King Fahd University of Petroleum & Minerals

DEPARTMENT OF MATHEMATICAL SCIENCES

Technical Report Series

TR 328

May 2005

The Expected Sample Variance in a General Situation

Anwar H. Joarder

The Expected Sample Variance in a General Situation

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 expected value of sample variance is often calculated by deriving the sampling distribution of sample variance which is difficult for many distributions. In this note it is demonstrated that only the distribution of the difference of any two observations is needed for the expected value of sample variance. An alternative method is also presented and some examples are provided for illustration.

Key Words: sample variance; expected value of sample variance; population variance

1. Introduction

Let $X_1, X_2, \cdots, X_n (n \geq 2)$ be a sample, not necessarily random (e.g. Joarder and Ahmed, 1998). Then the sample variance (s^2) is defined by $(n-1)s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2, n \geq 2$ while the second sample moment is given by $(1-1/n)s^2$. A matrix W showing the differences among observations can be prepared whose entries are $w_{ij} = x_i - x_j$ where i and j are integers $(i, j = 1, 2, ..., n)$ so that the set of elements of $W = \{w_{ij}: 1 \leq i \leq n, 1 \leq j \leq n\}$ can be 'decomposed' as W_l, W_d and W_u which are the elements in the lower triangle, diagonal and upper triangle of the matrix W where

$$\begin{aligned} W_l &= \{w_{ij}: 1 \leq i > j \leq n\} \\ &= \{w_{ij}: 1 \leq i \leq n, 1 \leq j \leq n; i > j\} \\ &= \{w_{ij}: 2 \leq i \leq n, 1 \leq j \leq i-1\} \\ &= \{w_{ij}: 1 \leq j \leq n-1, j+1 \leq i \leq n\} \end{aligned} \tag{1.1}$$

$$W_d = \{w_{ii} = 0: 1 \leq i \leq n\}$$

$$\begin{aligned} W_u &= \{w_{ij}: 1 \leq i < j \leq n\} \\ &= \{w_{ij}: 1 \leq i \leq n, 1 \leq j \leq n; i < j\} \\ &= \{w_{ij}: 1 \leq i \leq n-1, i+1 \leq j \leq n\} \\ &= \{w_{ij}: 2 \leq j \leq n, 1 \leq i \leq j-1\} \end{aligned} \tag{1.2}$$

The n^2 elements of the matrix W can be decomposed as $(n-1)n/2$, n and $(n-1)n/2$ where $(n-1)n/2$ is the number elements in W_l or W_u. Hence or otherwise

$$s^2 = \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i-1} \frac{w_{ij}^{\ 2}}{2} = \frac{1}{n(n-1)} \sum_{i=2}^{n} \sum_{j=1}^{i-1} (x_i - x_j)^2 = \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}^{\ 2} \tag{1.3}$$

See for example Kotz, Kozubowski and Podgorski (2002, 186) and Joarder (2003). This is also evident by the following algebra:

$$\begin{aligned} n\sum_{i=1}^{n}(x_i - \bar{x})^2 &= n \sum_{i=1}^{n} x_i^{\ 2} - \left(\sum_{i=1}^{n} x_i \right)^2 \\ &= n \sum_{i=1}^{n} x_i^{\ 2} - \left(\sum_{i=1}^{n} x_i^{\ 2} + 2 \sum_{i=2}^{n} \sum_{j=1}^{i-1} x_i x_j \right) \\ &= (n-1)\left(x_1^{\ 2} + x_2^{\ 2} + \cdots + x_n^{\ 2} \right) - 2 \sum_{1 \le i < j \le n} x_i x_j \end{aligned}$$

In case of independently and identically distributed randomly variables, often the expected value of sample variance is calculated by deriving the distribution of the random sample variance. If a sample is drawn from a normal population $N(\mu, \sigma^2)$, then, it is well known that the sample mean $(\bar{X})$ and variance (S^2) are independent and $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, a chi-square distribution with $(n-1)$ degrees of freedom and that $E(S^2) = (n-1)^{-1}\sigma^2 E[(n-1)S^2/\sigma^2] = (n-1)^{-1}\sigma^2 E(\chi^2_{n-1}) = \sigma^2$ (see for example Lindgren, 1993, 213).

