Spreadsheets in Education (eJSiE)

Volume 3 | Issue 2 Article 2

7-13-2009

Estimation Error in the Correlation of Two Random Variables: A Spreadsheet-Based Exposition

Clarence C. Y. Kwan
McMaster University, kwanc@mcmaster.ca

Estimation Error in the Correlation of Two Random Variables: A Spreadsheet-Based Exposition

Abstract
Although the statistical term *correlation* is well-known across many academic disciplines, estimation error in the correlation has traditionally been considered to be a topic too difficult for students outside statistical fields. This pedagogic study presents an approach for the estimation that does not require any advanced statistical concepts. By using familiar spreadsheet functions to facilitate the required computations, it intends to make the analytical material involved accessible to more students.

Keywords
correlation, estimation error, sampling variance

Estimation Error in the Correlation of Two Random Variables: A Spreadsheet-Based Exposition

1 Introduction

The statistical term *correlation* is well-known to students in many academic disciplines. As a dimensionless quantity, with potential values ranging from minus one to plus one, it captures how two random variables relate statistically to each other. Analytically, it is defined as the covariance of the two variables divided by the product of their standard deviations. Reliable information about the correlation allows useful implications to be drawn. In the context of investment, for example, if the random rate of return of an asset is known to be positively and highly correlated with that of another asset, a practical implication is that investing in both assets does not offer a much better risk-return trade-off than investing in only one of these assets.

When the correlation of two variables is estimated from a sample of observations, the observations can be viewed as random draws from a joint distribution of the two variables. To estimate the correlation from such a sample is straightforward. Currently available electronic spreadsheets all have functions for this task; in the case of Microsoft Excel, for example, the corresponding function is CORREL. However, to assess the accuracy of the correlation that the sample provides is not as simple. This is because the correlation, when expressed in terms of the variances and the covariance of the two variables, is in an analytically inconvenient form.

If the observations for estimating correlations are based on some experimental results, for example, an implicit requirement is that the experimental conditions for generating the observations be the same. Likewise, if the observations are deduced from empirical data, such as historical observations of some economic variables, it is implicitly assumed that the economic conditions for generating the observations be the same.[1] Thus, although estimation error in the correlation tends to decrease as the sample size increases, the reliance on large samples to bypass the issue of estimation error is not always a viable option.

[1] The assumption of stationary distributions of the underlying economic variables can be relaxed to allow time-varying variances, covariances, and correlations to be estimated. However, this pedagogic study is confined to simple cases where the stationarity assumption is deemed acceptable.

There are ways to bypass the analytical inconvenience as noted above. For example, it is well-known in statistics that, if the sample correlation r is based on N observations of a bivariate normal distribution, $r[(N-2)/(1-r^2)]^{1/2}$ follows approximately a t-distribution with $(N-2)$ degrees of freedom. This statistical feature allows the significance of the sample correlation to be tested. It is also well-known in statistics that the distribution of $z = \frac{1}{2}\ln[(1+r)/(1-r)]$ — Fisher's z-transformation of the sample correlation — is approximately normal with the standard error of z being $(N-3)^{-1/2}$. Here, the independence of the standard error of z from r requires the assumption of a bivariate normal distribution of the underlying random variables. Fisher's z-transformation allows the confidence intervals of the correlation to be established.[2] However, although these recipes are easy to follow, their derivations do require statistical concepts unfamiliar to most students outside statistical fields.

Then, from a pedagogic perspective, if the issue of estimation error is to be addressed, a challenging question for instructors to consider is whether the topic can still be taught to students outside statistical fields. This pedagogic study is a response to such a challenge; it extends the statistical approach in Schäfer and Strimmer (2005), originally for estimating the errors in the individual variances and covariances from finite samples, to estimating the error in the correlation of two random variables. Specifically, we treat each point estimate as a realization of the random variable in question. For example, if the random variable in question is a sample covariance, we express its sampling variance — an estimated variance of its sampling distribution — in terms of the observations of the two underlying random variables. A nice feature of the approach here is that there are no specific distributional requirements on the two underlying variables. More importantly, from a pedagogic perspective, the statistical concepts involved can be understood by students outside statistical fields. However, without specifying a bivariate distribution, the approach here does not directly facilitate any significance tests of the sample correlation; nor does it facilitate the establishment of any confidence intervals.

To bypass the analytical difficulty in expressing exactly the sampling variance of the correlation in terms of the observations of the two underlying random variables, we rely on a linear approximation that a first-order Taylor expansion provides. Specifically, we express it approximately as a linear combination of the sampling variances and covariances of various random

[2] See, for example, Warner (2007, chapter 7) for descriptions of the above t-test and Fisher's z-transformation.

variables, each of which can be estimated from observations of the two underlying random variables. This study, which seeks to assess the precision of the correlation that the sample provides, is a pedagogic version of some analytical material in Kwan (2008), Scheinberg (1966), and Stuart and Ord (1987, chapter 10).

In the approach here, we simplify notation (from that in the above references), if needed, not only for ease of exposition, but also for orderly arranging the corresponding data into arrays in a spreadsheet for computational convenience. Besides using familiar Excel functions such as VAR and COVAR for computing various variances and covariances during the intermediate steps, we also use matrix functions in Excel such as TRANSPOSE and MMULT for transposing and multiplying arrays of some data to compute the final result. For students who are unfamiliar with matrix algebra, we provide an equivalent, but more intuitive, computational approach as well. In so doing, we intend to make the computations involved accessible to more students in different academic disciplines.

The paper is organized as follows: Drawing on Schäfer and Strimmer (2005), section 2 describes how sampling variances and covariances of their point estimates (of variances and covariances) can be expressed in terms of the observations of the two underlying random variables. Section 3 extends the approach in section 2 to estimating the error in the sample correlation by expressing its sampling variance as approximately a linear combination of various sampling variances and covariances. Section 4 provides a spreadsheet-based illustration of the corresponding computations using Excel. Some concluding remarks are provided in section 5.

