The Mathematical Expectation of Sample Variance: A General Approach

Anwar H. Joarder and M. Hafidz Omar

Department of Mathematics and Statistics
King Fahd University of Petroleum and Minerals
Dhahran 31261, Saudi Arabia
Emails: anwarj@ kfupm.edu.sa, omarmh@kfupm.edu.sa

Abstract The expected value of sample variance based on independently and identically distributed normal observations is well known, and is often calculated by deriving its sampling distribution. However, the sampling distribution is difficult for other distributions, and more so if the observations are neither independently nor identically distributed. We demonstrate that the expected value in such a general situation depends on the second moment of the difference of pairs of its constituent random variables. We also prove, for this situation, an expression for expected variance that depends on the average of variances of observations, variation among true means and the average of covariances of pairs of observations. Many special cases are expressed as corollaries to illustrate ideas. Some examples that provide insights in mathematical statistics are considered. An application to textile engineering is presented.

Key words: sample variance; expected value of sample variance; population variance; correlation

MSC 2010: 60E99, 62F03, 62P30, 62K99

1. Introduction

The expected value of sample variance is often derived by deriving its sampling distribution which may be intractable in some situations. The objective of this paper is to derive a general formula for the mathematical expectation of sample variance.

One may wonder if there is any real world situation for which we need a generalization of the expected value formula for sample variance. Indeed this kind of situation arises when the observations are not necessarily independent, say, time series data or observations from a mixture distribution with parameters following some other distribution. See for example, Joarder and Ahmed (1998).

It was believed earlier that the rates of return on common stocks were adequately characterized by a normal distribution. But recently, it has been observed by several authors that the empirical distribution of rates of return of common stocks have somewhat thicker tails (larger kurtosis) than that of the normal distribution. The univariate *t*-distribution has got fatter tails and as such it is more appropriate than the normal distribution to characterize stock return rates. For example, if for a given $\Upsilon = \upsilon$, observations follow normal distribution, say, $X_i \sim N(0, \upsilon^2)$, $(i = 1, 2, \cdots, n)$, where $\nu\Upsilon^{-2} \sim \chi^2_\nu$, then the unconditional distribution of the sample follows a *t*-distribution (Example 3.7). Examples 3.3 and 3.7 show the usefulness of

the main result proved in Theorem 2.1 for a sample governed by *t*-distribution. Samples can be drawn from distributions where the components are dependent by 3 methods: the conditional distribution method, transformation method and the rejection method (Johnson, 1987, 43-48).

Blattberg and Gonedes (1974) assessed the suitability of the multivariate *t*-model and Zellner (1976) considered a regression model to study stock return data for a single stock. Interested readers may go through Sutradhar and Ali (1986) who considered a multivariate *t*-model for the price change data for the stocks of four selected firms: General Electric, Standard Oil, IBM and Sears. These are examples where expected sample variance or covariance matrix cannot be derived by appealing to independence.

Let $x_1, x_2, \cdots, x_n (n \geq 2)$ be a sample with variance (s^2) where $(n-1)s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2,\ n \geq 2.$

A matrix W showing the pair-wise 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

$W_l = \{w_{ij}: 1 \leq i \leq n,\ 1 \leq j \leq n;\ i > j\} = \{w_{ij}: 1 \leq i > j \leq n\},$

$W_u = \{w_{ij}: 1 \leq i \leq n,\ 1 \leq j \leq n;\ i < j\} = \{w_{ij}: 1 \leq i < j \leq n\}$ and $W_d = \{w_{ii} = 0: 1 \leq i \leq n\}$

which are the elements in the lower triangle, upper triangle and in the diagonal of the matrix W . Also

$$W_l = \{w_{ij}: 2 \leq i \leq n,\ 1 \leq j \leq i-1\} = \{w_{ij}: 1 \leq j \leq n-1,\ j+1 \leq i \leq n\},$$
$$W_u = \{w_{ij}: 1 \leq i \leq n-1,\ i+1 \leq j \leq n\} = \{w_{ij}: 2 \leq j \leq n,\ 1 \leq i \leq j-1\}. \quad (1.1)$$

Then it is easy to check that $(n-1)s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$ can also be represented by

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

See for example, Joarder (2003) and (2005). The following theorem is due to Joarder (2003).