But if X_i $(i = 1,2,..,n)$'s are uncorrelated random variables with finite mean $E(X_i) = \mu$ $(i = 1,2,...,n)$ and finite variance $V(X_i) = E(X_i - \mu)^2 = \sigma^2$ $(i = 1,2,\cdots,n)$, it also follows that $E\left(S^2\right) = \sigma^2$ by virtue of $(n-1)S^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$ and $E(\bar{X} - \mu)^2 = V(\bar{X}) = \sigma^2/n$.

In this note we demonstrate that the sampling distribution of the sample variance can be avoided to derive the expected value of sample variance in many general situations. An alternative method is also presented in Theorem 2.1 with some corollaries and examples.

2. Main Results

With the help of the representations in (1.3) of the above sample variance, we find an elegant expressions for the expected value of sample variance which seems to be very natural. In what follows we will need the following:

$$n(n-1)\sigma_\mu^2 = \sum_{i=2}^{n} \sum_{j=1}^{i-1} (\mu_i - \mu_j)^2 \tag{2.1}$$

$$n\overline{\sigma^2} = \sum_{i=1}^{n} \sigma_i^2 \tag{2.2}$$

Theorem 2.1 Let $X_i\ (i=1,2,..,n)$'s be random variables with finite mean $E(X_i)=\mu_i\ \ (i=1,2,...,n)$ and finite variance $V(X_i)=\sigma_i^2\ \ (I=1,2,\cdots,n)$ with $Cov(X_i,X_j)=\sigma_{ij}=\rho_{ij}\sigma_i\sigma_j\ ,\ \ (i=1,2,...n;j=1,2,...,n;i\neq j)$. Then

(a) $n(n-1)E\left(S^2\right)=\sum_{i=2}^{n}\sum_{j=1}^{i-1}E(X_i-X_j)^2=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}E(X_i-X_j)^2$

(b) $E\left(S^2\right)=\overline{\sigma^2}-\binom{n}{2}^{-1}\sum_{i=2}^{n}\sum_{j=1}^{i-1}\rho_{ij}\sigma_i\ \ \sigma_j+\sigma_\mu^2$

where σ_μ^2 and $\overline{\sigma^2}$ are defined by (2.1) and (2.2) respectively.

Proof The part (a) is obvious by (1.3). By writing $x_i-x_j=(x_i-\mu_i)-(x_j-\mu_j)+(\mu_i-\mu_j)$, it can be checked that

$$\begin{aligned} n(n-1)s^2=\sum_{i=2}^{n}\sum_{j=1}^{i-1}[&(x_i-\mu_i)^2+(x_j-\mu_j)^2+(\mu_i-\mu_j)^2 \\ &-2(x_i-\mu_i)(x_j-\mu_j)+2(x_i-\mu_i)(\mu_i-\mu_j)-2(x_j-\mu_j)(\mu_i-\mu_j)] \end{aligned} \tag{2.2}$$

Clearly $\sum_{i=2}^{n}\sum_{j=1}^{i-1}E(X_i-\mu_i)^2=\sum_{i=2}^{n}(i-1)\sigma_i^2$. Since $2\le j+1\le i\le n$ (See 1.1),

$$\sum_{i=2}^{n}\sum_{j=1}^{i-1}E(X_j-\mu_j)^2=\sum_{j=1}^{n-1}\sum_{i=j+1}^{n}E(X_j-\mu_j)^2=\sum_{j=1}^{n-1}(n-j)E(X_j-\mu_j)^2=\sum_{j=1}^{n-1}(n-j)\sigma_j^2,$$

$\sum_{i=2}^{n}\sum_{j=1}^{i-1}E(X_i-\mu_i)(X_j-\mu_j)=\sum_{i=2}^{n}\sum_{j=1}^{i-1}\rho_{ij}\sigma_i\ \ \sigma_j$ and $\sum_{i=2}^{n}\sum_{j=1}^{i-1}(\mu_i-\mu_j)^2=n(n-1)\sigma_\mu^2$ in analogy of (1.3), it follows from (2.1) that

$$n(n-1)E\left(S^2\right)=\sum_{i=2}^{n}(i-1)\sigma_i^2+\sum_{j=1}^{n-1}(n-j)\sigma_j^2+n(n-1)\sigma_\mu^2\ -2\sum_{i=2}^{n}\sum_{j=1}^{i-1}\rho_{ij}\sigma_i\ \ \sigma_j-0-0$$

The proof for part (b) is complete by noting that $\sum_{i=2}^{n}(i-1)\sigma_i^2+\sum_{j=1}^{n-1}(n-j)\sigma_j^2=(n-1)\sum_{i=1}^{n}\sigma_i^2$.