2 Estimation Errors in Variances and Covariances

Consider two random variables, x_1 and x_2, with N pairs of observations. Each pair is labeled as x_{1n} and x_{2n}, for $n = 1, 2, \ldots, N$. The two sample means are

$$\overline{x}_i = \frac{1}{N} \sum\nolimits_{n=1}^{N} x_{in}, \text{ for } i = 1 \text{ and } 2, \tag{1}$$

Each (i, j)-element of the 2×2 sample covariance matrix is

$$s_{ij} = \frac{1}{N-1} \sum\nolimits_{n=1}^{N} (x_{in} - \overline{x}_i)(x_{jn} - \overline{x}_j). \tag{2}$$

Here, s_{11} and s_{22} are the sample variances of the two variables, and $s_{12} = s_{21}$ is their sample covariance.

As in Schäfer and Strimmer (2005), we treat s_{11}, s_{12}, s_{21}, and s_{22} as random variables. Specifically, we introduce a random variable w_{ij} for each of i and j, which can be 1 or 2, and let

$$w_{ijn} = (x_{in} - \overline{x}_i)(x_{jn} - \overline{x}_j), \quad \text{for } n = 1, 2 \ldots, N, \tag{3}$$

be its N observations. This variable is the product of the underlying variables x_i and x_j with their sample means removed first. Noting that the sample mean of w_{ij} is

$$\overline{w}_{ij} = \frac{1}{N} \sum\nolimits_{n=1}^{N} w_{ijn}, \tag{4}$$

we can express the sample covariance of x_i and x_j, including cases where $i = j$ and $i \neq j$, equivalently as

$$s_{ij} = \frac{N}{N-1} \overline{w}_{ij}. \tag{5}$$

Given equation (5), the sampling variance of s_{ij} is

$$\widehat{Var}(s_{ij}) = \frac{N^2}{(N-1)^2} \widehat{Var}(\overline{w}_{ij}). \tag{6}$$

It is well-known in statistics that the distribution of the sample mean of a random variable based on N observations has a sampling variance that is only $1/N$ of the sampling variance of the variable.[3] Thus, it follows from equation (6) that

$$\widehat{Var}(s_{ij}) = \frac{N}{(N-1)^2} \widehat{Var}(w_{ij}). \tag{7}$$

With $\overline{w}_{kl}$ being the sample mean of the variable w_{kl}, we also obtain from equation (5) that

$$\widehat{Cov}\,(s_{ij}, s_{kl}) = \frac{N^2}{(N-1)^2} \widehat{Cov}(\overline{w}_{ij}, \overline{w}_{kl}), \tag{8}$$

[3] To explain this statistical concept to students, we take N random draws from the distribution of a random variable u, for example. With the draws being $u_1, u_2, \ldots, u_N$, the sample mean is $\overline{u} = (u_1 + u_2 + \ldots + u_N)/N$. The sampling variance of $\overline{u}$ is $\widehat{Var}(\overline{u}) = (1/N^2)\widehat{Var}(u_1 + u_2 + \ldots + u_N)$. The latter variance term can be written equivalently as a covariance involving two identical sums of terms, with u_p and u_q being their representative terms. Here, each of p and q can be any of $1, 2, \ldots, N$. Thus, when expressed explicitly, the latter variance term consists of the sum of $N \times N$ terms of the form $\widehat{Cov}(u_p, u_q)$, with each being the sampling covariance of u_p and u_q. If $p \neq q$, $\widehat{Cov}(u_p, u_q)$ is zero, as the two draws are independent of each other. Each case of $p = q$ pertains to the same draw, and thus the corresponding term $\widehat{Cov}(u_p, u_q)$ is the same as $\widehat{Cov}(u, u)$ or, equivalently, $\widehat{Var}(u)$. As there are N cases of $p = q$, it follows that $\widehat{Var}(\overline{u}) = (1/N)\widehat{Var}(u)$.

where each of i, j, k, and l can be 1 or 2. Likewise, in a bivariate setting, the joint distribution of the sample means of two random variables based on N pairs of observations has a sampling covariance that is $1/N$ of the sampling covariance of the two variables.[4] Thus, equation (8) also leads to

$$\widehat{Cov}\left(s_{ij}, s_{kl}\right) = \frac{N}{(N-1)^2}\widehat{Cov}\left(w_{ij}, w_{kl}\right). \tag{9}$$

Given the individual observations of the random variables w_{ij} and w_{kl} as equation (3) provides, to compute $\widehat{Var}(w_{ij})$ and $\widehat{Cov}\left(w_{ij}, w_{kl}\right)$ using Excel is straightforward. Given also equations (7) and (9), so are the computations of the sampling variances and covariances of s_{ij} and s_{kl} for various cases of i, j, k, and l.

Notice that, $N/(1-N)^2$ varies asymptotically as $1/N$. Notice also that, for a given joint distribution of the two underlying variables x_1 and x_2, as N approaches infinity, $\widehat{Var}(w_{ij})$ and $\widehat{Cov}\left(w_{ij}, w_{kl}\right)$ still remain finite. Thus, according to equations (7) and (9), an increase in the number of observations will tend to result in lower magnitudes of $\widehat{Var}(s_{ij})$ and $\widehat{Cov}\left(s_{ij}, s_{kl}\right)$. As it will soon be clear, some of these sampling variances and covariances are required for computing the sampling variance of the correlation. Further, for the same joint distribution of the two underlying variables, an increase in the number of observations will tend to result in a lower sampling variance of their correlation.