Theorem 1.1 Let $d_i = x_{i+1} - x_i$, $i = 1, 2, \cdots, n-1$ be the first-order differences of $n(\geq 2)$ observations. Then the variance (s^2) of n observations is given by $n(n-1)s^2 = d'Cd$ where $d = (d_1, d_2, \cdots, d_{n-1})'$ and $C = (c_{ij})$ is an $(n-1)\times(n-1)$ symmetric matrix with $c_{ij} = (n-i)j$ for $i, j = 1, 2, \cdots, n-1$ $(i \geq j)$.

Let the mean square successive difference (MSSD) of sample observations be given by $D = \sum_{i=1}^{n-1} d_i^2$. The ratio of the MSSD to the sample variance $T = D/[(n-1)S^2]$ was suggested by von Neumann, Kent, Bellinson and Hart (1941), Young (1941) and von Neumann (1941 and 1942) as a test statistic to test the independence of the random variables $X_1, X_2, \cdots, X_n (n \geq 2)$ which are successive observations on a stationary Gaussian time series.

In particular, the ratio actually studied by von Neumann was $nT/(n-1)$. Bingham and Nelson (1981) approximated the distribution of the von Neumann's T ratio.

In case of independently and identically distributed random 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 $(\overline{X})$ 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 $(n-1)E(S^2)/\sigma^2 = E(\chi^2_{n-1}) = n-1$, i.e., $E(S^2)=\sigma^2$ (see for example Lindgren, 1993, 213).

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. These situations include expectation of variance of observations that are not necessarily independent as mentioned earlier. Suppose that X_i's $(i=1,2,..,n)$ are uncorrelated random variables from an unknown distribution with finite mean $E(X_i)=\mu$ $(i=1,2,...,n)$ and finite variance $V(X_i)=E(X_i-\mu)^2=\sigma^2$ $(i=1,2,\cdots,n)$, $E(S^2)$ may not be obtained by utilizing the chi-square distribution. A more general approach is thus needed. In this case, it follows that $E(S^2)=\sigma^2$ by virtue of $(n-1)E(S^2)=\sum_{i=1}^{n}E(X_i-\mu)^2 - nE(\overline{X}-\mu)^2$, or,

$(n-1)E(S^2)=n\sigma^2-n(\sigma^2/n).$

In this paper we alternatively demonstrated that the expected value depends on the second moment of the difference of pairs of its constituent random variables. In theorem 2.1, a general formula for expected variance is derived in terms of some natural quantities depending on mean, variance and correlation. Some special cases are presented in Section 3 with examples. An application to textile engineering is presented in Section 4.

2. The Main Result

In what follows we will need the following:

$$n\overline{\mu}=\sum_{i=1}^{n}\mu_i,\ (n-1)\sigma_\mu^2=\sum_{i=1}^{n}(\mu_i-\overline{\mu})^2=\frac{1}{n}\sum_{i=2}^{n}\sum_{j=1}^{i-1}(\mu_i-\mu_j)^2,\ n\overline{\sigma^2}=\sum_{i=1}^{n}\sigma_i^2. \tag{2.1}$$

We define the covariance between X_i and X_j by

$$Cov(X_i,X_j)=\sigma_{ij}=E(X_i-\mu_i)(X_j-\mu_j),\ (i=1,2,...n;\ j=1,2,...,n;i\neq j).$$

Theorem 2.1 Let X_i's $(i=1,2,..,n)$ 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 $\sigma_{ij}=\rho_{ij}\sigma_i\sigma_j$, $(i=1,2,...n;\ j=1,2,...,n;i\neq j)$ and

$$\bar{\sigma}_{\cdot\cdot} = \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i-1} \rho_{ij} \sigma_i \ \sigma_j \ . \tag{2.2}$$

Then

a. $$n(n-1)E(S^2) = \sum_{i=2}^{n} \sum_{j=1}^{i-1} E(X_i - X_j)^2 = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} E(X_i - X_j)^2, \tag{2.3}$$

b. $$E(S^2) = \overline{\sigma^2} + \sigma_\mu^2 - \bar{\sigma}_{\cdot\cdot}, \tag{2.4}$$