Example 2.1: Consider the probability density function $f(x_1,x_2,x_3)=\frac{2}{3}(x_1+x_2+x_3),\ \ 0<x_1,x_2,x_3<1$. It can be checked that $f(x_i)=\frac{2}{3}(x_i+1),\ 0<x_i<1\ \ E(X_i)=5/9,\ \ E(X_i^2)=7/18,,\ E(X_iX_j)=11/36$, $V(X_i)=13/162$ and $Cov(X_i,X_j)=-1/324\ \ (i=1,2,3;j=1,2,3;i\neq j)$ (Hardle and Simar, 2003, 128). Then by Theorem 2.1(a)

$6E(S^2)=E(X_1-X_2)^2+E(X_1-X_3)^2+E(X_2-X_3)^2=3\left[7/18+7/18-2(11/36)\right]=1/2$

so that $E(S^2)=1/12$ where $S^2=\sum_{i=1}^{3}(X_i-\bar{X})^2/2$.

Example 2.2 Let X_j $(j=1,2,..,n)$'s be independently, identically and normally distributed as $N(\mu,\sigma^2)$. Since $X_i - X_j \sim N(0,\ 2\sigma^2)$, by Theorem 2.1(a), we have

$$n(n-1)E(S^2) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}E(X_i - X_j)^2 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(0^2 + 2\sigma^2\right) = n(n-1)\sigma^2 .$$

The following corollaries are obvious from Theorem 2.1(b).

Corollary 2.1 Let X_i $(i=1,2,..,n)$'s be random variables with $E(X_i)=\mu_i$, $V(X_i)=\sigma_i^2$, $Cov(X_i,X_j)=\rho\sigma_i^2$, $(i=1,2,...n; j=1,2,...,n; i\neq j)$ whenever they exist. Then $E\left(S^2\right)=\overline{\sigma^2}-\binom{n}{2}^{-1}\rho\sum_{i=2}^{n}(i-1)\sigma_i^2$ where $\overline{\sigma^2}$ is defined by (2.1). Note that $E\left(S^2\right)$ does not depend on μ_i $(i=1,2,\cdots,n)$.

Corollary 2.2: Let X_i $(i=1,2,..,n)$'s be uncorrelated random variables with finite mean $E(X_i)=\mu_i$ $(i=1,2,...,n)$ and finite variance $V(X_i)=\sigma_i^2$ $(I=1,2,\cdots,n)$. Then $E\left(S^2\right)=\overline{\sigma^2}+\sigma_\mu^2$ where σ_μ^2 and $\overline{\sigma^2}$ are defined by (2.1) and (2.2) respectively.

Corollary 2.3 Let X_i $(i=1,2,..,n)$'s be random variables with $E(X_i)=\mu_i$, $V(X_i)=\sigma^2$, $Cov(X_i,X_j)=\rho_{ij}\sigma^2$ $(i=1,2,...n; j=1,2,...,n; i\neq j)$ whenever they exist. Then $E\left(S^2\right)=(1-\rho_{ij}^2)\sigma^2+\sigma_\mu^2\leq 2\sigma^2+\sigma_\mu^2$ where σ_μ^2 is defined by (2.1)

Corollary 2.4 Let X_i $(i=1,2,..,n)$'s be random variables with $E(X_i)=\mu_i$, $V(X_i)=\sigma^2$, $Cov(X_i,X_j)=\rho\sigma^2$ $(i=1,2,...n; j=1,2,...,n; i\neq j)$ whenever they exist. Then $E\left(S^2\right)=(1-\rho)\sigma^2+\sigma_\mu^2\leq 2\sigma^2+\sigma_\mu^2$ where σ_μ^2 is defined by (2.1)

Corollary 2.5 Let X_i $(i=1,2,..,n)$'s be random variables with $E(X_i)=\mu_i$, $V(X_i)=\sigma^2$, $Cov(X_i,X_j)=0$ $(i=1,2,...n; j=1,2,...,n; i\neq j)$ whenever they exist. Then $E\left(S^2\right)=\sigma^2+\sigma_\mu^2$ where σ_μ^2 is defined by (2.1).

Example 2.3 Let X_i $(i=1,2,..,n)$'s be normally distributed as $N(\mu_i,\sigma^2)$,. $(i=1,2,\cdots,n)$ with $\rho_{ij}=0,\ (i,j=1,2,\cdots,n; i\neq j)$, The probability density function can be written along that in Example 2.7. By Corollary 2.5, we have $E\left(S^2\right)=\sigma^2+\sigma_\mu^2$ where σ_μ^2 is defined by (2.1).