3 Estimation Error in the Correlation of Two Random Variables

The sample correlation of the random variables x_1 and x_2 is

$$r = \frac{s_{12}}{\sqrt{s_{11}s_{22}}}. \tag{10}$$

[4] To explain this statistical concept to students, we consider a joint-distribution of two random variables, u and v, for example. We take N random draws from the distribution, which are $(u_1, v_1), (u_2, v_2), \ldots, (u_N, v_N)$. With the sample means of the two variables being $\overline{u} = (u_1 + u_2 + \ldots + u_N)/N$ and $\overline{v} = (v_1 + v_2 + \ldots + v_N)/N$, their sampling covariance is $\widehat{Cov}(\overline{u}, \overline{v}) = (1/N^2)\widehat{Cov}(u_1 + u_2 + \ldots + u_N, v_1 + v_2 + \ldots + v_N)$. When expressed explicitly, the latter covariance term is the sum of $N \times N$ terms of the form $\widehat{Cov}(u_p, v_q)$, each being the sampling covariance of u_p and v_q. Among them, all cases of $p \neq q$ will vanish, as the two draws are independent of each other. For each of the N remaining cases where $p = q$, which pertains to the same draw, the corresponding term $\widehat{Cov}(u_p, v_q)$ is the same as $\widehat{Cov}(u, v)$. Thus, we have $\widehat{Cov}(\overline{u}, \overline{v}) = (1/N)\widehat{Cov}(u, v)$.

Although this expression is in an analytically inconvenient form, we can still approximate it linearly, by treating s_{12}, s_{11}, and s_{22} as the underlying variables. The approach involved, commonly called the delta method, requires a first-order Taylor expansion of $s_{12}/\sqrt{s_{11}s_{22}}$ around $s_{12} = s^*_{12}$, $s_{11} = s^*_{11}$, and $s_{22} = s^*_{22}$, the corresponding point estimates that the sample provides. The truncated Taylor series is

$$\begin{aligned}\frac{s_{12}}{\sqrt{s_{11}s_{22}}} &= \frac{s^*_{12}}{\sqrt{s^*_{11}s^*_{22}}} - \frac{s^*_{12}}{2s^*_{11}\sqrt{s^*_{11}s^*_{22}}}(s_{11} - s^*_{11}) \\ &\quad - \frac{s^*_{12}}{2s^*_{22}\sqrt{s^*_{11}s^*_{22}}}(s_{22} - s^*_{22}) + \frac{1}{\sqrt{s^*_{11}s^*_{22}}}(s_{12} - s^*_{12}). \qquad (11)\end{aligned}$$

We will show below that this approximate expression allows the sampling variance of r to be estimated.

For students who are unfamiliar with multivariate differential calculus, here is a simple way to reach the above first-order Taylor expansion: For notational convenience, let $y_1 = s_{11}$, $y_2 = s_{22}$, $y_3 = s_{12}$, $y^*_1 = s^*_{11}$, $y^*_2 = s^*_{22}$, and $y^*_3 = s^*_{12}$. Let also $\triangle y_1 = y_1 - y^*_1$, $\triangle y_2 = y_2 - y^*_2$, and $\triangle y_3 = y_3 - y^*_3$ be the deviations from the corresponding point estimates. With $r = y_3/\sqrt{y_1y_2}$ and $r^* = y^*_3/\sqrt{y^*_1y^*_2}$, their difference is $\triangle r = r - r^*$. The idea of a first-order Taylor expansion here is to approximate $\triangle r$ as a linear function of $\triangle y_1$, $\triangle y_2$, and $\triangle y_3$. To this end, we write

$$\begin{aligned}\triangle r &= \frac{y^*_3 + \triangle y_3}{\sqrt{(y^*_1 + \triangle y_1)(y^*_2 + \triangle y_2)}} - \frac{y^*_3}{\sqrt{y^*_1y^*_2}} \\ &= \frac{(y^*_3 + \triangle y_3)\sqrt{y^*_1y^*_2} - y^*_3\sqrt{(y^*_1 + \triangle y_1)(y^*_2 + \triangle y_2)}}{\sqrt{(y^*_1 + \triangle y_1)(y^*_2 + \triangle y_2)y^*_1y^*_2}} \\ &= \frac{(y^*_3 + \triangle y_3)^2y^*_1y^*_2 - (y^*_3)^2(y^*_1 + \triangle y_1)(y^*_2 + \triangle y_2)}{d}, \qquad (12)\end{aligned}$$

where

$$\begin{aligned}d &= \sqrt{(y^*_1 + \triangle y_1)(y^*_2 + \triangle y_2)y^*_1y^*_2} \times \\ &\quad \left[(y^*_3 + \triangle y_3)\sqrt{y^*_1y^*_2} + y^*_3\sqrt{(y^*_1 + \triangle y_1)(y^*_2 + \triangle y_2)}\right]. \qquad (13)\end{aligned}$$

This expression allows us to view $\triangle r$ as approximately the difference between $(y^*_3 + \triangle y_3)^2y^*_1y^*_2$ and $(y^*_3)^2(y^*_1 + \triangle y_1)(y^*_2 + \triangle y_2)$ in the numerator scaled by the denominator d, which, as a scaling factor, can be approximated as $2y^*_1y^*_2y^*_3\sqrt{y^*_1y^*_2}$ by treating each of the $\triangle y_1$, $\triangle y_2$, and $\triangle y_3$ terms there as

zero. With all quadratic terms of $\triangle y_1$, $\triangle y_2$, and $\triangle y_3$ in the numerator — which include $(\triangle y_1)(\triangle y_2)$ and $(\triangle y_3)^2$ — ignored, equation (12) reduces to

$$\begin{aligned} r &= \frac{y_3^*}{\sqrt{y_1^* y_2^*}} + \frac{\left[(y_3^*)^2 + 2y_3^* \triangle y_3\right] y_1^* y_2^* - (y_3^*)^2 (y_1^* y_2^* + y_2^* \triangle y_1 + y_1^* \triangle y_2)}{2y_1^* y_2^* y_3^* \sqrt{y_1^* y_2^*}} \\ &= \frac{y_3^*}{\sqrt{y_1^* y_2^*}} - \frac{y_3^* \triangle y_1}{2y_1^* \sqrt{y_1^* y_2^*}} - \frac{y_3^* \triangle y_2}{2y_2^* \sqrt{y_1^* y_2^*}} + \frac{\triangle y_3}{\sqrt{y_1^* y_2^*}}, \end{aligned} \tag{14}$$

which is analytically equivalent to equation (11).