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

Proof . Part (a) is obvious by (1.2). Since $x_i - x_j = (x_i - \mu_i) - (x_j - \mu_j) + (\mu_i - \mu_j)$, it can be checked that

$$\begin{aligned} n(n-1)s^2 = \sum_{i=2}^{n} \sum_{j=1}^{i-1} [&(x_i - \mu_i)^2 + (x_j - \mu_j)^2 + (\mu_i - \mu_j)^2 \\ &- 2(x_i - \mu_i)(x_j - \mu_j) + 2(x_i - \mu_i)(\mu_i - \mu_j) - 2(x_j - \mu_j)(\mu_i - \mu_j)]. \end{aligned} \tag{2.5}$$

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),

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

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

$$\sum_{i=2}^{n} \sum_{j=1}^{i-1} E(X_i - \mu_i)(X_j - \mu_j) = \sum_{i=2}^{n} \sum_{j=1}^{i-1} \rho_{ij} \sigma_i \ \sigma_j \ \text{ by (2.2),}$$

it follows from (2.5) that

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

Then the proof for part (b) follows by virtue of

$$\sum_{i=2}^{n} (i-1)\sigma_i^2 + \sum_{j=1}^{n-1} (n-j)\sigma_j^2 = \sum_{i=2}^{n} (i-1)\sigma_i^2 + \sum_{i=1}^{n-1} (n-i)\sigma_i^2 = (n-1)\sum_{i=1}^{n} \sigma_i^2 \ . \qquad \square$$

Alternatively, readers acquainted with matrix algebra may prefer the following proof of Theorem 2.1.

Consider the vector $X : (n \times 1)$ of observations, and $\mu = E(X)$, the vector of means. Note also $\sum = \{\sigma_{ij}\}$, the $(n \times n)$ covariance matrix. Then it is easy to check that $(n-1)S^2 = X'MX$ where the $(n \times n)$ centering matrix $M = I_n - 1_n 1_n'/n$ with I_n the identity matrix of order n and 1_n is a $(n \times 1)$ vector of 1's. Then $(n-1)E(S^2) = E(X'MX)$. But $E(X'MX) = E(tr(X'MX))$ and $E(XX') = \sum + \mu\mu'$ so that

$$(n-1)E(S^2) = tr\left(M \sum\right) + \mu' M \mu.$$

That is, $E(S^2) = \overline{\sigma^2} - \bar{\sigma}_{..} + \sigma_\mu^2,\ (i \neq j)$, since $tr(M\sum) = (n-1)\overline{\sigma^2} - (n-1)\bar{\sigma}_{..},\ (i \neq j)$ and $\mu' M \mu = \sum_{i=1}^{n} (\mu_i - \bar{\mu})^2 = (n-1)\sigma_\mu^2$, where $\bar{\sigma}_{..} = \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i-1} \sigma_{ij}$ is the mean of the off-diagonal elements of $\sum$. □

3. Some Deductions and Mathematical Application of the Result

In this section, we deduce a number of corollaries from Theorem 2.1. But first we have two examples to illustrate part (a) of Theorem 2.1.

Example 3.1 Suppose that $f(x_i) = \frac{2}{3}(x_i + 1),\ 0 < x_i < 1\ \ (i = 1,2,3)$ and the *dependent* sample (X_1, X_2, X_3) is governed by the probability density function $f(x_1, x_2, x_3) = \frac{2}{3}(x_1 + x_2 + x_3)$, (cf. Hardle and Simar, 2003, 128). Then it can be checked that $E(X_i) = 5/9$, $E(X_i^2) = 7/18,\ E(X_i X_j) = 11/36$, $V(X_i) = 13/162$ and $Cov(X_i, X_j) = -1/324$ $(i = 1,2,3; j = 1,2,3; i \neq j)$. Let $S^2 = \sum_{i=1}^{3} (X_i - \bar{X})^2/2$ be the sample variance. Then by Theorem 2.1(a),

$$6E(S^2) = E(X_1 - X_2)^2 + E(X_1 - X_3)^2 + E(X_2 - X_3)^2.$$

Since $E(X_i - X_j)^2 = (7/18) + (7/18) - 2(11/36),\ (i = 1,2,3;\ j = 1,2,3; i \neq j)$, we have $E(S^2) = 1/12$.