Corollary 2.6 Let X_i $(i=1,2,..,n)$'s be random variables with $E(X_i)=\mu$, $V(X_i)=\sigma_i^2$, $Cov(X_i,X_j)=\rho_{ij}\sigma_i\sigma_j$, $(i=1,2,...n; j=1,2,...,n; i\neq j)$ whenever they exist. Then $E\left(S^2\right)=\overline{\sigma^2}-\binom{n}{2}^{-1}\sum_{i=2}^{n}\sum_{j=1}^{i-1}\rho_{ij}\sigma_i\ \sigma_j$ where $\overline{\sigma^2}$ is defined by (2.1).

Further if $Cov(X_i, X_j) = \rho\sigma_i^2$, $(i = 1,2,...n; j = 1,2,...,n; i \neq j)$, then

$E\left(S^2\right) = \overline{\sigma^2} - \binom{n}{2}^{-1} \rho \sum_{i=2}^{n} (i-1)\sigma_i^2$ where $\overline{\sigma^2}$ is defined by (2.1).

Corollary 2.7 Let X_i $(i = 1,2,..,n)$'s be random variables with $E(X_i) = \mu$, $V(X_i) = \sigma_i^2$, $Cov(X_i, X_j) = 0$, $(i = 1,2,...n; j = 1,2,...,n; i \neq j)$ whenever they exist. Then $E\left(S^2\right) = \overline{\sigma^2}$ where $\overline{\sigma^2}$ is defined by (2.2).

Corollary 2.8 Let X_j $(j = 1,2,..,n)$'s be independently distributed with finite mean $E(X_i) = \mu$ $(i = 1,2,...,n)$ and finite variance $V(X_i) = \sigma_i^2$ $(i = 1,2,\cdots,n)$. Then $E(S^2) = \overline{\sigma^2}$ where $\overline{\sigma^2}$ is defined by (2.2).

Corollary 2.9 Let X_i $(i = 1,2,..,n)$'s be random variables with $E(X_i) = \mu$, $V(X_i) = \sigma^2$, $Cov(X_i, X_j) = \rho_{ij}\sigma^2$, $(i = 1,2,...n; j = 1,2,...,n; i \neq j)$ whenever they exist. Then

$$E\left(S^2\right) = \left[1 - \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i-1} \rho_{ij}\right] \sigma^2 .$$

Corollary 2.10 Let X_j $(j = 1,2,..,n)$'s be identically distributed random variables i.e. $E(X_i) = \mu$, $V(X_i) = \sigma^2$ $(i = 1,2,...n)$ and $Cov(X_i, X_j) = \rho\sigma^2$ $(i = 1,2,...n; j = 1,2,...,n; i \neq j)$ whenever they exist. Then $E(S^2) = (1-\rho)\sigma^2 < 2\sigma^2$.

Example 2.4 Consider the following pdf $f(x_1, x_2, x_3) = \frac{2}{3}(x_1 + x_2 + x_3)$, $0 < x_1, x_2, x_3 < 1$. It can be checked that $E(X_i) = 5/9$, $V(X_i) = 13/162$ $(0 < x_i < 1 ; i = 1,2,3)$, $Cov(X_i, X_j) = -1/324$ and $\rho_{ij} = -1/26$ $(i = 1,2,3; j = 1,2,3; i \neq j)$ (Hardle and Simar, 2003, 128). Then by Corollary 2.10, we have $E(S^2) = (1-\rho)\sigma^2 = (1+1/26)(13/162) = 1/12$ where $S^2 = \sum_{i=1}^{3} (X_i - \bar{X})^2 / 2$ (cf. Example 2.1).

Example 2.5 Suppose that X_1 and X_2 have the joint density function $f(x_1, x_2) = \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2 + x_2^2)} \left(1 + x_1 x_2 e^{-\frac{1}{2}(x_1^2 + x_2^2 - 2)}\right), -\infty < x_1, x_2 < \infty$ (Hogg and Craig, 1978, 121). It can be proved that $X_i \sim N(0,1)$, $(i = 1,2)$, and $E(X_1 X_2) = E(X_1) + eI^2 = e/8 = Cov(X_1, X_2)$ where $I = \int_{-\infty}^{\infty} x^2 e^{-x^2} dx = \sqrt{\pi}/2$. Then, by virtue of Corollary 2.10, $E(S^2) = (1 - e/8)$ where $S^2 = (X_1 - X_2)^2 / 2$.