To estimate the sampling variance of r, it is more convenient to write equation (11) or equation (14) equivalently as

$$r = \alpha_0 + \alpha_1 y_1 + \alpha_2 y_2 + \alpha_3 y_3, \tag{15}$$

where

$$\alpha_0 = \frac{y_3^*}{\sqrt{y_1^* y_2^*}}, \tag{16}$$

$$\alpha_1 = -\frac{y_3^*}{2y_1^* \sqrt{y_1^* y_2^*}}, \tag{17}$$

$$\alpha_2 = -\frac{y_3^*}{2y_2^* \sqrt{y_1^* y_2^*}}, \tag{18}$$

$$\text{and} \quad \alpha_3 = \frac{1}{\sqrt{y_1^* y_2^*}} \tag{19}$$

are coefficients based on the point estimates y_1^*, y_2^*, and y_3^*. Given equation (15), the sampling variance of r is

$$\widehat{Var}(r) = \widehat{Cov}(\alpha_1 y_1 + \alpha_2 y_2 + \alpha_3 y_3, \alpha_1 y_1 + \alpha_2 y_2 + \alpha_3 y_3), \tag{20}$$

which, when expressed explicitly, is the sum of nine terms of the form $\alpha_i \alpha_j \widehat{Cov}(y_i, y_j)$, where each of i and j can be 1, 2, or 3. With $\widehat{Var}(y_i) = \widehat{Cov}(y_i, y_i)$ and $\widehat{Cov}(y_i, y_j) = \widehat{Cov}(y_j, y_i)$, we can write equation (20) explicitly as

$$\begin{aligned} \widehat{Var}(r) &= \alpha_1^2 \widehat{Var}(y_1) + \alpha_2^2 \widehat{Var}(y_2) + \alpha_3^2 \widehat{Var}(y_3) + 2\alpha_1 \alpha_2 \widehat{Cov}(y_1, y_2) \\ &\quad + 2\alpha_1 \alpha_3 \widehat{Cov}(y_1, y_3) + 2\alpha_2 \alpha_3 \widehat{Cov}(y_2, y_3). \end{aligned} \tag{21}$$

After substituting s_{11}, s_{22}, and s_{12} for y_1, y_2, and y_3, respectively, including

the corresponding point estimates, equation (21) becomes

$$\begin{aligned}\widehat{Var}\,(r) \;=\; & \left[\frac{(s_{12}^*)^2}{4(s_{11}^*)^3 s_{22}^*}\widehat{Var}\,(s_{11}) + \frac{(s_{12}^*)^2}{4s_{11}^*(s_{22}^*)^3}\widehat{Var}\,(s_{22}) + \frac{1}{s_{11}^* s_{22}^*}\widehat{Var}\,(s_{12})\right. \\ & + \frac{(s_{12}^*)^2}{2(s_{11}^* s_{22}^*)^2}\widehat{Cov}\,(s_{11}, s_{22}) - \frac{s_{12}^*}{(s_{11}^*)^2 s_{22}^*}\widehat{Cov}\,(s_{11}, s_{12}) \\ & \left. - \frac{s_{12}^*}{s_{11}^*(s_{22}^*)^2}\widehat{Cov}\,(s_{22}, s_{12})\right]. \end{aligned} \tag{22}$$

Analytically equivalent forms of equation (22) can be found in Kwan (2008), Scheinberg (1966), and Stuart and Ord (1987, chapter 10).

4 Spreadsheet-Based Computations

Electronic spreadsheets can facilitate an efficient computation of the sampling variance of the correlation of two variables, thus making the analytical material involved more accessible to students. Notably, spreadsheet functions such as VAR and COVAR in Excel allow us to compute variances and covariances directly. As the computation of $\widehat{Var}\,(r)$ involves various sampling variances and covariances of some underlying random variables, once the relevant data are arranged in matrix forms in a spreadsheet, we can use Excel functions such as TRANSPOSE and MMULT for matrix transposition and multiplication, respectively, to reduce the computational burden even further.

Specifically, by defining a 3-element column vector of coefficients $\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 \end{bmatrix}'$, where the prime indicates transposition of a matrix, and a 3×3 matrix $\mathbf{Z}$ with each (i, j)-element there being $\widehat{Cov}(y_i, y_j)$, we can write equation (21) more compactly as

$$\widehat{Var}(r) = \boldsymbol{\alpha}'\mathbf{Z}\boldsymbol{\alpha}. \tag{23}$$

This equation allows us to use the Excel function MMULT to compute $\widehat{Var}(r)$ directly.

For students who are unfamiliar with matrix algebra, equation (21) is best written as

$$\widehat{Var}(r) = \sum\nolimits_{i=1}^{3}\sum\nolimits_{j=1}^{3} \alpha_i \alpha_j \widehat{Cov}(y_i, y_j) \tag{24}$$

without using the results that $\widehat{Var}(y_i) = \widehat{Cov}(y_i, y_i)$ and $\widehat{Cov}(y_i, y_j) = \widehat{Cov}(y_j,\ y_i)$. As each of i and j can be 1, 2, or 3, the double summation

consists of the nine cases of $\alpha_i \alpha_j \widehat{Cov}(y_i, y_j)$ implicitly covered by equation (21). A pedagogic illustration of the computation involving the same algebraic form is available in Kwan (2007). In the current setting, we place the three coefficients and the nine covariances in a spreadsheet as follows:

	α_1	α_2	α_3
α_1	$\widehat{Cov}(y_1, y_1)$	$\widehat{Cov}(y_1, y_2)$	$\widehat{Cov}(y_1, y_3)$
α_2	$\widehat{Cov}(y_2, y_1)$	$\widehat{Cov}(y_2, y_2)$	$\widehat{Cov}(y_2, y_3)$
α_3	$\widehat{Cov}(y_3, y_1)$	$\widehat{Cov}(y_3, y_2)$	$\widehat{Cov}(y_3, y_3)$