Example 3.2 Let X_i's $(i = 1,2,..,n)$ be *independently, identically* and normally distributed as $N(\mu, \sigma^2)$. Since $X_i - X_j \sim N(0,\ 2\sigma^2),\ \ i \neq j$, by Theorem 2.1(a), we have

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

That is $E(S^2) = \sigma^2$.

The following corollary is a special case of Theorem 2.1(b) if $\rho_{ij} = 0,\ (i = 1, 2, \cdots, n;$ $j = 1, 2, \cdots, n)$.

Corollary 3.1 Let X_i's $(i = 1, 2, .., n)$ 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 $E(S^2) = \overline{\sigma^2} + \sigma_\mu^2$.

Note that if we form a matrix with the correlation coefficients $\rho_{ij}\ (i = 1, 2, \cdots, n; j = 1, 2, n)$, then by symmetry of the correlation coefficients, total number of the elements in the lower triangle (say ρ^*) would be the same as that of the upper triangle i.e.

$$\rho^* = \sum_{i=2}^{n}\sum_{j=1}^{i-1}\rho_{ij} = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\rho_{ij}\ .$$

Hence $\rho^* + n + \rho^* = n^2\bar{\rho}_{..}$ where

$$n^2\bar{\rho}_{..} = \sum_{i=1}^{n}\sum_{j=1}^{n}\rho_{ij}, \tag{3.1}$$

so that $\rho^* = n(n\bar{\rho}_{..} - 1)/2$.

If $V(X_i) = \sigma^2,\ \ (i = 1, 2, ...n)$ in Theorem 2.1 (b), then $\overline{\sigma^2} = \sigma^2$ and $\bar{\sigma}_{..} = (n\bar{\rho}_{..} - 1)\sigma^2/(n-1)$ where $\bar{\rho}_{..}$ is defined by (3.1) and we have the following corollary.

Corollary 3.2 Let X_i's $(i = 1, 2, .., n)$ 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 $E(S^2) = \left(1 - \dfrac{n\bar{\rho}_{..} - 1}{n-1}\right)\sigma^2 + \sigma_\mu^2$.

If $V(X_i) = \sigma^2$, $\rho_{ij} = \rho\ \ (i = 1, 2, ...n; j = 1, 2, ..., n; i \neq j)$, in Theorem 2.1 (b), then $\overline{\sigma^2} = \sigma^2$ and $\bar{\sigma}_{..} = \rho\sigma^2$, we have the following corollary.

Corollary 3.3 Let X_i's $(i = 1, 2, .., n)$ 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 $E(S^2) = (1-\rho)\sigma^2 + \sigma_\mu^2 \leq 2\sigma^2 + \sigma_\mu^2$.

If $V(X_i) = \sigma^2$, $\rho_{ij} = 0\ \ (i = 1, 2, ...n; j = 1, 2, ..., n; i \neq j)$, in Theorem 2.1 (b), then $\overline{\sigma^2} = \sigma^2$ and $\bar{\sigma}_{..} = 0$, then we have the following corollary.

Corollary 3.4 Let X_i $(i=1,2,..,n)$'s be random variables with $E(X_i)=\mu_i$, $V(X_i)=\sigma^2$, $Cov(X_i,X_j)=0$, $(i=1,2,...n;\ j=1,2,...,n;\ i\neq j)$ whenever they exist. Then $E(S^2)=\sigma^2+\sigma_\mu^2$.

An example is provided below to illustrate the situation.