Corollary 2.11 Let X_j $(j = 1,2,..,n)$'s be uncorrelated but identically distributed random variables i.e. $E(X_i) = \mu$, $V(X_i) = \sigma^2$ $(i = 1,2,...n)$ and $Cov(X_i, X_j) = 0$ $(i = 1,2,...n; j = 1,2,...,n; i \neq j)$ whenever they exist. Then $E(S^2) = \sigma^2$.

Example 2.6 Suppose that X_1, X_2 and X_3 have the joint probability density function

$$f(x_1,x_2,x_3) = \frac{\Gamma((\nu+3)/2)}{(\nu\pi)^{3/2}\Gamma(\nu/2)} \left(1+\tfrac{1}{\nu}\left(x_1^2+x_2^2+x_3^2\right)\right)^{-(\nu+3)/2}, -\infty < x_i < \infty (i=1,2,3)$$

(Anderson, 2003, 55). That is $X_i \sim t_\nu$ $(i=1,2,3)$ and they are pairwise uncorrelated with each pair having a standard bivariate t-distribution with probability density function

$$f(x_i,x_j) = \frac{1}{2\pi} \left(1+\tfrac{1}{\nu}\left(x_i^2+x_j^2\right)\right)^{-\nu/2+1}, -\infty < x_i, x_j < \infty \ (i,j=1,2,3).$$

Since $Cov(X_i,X_j)=0$ $(i=1,2,3; j=1,2,3; i\neq j)$, by virtue of Corollary 2.11, we have $E(S^2)=\nu/(\nu-2),\ \nu>2$ where $S^2=\sum_{i=1}^{3}(X_i-\bar{X})^2/2$.

Example 2.7 Suppose that X_1, X_2 and X_3 have the joint density function

$$f(x_1,x_2,x_3) = \left(\tfrac{1}{2\pi}\right)^{3/2} e^{-(x_1^2+x_2^2+x_3^2)/2}\left(1+x_1x_2x_3\ e^{-(x_1^2+x_2^2+x_3^2)}\right), -\infty < x_i < \infty (i=1,2,3)$$

(Hogg and Craig, 1978, 121). Then it can be proved that $X_i \sim N(0,1)$, $(i=1,2,3)$ and that they are pairwise statistically independent with each pair having a standard bivariate normal distribution. Since $Cov(X_i,X_j)=0$ $(i=1,2,3; j=1,2,3; i\neq j)$, by virtue of Corollary 2.11, we have $E(S^2)=1$ where $S^2=\sum_{i=1}^{3}(X_i-\bar{X})^2/2$.

Corollary 2.12 Let X_j $(j=1,2,..,n)$'s be independently and identically distributed random variables i.e. $E(X_i)=\mu,\ V(X_i)=\sigma^2$ $(i=1,2,...n)$ whenever they exist. Then by Theorem 2.1(b) $E(S^2)=\sigma^2$ which can also be written as $E(S^2)=\frac{1}{2}E(X_1-X_2)^2=\sigma^2$ Theorem 2.1(a).

Example 2.8 Let X_j $(j=1,2,..,n)$'s be independently and identically distributed Bernoulli random variables $B(1,p)$. Then by Corollary 2.12, $E(S^2)=p(1-p)$.

Example 2.9 Let X_j $(j=1,2,..,n)$'s be independently and identically distributed as $N(\mu,\sigma^2)$. Then by Corollary 2.12, $E(S^2)=\sigma^2$ which is well known (Lindgren, 1993).

Similarly, the expected sample variance is the population variance in exponential population with mean $E(X)=\beta$, and gamma population $G(\alpha,\beta)$ with mean $E(X)=\alpha\beta$ and variance $V(X)=\alpha\beta^2$ **.**

Acknowledgements

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

References

Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley-Interscience.

Hogg, R.V. and Craig, A.T. (1978). *Introduction to Mathematical Statistics.* Macmillan Publishing Co.

Joarder, A.H. (2003). Sample Variance and first-order differences of observations. *Math Scientist*, **28**, 129-133.

Joarder, A.H. and Ahmed, S.E. (1998). Estimation of the scale matrix of a class of elliptical distributions. *Metrika* **48**, 149-160.

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

Lindgren, B.W. (1993). *Statistical Theory*. Chapman and Hall.

Hardle, W. and Simar, L. (2003). *Applied Multivariate Statistical Analysis*. Springer.