Once we multiply each covariance term by the corresponding coefficients in the same row and in the same column, we have the following:

$\alpha_1 \alpha_1 \widehat{Cov}(y_1, y_1)$	$\alpha_1 \alpha_2 \widehat{Cov}(y_1, y_2)$	$\alpha_1 \alpha_3 \widehat{Cov}(y_1, y_3)$
$\alpha_2 \alpha_1 \widehat{Cov}(y_2, y_1)$	$\alpha_2 \alpha_2 \widehat{Cov}(y_2, y_2)$	$\alpha_2 \alpha_3 \widehat{Cov}(y_2, y_3)$
$\alpha_3 \alpha_1 \widehat{Cov}(y_3, y_1)$	$\alpha_3 \alpha_2 \widehat{Cov}(y_3, y_2)$	$\alpha_3 \alpha_3 \widehat{Cov}(y_3, y_3)$

With each element in this 3×3 block being one of the nine cases of $\alpha_i \alpha_j$ $\widehat{Cov}(y_i, y_j)$ in equation (24), the sum is the sampling variance of r. The computations involved can easily be performed on Excel as well.

Notice that, as equations (23) and (24) show, $\widehat{Var}(r)$ is a linear function of nine individual terms of the form $\widehat{Cov}(y_i, y_j)$, where each of i and j can be 1, 2, or 3. It is equivalent to a linear combination of various cases of $\widehat{Cov}\left(s_{ij}, s_{kl}\right)$, where each of i, j, k, and l can be 1 or 2. Given equations (7) and (9), as well as $y_1 = s_{11}$, $y_2 = s_{22}$, and $y_3 = s_{12}$, we can also write $\widehat{Var}(r)$ as $N/(1-N)^2$ multiplied by $\boldsymbol{\alpha}' \mathbf{W} \boldsymbol{\alpha}$, where each element of the 3×3 matrix $\mathbf{W}$ is one of the various cases of $\widehat{Cov}\left(w_{ij}, w_{kl}\right)$. With $N/(1-N)^2$ varying asymptotically as $1/N$ and with $\boldsymbol{\alpha}' \mathbf{W} \boldsymbol{\alpha}$ being always positive and finite, it follows that, for a given joint distribution of the two underlying variables x_1 and x_2, an increase in the number of observations tends to result in a lower $\widehat{Var}(r)$.

For a numerical illustration with Excel, we use daily return data of the Dow Jones Industrial Average (DJIA) of 30 U.S. stocks and the Financial Times Stock Exchange Index (FTSE) of 100 U.K. stocks, over 25 trading days from November 24, 2008 to December 31, 2008.[5] As our main purpose

[5] Daily closing values of the two indices, under the ticker symbols ^DJI and ^FTSE, are freely available from Yahoo! Finance <http://finance.yahoo.com/> on the internet. For this numerical illustration, a day is considered a trading day when the closing values for both indices are available. For each index, the return on trading day n is the change in the closing index values from trading day $n-1$ to trading day n, as a proportion of the closing index value at trading day $n-1$.

here is to illustrate the computational procedure, whether the use of only 25 pairs of return observations is adequate for the estimation is not an issue. To illustrate the impact of changes in the number of observations on the estimation results, we also attempt as many as 250 pairs of daily return observations, from December 31, 2007 to December 31, 2008. Specifically, for each of the cases involving $25, 26, 27, \ldots, 250$ consecutive trading days, the observations always end at December 31, 2008. Some essential results will be provided subsequent to the Excel example.

In the Excel example in Figure 1, the daily return data, including the corresponding dates, are displayed in A3:C27 of the spreadsheet. The number of observations, the sample mean, the sample variance, and the sample standard deviation of each of the index returns are shown in B29:C32. The corresponding cell formulas, along with those for the remaining computations in this example, are listed in B56:C92. The sample covariance of the two variables is provided in B34.[6] The sample correlation is computed in two equivalent ways; while B35 is based on the sample covariance divided by the product of the two sample standard deviations, B36 is computed directly by using the Excel function CORREL.

The 25 observations of w_{11}, w_{22}, and w_{12} under the headings of "1: DJIA, DJIA, 2: FTSE,FTSE," and "3: DJIA,FTSE," each being a product of mean-removed returns as defined in equation (3), are shown in E3:E27, F3:F27, and G3:G27, respectively. As these headings contain the labels 1, 2, and 3, it is implicit that $y_1 = s_{11}$, $y_2 = s_{22}$, and $y_3 = s_{12}$. The numbers of observations as required for subsequent computations are displayed in E29:G29.

To compute the sampling variance of the correlation by matrix multiplications, we first set up the row vector of coefficients $\boldsymbol{\alpha}' = \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 \end{bmatrix}$ in E38:G38, where the three individual elements are as defined in equations (17)-(19). The corresponding column vector $\boldsymbol{\alpha}$, as displayed in I40:I42, is obtained by using the Excel function TRANSPOSE. The 3×3 matrix $\mathbf{Z}$, with each (i, j)-element being $\widehat{Cov}(y_i, y_j)$, is placed in E40:G42. The computations of these matrix elements are based on equations (7) and (9). The sampling variance of r, as shown in E44, is the result of the matrix multiplication $\boldsymbol{\alpha}'\mathbf{Z}\boldsymbol{\alpha}$ by using the Excel function MMULT repeatedly. To show more explicitly the magnitude of the error relative to the estimated value r^*, we also provide in E45:E46 the standard error $SE(r) = \sqrt{\widehat{Var}(r)}$ and the coefficient of variation, which is the ratio $SE(r)/r^*$.

[6] As the Excel function COVAR treats each sample as its population, a multiplicative factor $N/(N-1)$ is required to correct the bias in the estimated covariance from N pairs of observations. (See B63:C63 and B78:C80.)