Example 3.3 Let $f(x_i)=\dfrac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\sigma\Gamma(\nu/2)}\left(1+\dfrac{1}{\nu\sigma^2}(x_i-\mu_i)^2\right)^{-(\nu+1)/2}$, $\nu>2$, $(i=1,2,3)$ and the dependent observations (X_1,X_2,X_3) be governed by the probability density function

$$f(x_1,x_2,x_3)=\frac{\Gamma((\nu+3)/2)}{(\nu\pi)^{3/2}\sigma^3\Gamma(\nu/2)}\left(1+\frac{1}{\nu\sigma^2}(x_1-\mu_1)^2+(x_2-\mu_2)^2+(x_3-\mu_3)^2)\right)^{-(\nu+3)/2}, \quad (3.2)$$

$-\infty<x_i<\infty$ $(i=1,2,3)$, $\sigma>0$, $\nu>2$ (cf. Anderson, 2003, 55). Since $V(X_i)=\dfrac{\nu\sigma^2}{\nu-2}$, $\nu>2$, $(i=1,2,3)$, and $Cov(X_i,X_j)=0$, $(i=1,2,3;\ j=1,2,3;\ i\neq j)$, it follows from Corollary 3.4 that $E(S^2)=\dfrac{\nu\sigma^2}{\nu-2}+\sigma_\mu^2$, $\nu>2$ where S^2 is the sample variance and $2\sigma_\mu^2=(\mu_1-\bar{\mu})^2+(\mu_2-\bar{\mu})^2+(\mu_3-\bar{\mu})^2$, $3\bar{\mu}=\mu_1+\mu_2+\mu_3$.

Corollary 3.5 Let X_i's $(i=1,2,..,n)$ be random variables with $E(X_i)=\mu$, $V(X_i)=\sigma_i^2$, $\sigma_{ij}=\rho_{ij}\sigma_i\sigma_j$, $(i=1,2,...n;\ j=1,2,...,n;\ i\neq j)$ whenever they exist. Then $E(S^2)=\overline{\sigma^2}-\bar{\sigma}_{..}$ where $\overline{\sigma^2}$ is defined by (2.1) and $\bar{\sigma}_{..}$ is defined in (2.2).

Corollary 3.6 Let X_i's $(i=1,2,..,n)$ be random variables with $E(X_i)=\mu$, $V(X_i)=\sigma_i^2$, $Cov(X_i,X_j)=\rho\sigma_i^2$, $(i=1,2,...n;\ j=1,2,...,n;\ i\neq j)$ whenever they exist. Then $E(S^2)=\overline{\sigma^2}-\binom{n}{2}^{-1}\rho\sum_{i=2}^{n}(i-1)\sigma_i^2$.

Corollary 3.7 Let X_i's $(i=1,2,..,n)$ 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 $E(S^2)=\overline{\sigma^2}$.

Corollary 3.8 Let X_i's $(i=1,2,..,n)$ be *independently* distributed random variables 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 $E(S^2)=\overline{\sigma^2}$.

Corollary 3.9 Let X_i's $(i=1,2,..,n)$ 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

$E(S^2) = \left(1 - \frac{n\bar{\rho}_{..} - 1}{n-1}\right)\sigma^2$ where $\bar{\rho}_{..}$ is defined by (3.1).

Corollary 3.10 Let X_i's $(i = 1,2,..,n)$ 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 $E(S^2) = (1-\rho)\sigma^2$ (cf. Rohatgi and Saleh).

Two examples are given below to illustrate the above corollary.

Example 3.4 In Example 3.1, $\sigma^2 = 13/162$ and $\rho = -1/26$. Then by Corollary 3.10, we have $E(S^2) = (1-\rho)\sigma^2 = 1/12$ where $S^2 = \sum_{i=1}^{3}(X_i - \bar{X})^2/2$.

Example 3.5 Let $X_i \sim N(0,1),\ (i = 1,2)$ and the sample (X_1, X_2) (not necessarily independent) have the joint density function

$$f(x_1, x_2) = \frac{1}{2\pi}\exp\left[-\tfrac{1}{2}(x_1^2 + x_2^2)\right]\left[1 + x_1 x_2 \exp\left(-\tfrac{1}{2}(x_1^2 + x_2^2 - 2)\right)\right],\ -\infty < x_1, x_2 < \infty, \qquad (3.3)$$

(cf. Hogg and Craig, 1978, 121). Writing out the joint density function (3.3) into two parts, we can easily prove that

$E(X_1X_2) = E(X_1)E(X_2) + \frac{e}{2\pi}\ I(x_1)I(x_2)$ where $I(x) = \int_{-\infty}^{\infty} x^2 e^{-x^2} dx = \sqrt{\pi}/2$. But $E(X_i) = 0\ (i = 1,2)$, $V(X_i) = 1\ (i = 1,2)$, hence $Cov(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2)$ which simplifies to $e/8$ is also the correlation coefficient ρ between X_1 and X_2. Hence by virtue of Corollary 3.10, we have $E(S^2) = (1 - e/8)$.

