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

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 well known 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 observations based on $n(n-1)/2$ distinct subsets of units 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 provide some insight into it. Contribution of a new observation to variance is highlighted which is important in sequential sampling. Notions are illustrated with examples.

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 notions a student learns 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 it is calculated. Can we explain it in easy-to-understand terms? It is based on deviations of observations from the sample mean denoted by $x_i - \bar{x},\ (i = 1, 2, ..., n)$. These do help understand the variation in the sample observations. For a sample of (4,5,11,14), the sample mean $\bar{x} = 8.5$ so that the deviations are given by $x_1 - \bar{x} = 4 - 8.5 = -4.5,$

[*] *Published in International Journal of Mathemaatical Education in Science and Technology, 33(5), 772-784.(London, UK)*

$x_2 - \bar{x} = 5 - 8.5 = -3.5,\ \ x_3 - \bar{x} = 11 - 8.5 = 2.5,\ \ x_4 - \bar{x} = 14 - 8.5 = 5.5$. Another sample with the same mean of 8.5 may have different variability e.g. the sample (7, 10, 8, 9) also has a mean of 8.5 but the deviations are $7 - 8.5 = -1.5,$ $10 - 8.5 = 1.5$, $8 - 8.5 = -0.5,\ \ 9 - 8.5 = 0.5$ which do not exhibit as much variability as they do in the previous sample.

What information do these distances or deviations $x_i - \bar{x},\ (i = 1, 2, \cdots, n)$ 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

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

To gauge the variability of the observations in a sample all we care about is whether an observation is away from the sample mean, be it below or above it. 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 the sample 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:

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

where the sum is over all the sample observations. This denotes the total variation among observations in a sample. For the sample (4, 5, 11, 14), 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 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 as defined below:

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

Obviously $0 \leq s_n^2 < \infty$. In case $n = 1,$ the variance is usually defined to be 0. If all the observations were the same, each deviation would have been zero, so would have been the variance. If , however, the observations are widely apart, so will be the deviations producing positive TSS or positive variance. Thus the smaller (larger) the deviations in absolute value, the smaller (larger) is the variance, and vice versa.

The variance of the sample (4, 5, 11, 14) is $s_4^2 = 69/(4-1) = 23$. The variance of the second sample, producing deviations that are relatively less widely apart compared to that of the first sample, is approximately 1.67 which is, as expected, much lesser than that of the first sample.

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 into the nature of variance. 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 guarantees 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 observations based on $n(n-1)/2$ distinct subsets of units of size 2 from the sample, it is argued that this sense of pairing may result in precision. The result is implicit in many texts (see e.g. Lindgren , 1993, 206). Recurrence relation of variance, which is important in sequential sampling for quality control 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 the 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

$(n-1)s_n^2 = TSS$ can be represented by the following equivalent forms:

$$\begin{aligned}(n-1)s_n^2 &= \frac{1}{n}\sum_{1\le i<j\le n}(x_i-x_j)^2 = \frac{1}{n}\sum_{1\le j<i\le n}(x_i-x_j)^2 = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-x_j)^2 \\ &= \frac{1}{n}\sum_{i=2}^{n}\sum_{j=1}^{i-1}(x_i-x_j)^2 = \frac{1}{n}\sum_{i=1}^{n-1}\sum_{j=1}^{i+1}(x_i-x_j)^2 = \sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2 .\end{aligned} \tag{2.1}$$

If sample observations are integers but not large in size, the last representation allows one to do the calculation mentally. Since $\sum_{i=1}^{n}x_i = n\bar{x}_n$, it follows from (2.1) that

$$\sum_{i=1}^{n}x_i^2 = (n-1)s_n^2 + n\,\bar{x}_n^2 . \tag{2.2}$$

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 $n \ge 2$ the following recurrence relation holds:

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

Proof: It is easy to check that :

$$\begin{aligned} n\,s_{n+1}^2 &= \sum_{i=1}^{n+1}(x_i-\bar{x}_{n+1})^2 \\ &= \left[(x_1-\bar{x}_{n+1})^2 + (x_2-\bar{x}_{n+1})^2 + \cdots + (x_n-\bar{x}_{n+1})^2\right] + (x_{n+1}-\bar{x}_{n+1})^2 \\ &= \sum_{i=1}^{n}(x_i-\bar{x}_{n+1})^2 + (x_{n+1}-\bar{x}_{n+1})^2 \end{aligned} \tag{2.4}$$

where $\bar{x}_{n+1}$ is the sample mean based on $n+1$ observations. Since $(n+1)\bar{x}_{n+1} = \sum_{i=1}^{n+1} x_i = n\bar{x}_n + x_{n+1}$ it follows that $\bar{x}_{n+1} = \frac{1}{n+1}(n\bar{x}_n + x_{n+1}) = \bar{x}_n + \frac{x_{n+1} - \bar{x}_n}{n+1}$ and consequently we have:

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

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).

Thus if the first n observations are known, a value x_{n+1} can be obtained if a particular variance s_{n+1}^2 is desired.

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 present a neater proof of a theorem implicit in many texts (see e.g. Lindgren, 1993, 206) that the variance of a sample of n observations can be easily calculated by calculating the variances of $\binom{n}{2}$ distinct pairs of observations and then averaging them.

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

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

Proof. Since

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

and $\left(\sum_{i=1}^{n} x_i\right)^2 = \sum_{i=1}^{n} x_i{}^2 + 2\ \sum_{i=2}^{n}\sum_{j=1}^{i-1} x_i x_j$, (2.7)

it follows that :

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

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 prepared whose entries are $w_{ij} = x_i - x_j \ \ (i, j = 1, 2, ..., n)$. Then

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

(2.9)

where the factor $n(n-1) = n^2 - n$ is the number of off-diagonal elements of the matrix with elements $w_{ij} \ (i, j = 1, 2, ..., n)$. Note that $\frac{(n-1)s_n^2}{n} = V$ (see 2.8), is the second sample moment reported by Lindgren (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) .

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

2.5 Recurrence Relation of Variance Without Sample Mean

The following Recurrence Relation follows from Theorem 2.2.:

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

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.

2.6 Variance by Grouping

The variance or 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 further neater by the use of identity in (2.2).

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

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

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:

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

Similarly for 3 groups we have:

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

The proof for k groups is thus obvious.

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

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

which will be reduced to $(n-1)\sum_{i=1}^{k} s_{(i)}^2 + \frac{1}{n}\sum_{i(>l)=1}^{k}\left(T_{(i)} - T_{(l)}\right)^2$ if the group sizes are the same. For ease of calculation by hand the following representation may be better:

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

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 the 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 the sample variance is given by :

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{k} (n_i - 1)\, s_{(i)}^2$$

which is a weighted sum of group variances.

2.7 Variance by Pairing

If sample size is 2, the sample variance is ½ times the square of the range. This suggests us that the grouping technique in (2.14) can be used to calculate variance by choosing $n_i = 2\,(i = 1, 2, \cdots, k-1)$ 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 follows from (2.14), may be better:

$$TSS = \sum_{i=1}^{k} \frac{w_i^2}{2} + \sum_{i(>l)=1}^{k} \frac{1}{n\, n_i\, n_l} \begin{vmatrix} n_i & T_{(i)} \\ n_l & T_{(l)} \end{vmatrix}^2 \qquad (2.15)$$

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

$TSS = \sum_{i=1}^{k} \frac{w_i^2}{2}$ so that the variance will be $s^2 = \frac{1}{n-1} \sum_{i=1}^{k} \frac{w_i^2}{2}$. Note that $w_i = 0$ 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 $x_i\ (i = 1,2,\cdots,n)$, we want to find the value of t that minimizes $\sum_{i=1}^{n} (x_i - t)^2$. Note that the sums of squares $\sum_{i=1}^{n} (x_i - t)^2$ is the square of the Euclidean distance in n dimensions between n-dimensional point $(t, t, \cdots, t)$ and the observations expressed as the point $(x_1, x_2, \cdots, x_n)$. 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 $\sum_{i=1}^{n}(x_i - t)^2 = \sum_{i=1}^{n}(x_i^2 - 2x_i\, t + t^2) = \sum_{i=1}^{n} x_i^2 - 2t\sum_{i=1}^{n} x_i + nt^2$ is quadratic in t. Its minimum occurs at its vertex and algebra shows that the minimum is $\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$. 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