	A	B	C	D	E	F	G	H	I
1		Returns			Products of Mean-Removed Returns				
2	Date	DJIA	FTSE		1: DJIA,DJIA	2: FTSE,FTSE	3: DJIA,FTSE		
3	24/11/2008	0.049335	0.098387		0.002069292	0.008392686	0.004167364		
4	25/11/2008	0.004273	0.004406		0.000000183	0.000005610	-0.000001013		
5	26/11/2008	0.029146	-0.004459		0.000640099	0.000126205	-0.000284225		
6	28/11/2008	0.011738	0.032581		0.000062286	0.000665957	0.000203666		
7	01/12/2008	-0.077013	-0.051889		0.006538083	0.003441472	0.004743483		
8	02/12/2008	0.033133	0.014119		0.000857730	0.000053931	0.000215076		
9	03/12/2008	0.020501	0.011424		0.000277406	0.000021613	0.000077430		
10	04/12/2008	-0.025077	-0.001535		0.000836486	0.000069053	0.000240338		
11	05/12/2008	0.030942	-0.027428		0.000734235	0.001169862	-0.000926798		
12	08/12/2008	0.034597	0.061910		0.000945657	0.003039906	0.001695496		
13	09/12/2008	-0.027182	0.018883		0.000962714	0.000146609	-0.000375689		
14	10/12/2008	0.008064	-0.003195		0.000017799	0.000099410	-0.000042064		
15	11/12/2008	-0.022408	0.004900		0.000689271	0.000003516	0.000049227		
16	12/12/2008	0.007541	-0.024677		0.000013657	0.000989233	-0.000116233		
17	15/12/2008	-0.007550	-0.000654		0.000129847	0.000055193	0.000084656		
18	16/12/2008	0.041988	0.007364		0.001454872	0.000000347	0.000022461		
19	17/12/2008	-0.011183	0.003504		0.000225861	0.000010698	0.000049157		
20	18/12/2008	-0.024857	0.001503		0.000823856	0.000027793	0.000151319		
21	19/12/2008	-0.003008	-0.010114		0.000046965	0.000285235	0.000115741		
22	22/12/2008	-0.006926	-0.008794		0.000116028	0.000242403	0.000167707		
23	23/12/2008	-0.011761	0.001600		0.000243563	0.000026778	0.000080760		
24	24/12/2008	0.005819	-0.009258		0.000003893	0.000257044	-0.000031634		
25	29/12/2008	0.001824	0.024380		0.000004085	0.000309928	-0.000035581		
26	30/12/2008	0.021742	0.016970		0.000320294	0.000103936	0.000182456		
27	31/12/2008	0.012459	0.009447		0.000074193	0.000007142	0.000023019		
28									
29	No. of Obs.	25	25		25	25	25		
30	Mean	0.003846	0.006775						
31	Variance	0.000754	0.000815						
32	St. Dev.	0.027453	0.028542						
33									
34	Covariance	0.000436							
35	Correlation	0.556007							
36	Cor. (Direct)	0.556007							
37									
38				Coef.	-368.8607	-341.2558	1276.2066		
39									Coef.
40	Matrix of Sampl. Variances & Covariances				0.00000007550	0.00000005419	0.00000006001		-368.8607
41					0.00000005419	0.00000014358	0.00000008542		-341.2558
42					0.00000006001	0.00000008542	0.00000007186		1276.2066
43									
44	Sampling Var. of Cor. (by Mat. Mult.)				0.026770				
45	Standard Error of Correlation				0.163616				
46	Coefficient of Variation				0.294269				

Figure 1: An Excel Example of Estimation Error in the Correlation of Two Variables.

	A	B	C	D	E	F	G	H	I
47									
48	Prod. of Coef.'s & Sampl. Var.'s or Cov.'s				0.010273	0.006821	-0.028252		
49					0.006821	0.016721	-0.037201		
50					-0.028252	-0.037201	0.117040		
51									
52	Sampling Var. of Cor. (by Summation)				0.026770				
53	Standard Error of Correlation				0.163616				
54	Coefficient of Variation				0.294269				
55									
56	Formulas:	B29	=COUNT(B3:B27)						
57		B30	=AVERAGE(B3:B27)						
58		B31	=VAR(B3:B27)						
59		B32	=SQRT(B31)						
60			Copy B29:B32 to B29:C32						
61			Copy B29 to E29:G29						
62									
63		B34	=COVAR(B3:B27,C3:C27)*B29/(B29-1)						
64		B35	=B34/(B32*C32)						
65		B36	=CORREL(B3:B27,C3:C27)						
66									
67		E3	=(B3-B$30)*(B3-B$30)						
68		F3	=(C3-C$30)*(C3-C$30)						
69		G3	=(B3-B$30)*(C3-C$30)						
70			Copy E3:G3 to E3:G27						
71									
72		E38	=-B34/(2*B31*SQRT(B31*C31))						
73		F38	=-B34/(2*C31*SQRT(B31*C31))						
74		G38	=1/SQRT(B31*C31)						
75									
76		I40:I42	{=TRANSPOSE(E38:G38)}						
77									
78		E40	=COVAR($E$3:$E$27,E$3:E$27)*E$29*E$29/((E$29-1)*(E$29-1)*(E$29-1))						
79		E41	=COVAR($F$3:$F$27,E$3:E$27)*E$29*E$29/((E$29-1)*(E$29-1)*(E$29-1))						
80		E42	=COVAR($G$3:$G$27,E$3:E$27)*E$29*E$29/((E$29-1)*(E$29-1)*(E$29-1))						
81			Copy E40:E42 to E40:G42						
82									
83		E44	=MMULT(E38:G38,MMULT(E40:G42,I40:I42))						
84		E45	=SQRT(E44)						
85		E46	=E45/B36						
86									
87		E48	=E$38*$I40*E40						
88			Copy E48 to E48:G50						
89									
90		E52	=SUM(E48:G50)						
91		E53	=SQRT(E52)						
92		E54	=E53/B36						

Figure 1: An Excel Example of Estimation Error in the Correlation of Two Variables (continued).