Part (b) of Theorem 2.1 is specialized below for *uncorrelated but identically* distributed random variables.

Corollary 3.11 Let X_i's $(i = 1,2,..,n)$ 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 $E(S^2) = \sigma^2$.

Two examples are given below to illustrate Corollary 3.11.

Example 3.6 Let $X_i \sim N(0,1),\ (i = 1,2,3)$ and the sample (X_1, X_2, X_3) be governed by the joint density function

$$f(x_1, x_2, x_3) = \left(\frac{1}{2\pi}\right)^{3/2}\exp\left[-\frac{1}{2}(x_1^2 + x_2^2 + x_3^2)\right]\left[1 + x_1x_2x_3\ \exp\left(-\frac{1}{2}(x_1^2 + x_2^2 + x_3^2)\right)\right], \qquad (3.4)$$

$-\infty < x_i < \infty\ (i = 1,2,3)$ (cf. Hogg and Craig, 1978, 121). Then it can be proved that the sample observations are pair-wise statistically independent with each pair having a standard

bivariate normal distribution. We thus have $E(X_i)=0$, $V(X_i)=1$ and $Cov(X_i,X_j)=0$ $(i=1,2,3; j=1,2,3; i\neq j)$. By virtue of Corollary 3.11, we have $E(S^2)=1$ where $S^2=\sum_{i=1}^{3}(X_i-\bar{X})^2/2$.

Example 3.7 Let $X_i\,(i=1,2,3)$ have a univariate *t*-distribution with density function

$$f(x_i)=\frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}\left(1+\frac{1}{\nu}x_i^2\right)^{-(\nu+1)/2},\ \nu>2,\ (i=1,2,3)$$

and the sample (X_1,X_2,X_3) be governed by the joint density function of a trivariate *t*-distribution given by

$$f(x_1,x_2,x_3)=\frac{\Gamma((\nu+3)/2)}{(\nu\pi)^{3/2}\Gamma(\nu/2)}\left(1+\frac{1}{\nu}(x_1^2+x_2^2+x_3^2)\right)^{-(\nu+3)/2}, \tag{3.5}$$

$-\infty<x_i<\infty$ $(i=1,2,3)$ (Anderson, 2003, 55).

Obviously X_1,X_2 and X_3 are independent if and only if $\nu\to\infty$. It can be proved that $(X_i\mid \Upsilon=\upsilon)\sim N(0,\upsilon^2)$, $(i=1,2,3)$, where $\nu/\Upsilon^2\sim\chi_\nu^2$ It can be proved that $X_i\,(i=1,2,3)$'s are pair-wise *uncorrelated* with each pair having a standard bivariate *t*-distribution with probability density function

$$f(x_i,x_j)=\frac{1}{2\pi}\left(1+\tfrac{1}{\nu}(x_i^2+x_j^2)\right)^{-(\nu+2)/2}, \tag{3.6}$$

$-\infty<x_i,x_j<\infty$ $(i,j=1,2,3; i\neq j)$. Since $E(X_i)=0$, $V(X_i)=\nu/(\nu-2)$, $\nu>2$, $(i=1,2,3)$, $Cov(X_i,X_j)=0$ $(i=1,2,3; j=1,2,3; i\neq j)$. Then by virtue of Corollary 3.11, we have $E(S^2)=\nu/(\nu-2)$, $\nu>2$ where $S^2=\sum_{i=1}^{3}(X_i-\bar{X})^2/2$ is the sample variance. A realistic example based on stock returns is considered in Sutradhar and Ali (1986).

Corollary 3.12 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 $E(S^2)=\frac{1}{2}E(X_1-X_2)^2=\sigma^2$ by Theorem 2.1(a).

Example 3.8 Let $X_j\,(j=1,2,..,n)$'s be *independently and identically* distributed Bernoulli random variables $B(1,p)$. Then by Corollary 3.12, we have $E(S^2)=p(1-p)$.