Let c_{n+1} be the contribution of a new observation x_{n+1} to any variance formula based on n observations so that

$$s_{n+1}^2 = s_n^2 + c_{n+1}\,. \tag{3.1}$$

3.1 The contribution of a new observation $x_{n+1} = \bar{x}_n \mp d$ to the variance formula of Theorem 2.1 is given by:

$$c_{n+1} = -\frac{{s_n}^2}{n} + \frac{d^2}{n+1} \tag{3.2}$$

which is minimized if $d = 0$ i.e. if $x_{n+1} = \bar{x}_n$. If the new observation is at most d units away from $\bar{x}_n$, i.e. $\bar{x}_n - d \le x_{n+1} \le \bar{x}_n + d$, then it follows from (3.2) that c_{n+1} satisfies the following bounds:

$$-\frac{1}{n}\; s_n^2 \le c_{n+1} \le -\frac{{s_n}^2}{n} + \frac{d^2}{n+1}\,. \tag{3.3}$$

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

$$c_{n+1} = \frac{1}{(n+1)n}\,\sum_{i=1}^{n}(x_{n+1} - x_i)^2 \;-\; \frac{2}{n+1}\; s_n^2 \tag{3.4}$$

which also satisfies the bounds given by (3.3).

Theorem 3.1 If the new observation is $x_{n+1} = \bar{x}_n \mp d$, then for any expression of variance s_n^2 , the following recurrence relation holds:

$$s_{n+1}^2 = \frac{n-1}{n}\ s_n^2 + \frac{d^2}{n+1} = \frac{n-1}{n}\ s_n^2 + \frac{\left(x_{n+1} - \bar{x}_n\right)^2}{n+1} .$$
(3.5)

The contribution $c_{n+1} = s_{n+1}^2 - s_n^2$ of a new observation x_{n+1} is

(ii) negative if $\bar{x}_n - \sqrt{\frac{n+1}{n}}\ s_n < x_{n+1} < \bar{x}_n + \sqrt{\frac{n+1}{n}}\ s_n$,

(ii) zero if $x_{n+1} = \bar{x}_n \mp \sqrt{\frac{n+1}{n}}\ s_n$ and

(iii) positive elsewhere.

and is minimized at $x_{n+1} = \bar{x}_n$ (in which case it is $(n-1)n^{-1}\ s_n^2 < s_n^2$)

4. Some Illustrations

4.1 Variance by Recurrence Relation Depending on Sample Mean

To calculate the variance of the sample (4, 5, 11, 14) sequentially by Theorem 2.1, we have:

$$s_2^2 = \frac{1}{2}\left[(4\text{-}9/2)^2 + (5-9/2)^2\right] = 1/2\ , \quad \bar{x}_2 = 9/2\ ,$$

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

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

To see the contribution of a new observation to the variance formula in Theorem 2.1 let us assume that we already have 3 observations (4, 5, 11) with variance $s_3^2 = 43/3$ and a new observation say $x_{3+1} = 14$. It follows from (3.2) that:

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

so that by (3.1) we have $s_{3+1}^2 = s_3^2 + c_{3+1} = 43/3 + 26/3 = 23.$

4.2 Variance by Distinct Pairing (Variance Without Sample Mean)

To calculate the variance of the sample (4, 5, 11, 14) sequentially by Theorem 2.2, we have:

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

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

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

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

x	4	5	11	14
4				
5	$5-4=1$			
11	$11-4=7$	$11-5=6$		
14	$14-4=10$	$14-5=9$	$14-11=3$	

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

To see the contribution of a new observation $x_{3+1} = 14$ in the variance formula in Theorem 2.2 let us again assume that we already have 3 observations (4, 5, 11) with variance $s_3^2 = 43/3$ and a new observation say $x_{3+1} = 14$. It follows from (3.4) that:

$c_{3+1} = \frac{1}{4(3)}\left[\frac{(10)^2}{2} + \frac{(9)^2}{2} + \frac{(3)^2}{2}\right] - \frac{2}{3+1}s_3{}^2 = 104/12 = 26/3$ so that by (3.1) we have

$$s_{3+1}^2 = 43/3 + 26/3 = 23.$$

4.3 Variance Depending on Distinct Pairing and Sample Mean

To calculate the variance of the sample (4, 5, 11, 14) sequentially by Theorem 2.3, we have:

$$s_2^2 = 0 + \frac{1}{2}\ (5-4)^2 = \frac{1}{2}\,\mathbf{,}$$

$$s_{2+1}^2 = \frac{1}{2^2}(x_2 - x_1)^2 + \frac{1}{2+1}(x_3 - \bar{x}_2)^2$$

$$= \frac{1}{4}(5-4)^2 + \frac{1}{3}\left(11 - \frac{9}{2}\right)^2 = 43/3,$$

$$s_{3+1}^2 = \frac{1}{3^2}\left[(x_2 - x_1)^2 + (x_3 - x_1)^2 + (x_3 - x_2)^2\right] + \frac{1}{3+1}(x_4 - \bar{x}_3)^2$$

$$= \frac{1}{4}\left(1^2 + 7^2 + 6^2\right) + \frac{1}{4}\left(14 - \frac{20}{3}\right)^2 = 23.$$

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

To calculate the variance of the sample (4, 5, 11, 14) sequentially by equation 2.11, we have

$$s_2^2 = 0 + \frac{(5-4)^2}{2}\ ,$$

$$\binom{2+1}{2} s_{2+1}^2 = \binom{2}{2}\ s_2^2 + \frac{1}{2}\sum_{i=1}^{3}(x_{2+1} - x_i)^2$$

$$= \frac{1}{2} + \frac{1}{2}\left[(x_3 - x_1)^2 + (x_3 - x_2)^2\right] = \frac{1}{2} + \frac{1}{2}\ (7^2 + 6^2) = 43\ ,$$

$$\binom{3+1}{2} s_{3+1}^2 = \binom{3}{2}\ s_3^2 + \frac{1}{2}\sum_{i=1}^{3}(x_{n+1} - x_i)^2$$

$$= 3\ \left(\frac{43}{3}\right) + \left[\frac{(10)^2}{2} + \frac{(9)^2}{2} + \frac{(3)^2}{2}\right] = 43 + 95$$

so that $s_2^2 = 1/2$, $s_3^2 = 43/3$ and $s_{3+1}^2 = \frac{1}{6}\left(43 + 95\right) = 23$.

4.5 Variance by Grouping

To calculate the variance of grades of 9 students (40, 70, 95, 70, 50, 70, 90, 70, 70) by (2.14), the sample may be grouped as (40, 50), (90, 95) and (70, 70, 70, 70, 70) for smaller VWG.

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

The variance is given by $s^2 = (62.5 + 2259.722...)/8 \approx 290.278$.

4.6 Variance by Pairing

To calculate the variance of (4,5,11,14,20) by (2.15), 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	$\frac{1}{5(2)(2)}\begin{vmatrix}2 & 24\\2 & 19\end{vmatrix}^2 = 5$
$40.5 = (5-14)^2/2$	2	(5, 14)	2	19	$\frac{1}{5(2)(1)}\begin{vmatrix}2 & 24\\1 & 11\end{vmatrix}^2 = 0.4$
0	3	(11)	1	11	$\frac{1}{5(2)(1)}\begin{vmatrix}2 & 19\\1 & 11\end{vmatrix}^2 = 0.9$
168.5					6.3

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

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 last representation of equation (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 along the line may also be devised for the same. It remains open to check the relative efficiency of various methods by computer programs.

Acknowledgements

The author acknowledges the excellent research facilities available at King Fahd University of Petroleum and Minerals, Saudi Arabia. The author also thanks an anonymous referee, and my colleague Dr. A. Laradji for constructive suggestions that have improved the presentation of the paper.

References

Farnsworth, D.L., 2000, The Geometry of Statistics. *The College Mathematics Journal*. **31**(3), 200-204.

Fisher, R.A., 1947, *The Design of Experiments* (4th ed.), (Edinburgh :Oliver and Boyd).

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*. (London: Chapman and Hall).

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

File: c \ paper\ p30a.doc