As an alternative to matrix multiplication, we compute $\widehat{Cov}(r)$ by summing the nine cases of $\alpha_i \alpha_j \widehat{Cov}(y_i, y_j)$ in E48:G50. As expected, the sum in E52 is the same as what is shown in E44. The computations of the standard error $SE(r)$ and the coefficient of variation $SE(r)/r^*$, which have been performed in E45:E46, are repeated in E53:E54. Students who are unfamiliar with matrix operations can skip rows 44 to 46 of the spreadsheet.

With the estimated correlation $r^* = 0.556007$, the sampling variance $\widehat{Var}(r) = 0.026770$, the standard error $SE(r) = 0.163616$, and the coefficient of variation $SE(r)/r^* = 0.294269$, estimation error in the correlation based on 25 pairs of observations accounts for nearly 30% of its estimated value. Once 250 pairs of observations are used instead, we have $r^* = 0.526081$, $\widehat{Var}(r) = 0.004497$, $SE(r) = 0.067058$, and $SE(r)/r^* = 0.127467$. Not surprisingly, an increase of the number of observations by an order of magnitude (from 25 to 250) has resulted in a decrease of $\widehat{Var}(r)$ by nearly an order of magnitude (from 0.026770 to 0.004497). As the standard error based on 250 pairs of observations still accounts for over 12% of its estimated value, estimation error in the correlation is far from being negligible here.

Figure 2 shows graphically how the standard error $SE(r)$ varies with the number of observations N. As N increases gradually from 25, the graph of $SE(r)$ exhibits some initial fluctuations but with a clear downward trend; the trend becomes much less prominent when N reaches about 75. Such results suggest that, under the stationarity assumption of the joint distribution of daily returns of the two market indices considered, at least approximately 75 observations are required for a reasonable estimate of the correlation. However, as the case of $N = 75$ — which is based on daily returns from September 15, 2008 to December 31, 2008 — gives us $r^* = 0.546767$, $\widehat{Var}(r) = 0.007356$, $SE(r) = 0.085766$, and $SE(r)/r^* = 0.156861$, the standard error still accounts for over 15% of the estimated correlation.[7]

[7] Given the recent changes in the global economic conditions, which have evolved into a global stock market turmoil, the return observations since October 2008 could be viewed as being from a different joint distribution of the two underlying variables. Under such a view, the reliance on more historical return data for the estimation to bypass the issue of estimation error would not be a viable option, as long as one's interest is in knowing the current correlation. As indicated earlier, this numerical example is intended to illustrate pedagogically the computational detail. Thus, although the stationarity assumption of the joint distribution of the two underlying variables is required for the estimation results to be meaningful, to test for the validity of such an assumption is outside the scope of this pedagogic study.

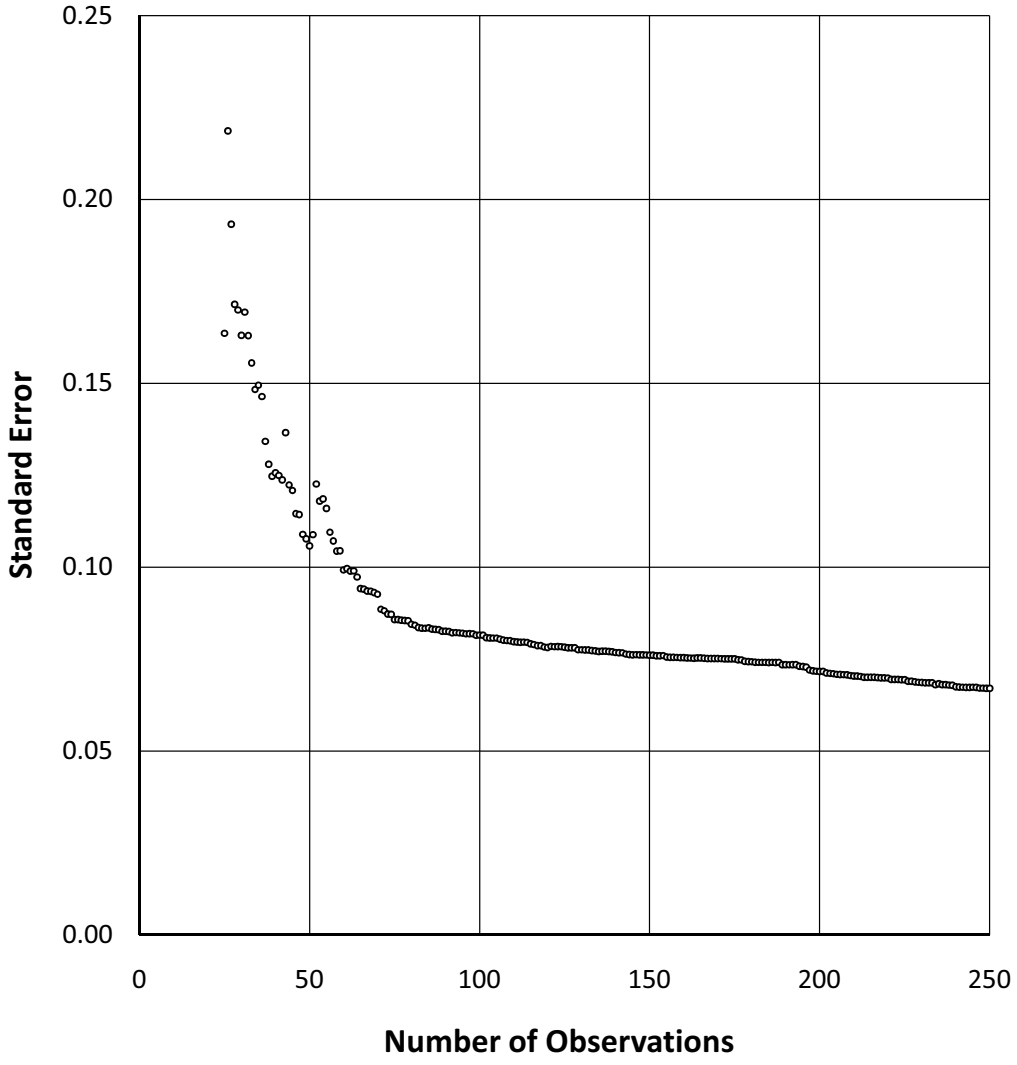

Figure 2: A Graph of the Standard Error (the Square Root of the Sampling Variance of the Estimated Correlation) versus the Number of Observations, Based on 25 to 250 Daily Returns of the Dow Jones Industrial Average (DJIA) and the Financial Times Stock Exchange Index (FTSE).