Example 3.9 Let $X_j\ (j=1,2,..,n)$'s be *independently and identically* distributed as $N(\mu,\sigma^2)$. Then by Corollary 3.12, we have $E(S^2)=\sigma^2$ which is well known (Lindgren, 1993).

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

4. An Application in Textile Engineering

A textile company weaves fabric on a large number of looms. They would like the looms to be homogenous so that they obtain fabric of uniform strength. The process engineer suspects that, in addition to the usual variation in strength for samples of fabric from the same loom, there may be significant variations in mean strengths between looms. To investigate this, the engineer selects four looms at random and makes four strength determinations on the fabric manufactured on each loom. This experiment is done in random order, and the data obtained are shown below.

Looms	Observations				Total
1	98	97	99	96	390
2	91	90	93	92	366
3	96	95	97	95	383
4	95	96	99	98	388

Montgomery (2001, 514).

Consider a random effects linear model given by

$$y_{ij}=\mu+\tau_i+\varepsilon_{ij},\ \ i=1,2,\cdots,a;\ \ j=1,2,\cdots,n_i$$

where μ is some constant, τ_i has mean 0 and variance σ_τ^2, the errors ε_{ij} have mean 0 and variance σ^2. Also assume that τ_i and ε_{ij} are uncorrelated. Then $\bar{Y}_{i.}=\frac{1}{n_i}\sum_{j=1}^{n_i}Y_{ij}$, has mean $E(\bar{Y}_{i.})=\mu$, and variance

$$V(\bar{Y}_{i.})=\sigma_\tau^2+\frac{1}{n_i}\sigma^2. \qquad (4.1)$$

If we write $S^2_{\bar{Y}_{i.}}=\frac{1}{a-1}\sum_{i=1}^{a}(\bar{Y}_{i.}-\bar{Y}_{..})^2$, then by Corollary 3.8, $E(S^2_{\bar{Y}_{i.}})=\frac{1}{a}\sum_{i=1}^{a}V(\bar{Y}_{i.})$, which, by virtue of (4.1), can be written as

$$E(S^2_{\bar{Y}_{i.}})=\frac{1}{a}\sum_{i=1}^{a}\left(\sigma_\tau^2+\frac{1}{n_i}\sigma^2\right),$$

so that

$$E(S^2_{\overline{Y}_{i.}}) = \sigma^2_\alpha + \frac{\sigma^2}{a}\sum_{i=1}^{a}\frac{1}{n_i}. \tag{4.2}$$

The Sum of Squares due to Treatment (looms here) is $SST = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(\overline{y}_{i.} - \overline{y}_{..})^2$ which can be written as $SST = (a-1)\sum_{j=1}^{n_i} S^2_{\overline{Y}_{i.}}$ so that by (4.2), we have

$$E(SST) = (a-1)\sum_{j=1}^{n_i}\left(\sigma^2_\tau + \frac{\sigma^2}{a}\sum_{i=1}^{a}\frac{1}{n_i}\right) \tag{4.3}$$

which, in the balanced case, simplifies to $E(SST) = (a-1)(n\sigma^2_\tau + \sigma^2).$ Then the expected mean Sum of Squares due to Treatment is given by

$$E(MST) = n\sigma^2_\tau + \sigma^2. \tag{4.4}$$

The Sum of Squares due to Errors is given by $SSE = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i.})^2$ which can be written as $SSE = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(\varepsilon_{ij} - \overline{\varepsilon}_{i.})^2$. Since $E(\varepsilon_{ij}) = 0$ and $V(\varepsilon_{ij}) = \sigma^2,$ by Corollary 3.12, we have $E\left(\frac{1}{n_i - 1}\sum_{j=1}^{n_i}(\varepsilon_{ij} - \overline{\varepsilon}_{i.})^2\right) = \sigma^2,$ so that $E(SSE) = \sum_{i=1}^{a}(n_i - 1)\sigma^2$ which in the balanced case, simplifies to $E(SSE) = a(n-1)\sigma^2$ so that the mean Sum of Squares due to Error is given by

$$E(MSE) = \sigma^2. \tag{4.5}$$

By virtue of (4.4) and (4.5), a test of $H_0 : \sigma^2_\tau = 0$ against $H_1 : \sigma^2_\tau > 0$ can be based on $F = \frac{MST}{MSE}$ where the variance ratio has usual F distribution with $(a-1)$ and $a(n-1)$ degrees of freedom if the errors are normally distributed.

