, D.L. (2000). The Geometry of Statistics. *The College Mathematics Journal*. 31(3), 200-204.
- Fisher, R.A. (1947). *The Design of Experiments* (4<sup>th</sup> ed.), Oliver and Boyd, Edinburgh.
- Kolman, B., Anton, H. and Averbach, B. (1992), *Mathematics with Applications for the Management, life and Social Sciences*, (Philadelphia: Saunders College Publishing)
- Lindgren, B.W. (1993). *Statistical Theory*. Chapman and Hall.
- Ross, S.M. (1987), *Introduction to Probability and Statistics for Engineers and Scientists*. (New York: Wiley).