5 Concluding Remarks

The statistical term *correlation* is well-known across many academic disciplines. Researchers use estimated correlations of variables from experimental or empirical data to draw implications relevant to their own research fields. Students are taught how to estimate correlations and to interpret the correlation results. However, the derivations of significance tests (for the sample correlation) and confidence intervals, even under simplifying assumptions for analytical convenience, still require statistical concepts that are unfamiliar to most students outside statistical fields. Given the practical importance of the concept of correlation, therefore, a challenging question for instructors to consider is whether it is possible to teach students estimation error in the correlation by using only familiar mathematical and statistical tools.

In this pedagogic study, we have presented an approach to estimate the error in the sample correlation, with the required statistical concepts set at a level suitable for students outside statistical fields. We have presented the same approach with and without using multivariate differential calculus. Thus, prior knowledge of advanced calculus is not essential for understanding the analytical material here. Students with general algebraic skills are expected to be able to follow all the material involved. Consider, for example, undergraduate commerce students who have taken an introductory finance course. These students already know how to express the variance of a linear combination of a small number of variables — in the context of expressing an investment portfolio's random return as a weighted average of the random returns of the underlying assets — in terms of the variances and covariances of these variables. In fact, the expressions in equations (23) and (24) are in the same algebraic forms as those for an investment portfolio's variance of returns, although the coefficients and the covariances here and those pertaining to an investment portfolio are in two very different contexts.[8]

The analytical expression of estimation error in the correlation as derived in this study has no specific distributional requirements on the two random variables, apart from the implicit assumption of stationarity, which allows the individual observations in a sample to be treated as random draws. Under the stationarity assumption, this study has assessed the accuracy of the sample correlation (in terms of its sampling variance) as an estimator of the unknown population correlation. Further, this study has applied familiar statistical concepts to a setting that utilizes analytical tools for error prop-

[8] See, for example, Kwan (2007) for the corresponding algebraic expressions in the context of portfolio theory.

agation. Science and engineering students, for example, are expected to be aware of various experimental settings where each variable of interest can be expressed as a function of some other variables, for which experimental values are available. As experimental values are inevitably subject to measurement errors, so are the variables of interest. A common approach in analyzing error propagation is by using a first-order Taylor expansion of the function involved. In the case of the estimated correlation, it is a function of the estimated variances and covariance of the two underlying variables, for which the individual estimation errors are available. From a pedagogic perspective, therefore, it is useful for instructors to remind students the similarity between the approach here (pertaining to the Taylor expansion) and what students likely have learned about error propagation elsewhere.

As electronic spreadsheets such as Excel are now well-known computational tools, we as instructors are becoming less constrained in our efforts to cover relevant topics that have traditional been considered to be too advanced or computationally too tedious for students. Estimation error in the correlation of two variables is one of such topics.[9] What this pedagogic study intends to achieve does go beyond providing a simple recipe for estimating the error in the sample correlation. At a more fundamental level, it reminds students that point estimates from a sample of observations — whether they pertain to estimates of variances, covariances, or correlations — are subject to estimation errors and that the magnitudes of such errors depend on the data involved. It then provides the analytical detail for computing such errors. By using Excel functions that students are already familiar with for the required computations, it also makes the corresponding analytical material

[9]The choice of computer software for pedagogic purposes depends on many factors. Besides cost and functionality considerations, an important factor is the familiarity of its available features, not only to the instructor of the course involved, but also to the teaching assistants and technical support staff. As Excel is part of Microsoft Office, installed by many educational institutions for their students to access, its basic features are likely to have been familiar to many students (prior to enrolling in courses requiring certain specific knowledge of its more advanced features). With some technical support, students can acquire the necessary skills to perform the computational tasks in such courses. Although spreadsheets, such as Excel, are not as versatile as many other available computational packages in terms of functionality, their operational simplicity, nonetheless, is a practical advantage. Indeed, from the classroom experience of the author, as an instructor of various finance and investment courses, familiar Excel features are already adequate for facilitating the delivery of many advanced topics in these courses. In contrast, if the selected computational software also requires students to learn a new programming language, then much more extensive technical support would be required. This language burden could potentially undermine the instructor's efforts to introduce challenging but relevant topics to students.

less abstract and thus accessible to more students across different academic disciplines.

Acknowledgement: The author wishes to thank the anonymous reviewers for helpful comments and suggestions.

References

Kwan, C.C.Y. (2007) A Simple Spreadsheet-Based Exposition of the Markowitz Critical Line Method for Portfolio Selection. *Spreadsheets in Education*, 2(3): 253-280.

Kwan, C.C.Y. (2008) Estimation Error in the Average Correlation of Security Returns and Shrinkage Estimation of Covariance and Correlation Matrices. *Finance Research Letters*, 5: 236-244.

Schäfer, J., and Strimmer, K. (2005) A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1): Article 32.

Scheinberg, E. (1966) The Sampling Variance of the Correlation Coefficients Estimated in Genetic Experiments. *Biometrics*, 22(1): 187-191.

Stuart, A., and Ord, J.K. (1987) *Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory*, 5th Edition, Charles Griffin & Co., London.

Warner, R.M. (2007) *Applied Statistics: From Bivariate Through Multivariate Techniques*, SAGE Publications, Thousand Oaks, CA.