It can be easily checked that the Centered Sum of Squares $(CSS = a(n-1) \times MST)$ and Sum of Squares due to treatments $(SST = (a-1)MST)$ for our experiment above are given by

$$CSS = (98)^2 + (97)^2 + \cdots + (98)^2 - \left((1527)^2 / 4(4)\right) \approx 111.94,$$

$$SST = \frac{1}{4}\left((390)^2 + (366)^2 + \cdots + (388)^2\right) - \left((1527)^2 / 4(4)\right) = 89.19,$$

respectively, and $F = \frac{MST}{MSE} = \frac{89.19/3}{22.75/12} = 15.68$

which is much larger than $f_{0.05} = 3.52$. The p-value is smaller than 0.001. This suggests that there is a significant variation in the strength of the fabrics between the looms.

5. Conclusion

The general method for the expectation of sample variance that has been developed here is important if observations have *non-identical* distributions be it in means, variances or covariances. While part (a) of Theorem 2.1 states that expected variance depends on that of the squared difference of pairs of observations, part (b) of the theorem states that expected variance depends on the average of variances of observations, variation among true means and the average of covariances of pairs of observations. The theorem has potential to be useful in time series analysis, design of experiments and psychometrics where the observations are not necessarily independently and identically distributed. Because no distributional form is assumed to obtain the main results, the theorem can also be applied even without requiring strict adherence to the normality assumption. The results (4.4) and (4.5) are usually derived in Design and Analysis or other statistics courses by distribution theory based on strong distributional assumptions, mostly normality. It is worth mentioning that the assumption of normality of the errors can be relaxed to broader class of distributions, say, elliptical distributions. See for example Joarder (2013).

Acknowledgements

The authors are grateful to their colleague Professor A. Laradji and anonymous referees for providing constructive suggestions that have improved the motivation and content of the paper. The authors gratefully acknowledge the excellent research facilities available at King Fahd University of Petroleum & Minerals, Saudi Arabia.

References

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

Bingham, C. and Nelson, L. (1981). An approximation for the distribution of the von Neumann ratio. *Technometrics*, **23**(3), 285-288.

Blattberg, R.C. and Gonedes, N.J. (1974). A comparison of the stable and Student distributions as statistical models for stock prices. *Journal of Business*, **47**, 224-280.

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

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

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

Joarder, A.H. (2005). The Expected Sample Variance in a General Situation. *Technical Report No. 308,* Department of Mathematics and Statistics, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia.

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

Joarder, A.H. (2013). Robustness of correlation coefficient and variance ratio under elliptical symmetry. To appear in *Bulletin of Malaysian Mathematical Sciences Society.*

Johnson, M.E. (1987). *Multivariate Statistical Simulation*. John Wiley and Sons.

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

Montgomery, D.C. (2001, 514). *Design and Analysis of Experiments*. John Wiley and Sons.

Rohatgi, V.K. and Saleh, A.K.M.E. (2001). *An Introduction to Probability and Statistics*. John Wiley.

Sutradhar, B.C. and Ali, M.M. (1986). Estimation of the parameters of a regression model with a multivariate t error variable. *Communications in Statistics – Theory and Methods,* **15**, 429- 450.

Young, L.C. (1941). On randomness in ordered sequences. *Annals of Mathematical Statistics,* **12,** 293-300.

Von Neumann, J; Kent, R.H.; Bellinson, H.R. and Hart, B.I. (1941). The mean square successive difference. *Annals of Mathematical Statistics*, 12, 153-162.

Von Neumann, J. (1941). Distribution of the ratio of the mean square successive difference to the variance. *Annals of Mathematical Statistics*, **12**, 367-395.

Von Neumann, J. (1942). A further remark concerning the distribution of the ratio of the mean square successive difference to the variance. *Annals of Mathematical Statistics*, **13**, 86-88.

Zellner, A. (1976). Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error term. *Journal of American Statistical Association*, 71, 400-405 (Correction, 71, 1000).