Fixed Accuracy Interval Estimation of the Common Variance in a Equi-Correlated Normal Distribution

Shyamal K. De[1] **and Nitis Mukhopadhyay**[2]

[1]School of Mathematical Sciences, National Institute of Science Education and Research, Bhubaneswar, Odisha, India

[2]Department of Statistics, University of Connecticut, Storrs, Connecticut, USA

Abstract: Two-stage procedures are proposed to construct *fixed-accuracy* confidence intervals for the common variance of a equi-correlated normal distribution. The exact distributions of the stopping variable and the estimator of the common variance at stopping are derived. We also derive exact formulas for the expectations of functions of the stopping variable and the estimator of the common variance at stopping. The coverage probabilities of the proposed interval estimator are computed exactly and are shown to be nearly the same as the prescribed level.

Keywords: Common variance; Equi-correlated normal; Exact computations; Fixed-accuracy; Simulations; Stopping variable; Two-stage sampling.

Subject Classifications: 60G51; 60K15; 60K40.

1. INTRODUCTION

Correlated multivariate normal distributions appear in many areas of statistics such as ANOVA design with repeated measurements. Suppose that there are n experimental units, and each unit is measured on a certain variable at m time periods. Thus, m measurements from one unit are correlated. However, the measurements from different units can be assumed independent.

In an ANOVA design with repeated measurements, the error part can be modeled by an m-variate normal distribution with mean vector $\mathbf{0}$ and covariance matrix Σ. We investigate two-stage estimation problems for the elements in Σ when Σ has a special structure. Multi-stage sampling designs were developed by Mahalanobis (1940) followed by pathbreaking constructions of two-stage methodologies due to Stein (1945, 1949). For a review, one may refer to Siegmund (1985), Ghosh and Sen (1991), Ghosh et al. (1997), and Mukhopadhyay and de Silva (2009), Zacks (2009a) among other sources.

Address correspondence to Shyamal K. De, School of Mathematical Sciences, National Institute of Science Education and Research, Bhubaneswar, Odisha 751005, India; Tel: +91 (674) 230-4143; Fax: +91 (674) 230-4081; E-mail: sde@niser.ac.in

Let us now formulate our specific problem on hand precisely. Consider n independently and identically distributed m-dimensional vectors $\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n}$ from a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\Sigma = \sigma^2 R$, where R is an unknown correlation matrix. We consider the equicorrelated model considered by Rao (1973, p. 67), Mukhopadhyay (1982), Zacks and Ramig (1987), Ghezzi and Zacks (2005), Haner and Zacks (2013), and De (2014). That is, each entry of R is the common correlation, $\rho \in (-\frac{1}{m-1},\ 1)$, between any two components of $\mathbf{X_1}$. Let us denote the parameter vector $\theta = (\rho,\ \sigma^2)$.

Mukhopadhyay and Banerjee (2014) and Banerjee and Mukhopadhyay (2014) drew attention to certain important drawbacks inherent in the constructions of *fixed-width* (Chow and Robbins 1965, Khan 1969) and *proportional accuracy* (Zacks 1966, Nadas 1969, Willson and Folks 1983) confidence intervals for an unknown parameter whose support is, for example, R^+ or $(0, 1)$. Mukhopadhyay and Banerjee (2014) and Banerjee and Mukhopadhyay (2014) then resolved such drawbacks by introducing the idea of a *fixed accuracy* confidence interval which always lies in a subset of the parameter space under consideration.

Our goal is to estimate σ^2 with a prescribed *fixed-accuracy* in the presence of the nuisance parameter ρ in the spirit of Mukhopadhyay and Banerjee (2014) and Banerjee and Mukhopadhyay (2014). That is, for fixed $\delta > 1$ and $\alpha \in (0, 1)$, we would like to have

$$P_\theta\left(\delta^{-1} < \frac{\widehat{\sigma}_n^2}{\sigma^2} < \delta\right) \geq 1 - \alpha \quad \text{ for all fixed } \theta. \tag{1.1}$$

The associated interval estimator $(\delta^{-1}\widehat{\sigma}_n^2, \delta\widehat{\sigma}_n^2)$ is called a fixed-accuracy confidence interval of σ^2 in the sense of Mukhopadhyay and Banerjee (2014) and Banerjee and Mukhopadhyay (2014). In some sense, the probability in (1.1) measures the chance of closeness between $\widehat{\sigma}_n^2\sigma^{-2}$ and 1, but this construction is different from interval estimators with fixed-width or proportional closeness.

The exact formulas for evaluating risk functions and other useful functionals in the context of a number of two-stage and purely sequential methodologies in gamma and exponential distributions were systematically developed by Zacks (2009b), Zacks and Mukhopadhyay (2006a,b), and Zacks and Khan (2011). Fixed-sample-size estimation of the common variance of correlated multinormal distributions was first considered by Zacks and Ramig (1987). They developed the uniform minimum variance unbiased estimators and the maximum likelihood estimators of σ^2 and ρ. The dissertation of Haner (2012) and the subsequent paper by Haner and Zacks (2013) developed theory and methodologies for fixed-width interval estimation of σ^2. Recently, De (2014) developed a modified three-stage procedure for fixed-width interval estimation of σ^2.

One drawback of these approaches is that one needs to have some prior information about σ^2 in order to select the fixed-width appropriately. For example, when the true value of σ^2 is 0.1, then it does not make sense to choose the fixed-width as 0.2. Since in many practical situations, one may not have a prior information about the true value of σ^2, and it may be difficult to select a certain width that will work across the board.

Moreover, with preassigned $d > 0$ and a fixed-width confidence interval $(\widehat{\sigma}_n^2 - d, \widehat{\sigma}_n^2 + d)$, there is a positive probability that $\widehat{\sigma}_n^2 - d < 0$ for all σ^2 and ρ. The proposed estimator in (1.1) overcomes such drawbacks.

In this paper, we begin with some preliminaries (Section 2) and then move forward quickly to develop

an appropriate two-stage methodology (Section 3) to construct a fixed-accuracy confidence interval for σ^2. The exact distributions of the stopping variable and the estimator of σ^2 at stopping are derived (Section 4). The exact formulas for coverage probability and other useful functionals are derived by significantly modifying the techniques from Zacks and Ramig (1987), Zacks and Mukhopadhyay (2007), Haner and Zacks (2013), De (2014), and it is shown (Section 6) that the attained coverage probability associated with our proposed interval estimator is nearly the same as the prescribed level. Choice of a pilot size plus a number of desirable asymptotic properties are developed in Section 5. We establish the asymptotic first-order efficiency and asymptotic consistency properties of the proposed methodology. A selected batch of numerical results obtained using both exact calculations and simulations are presented critically in Section 6 which are followed by some concluding comments (Section 7).

2. PRELIMINARY THEORY

This section discusses some preliminary theory needed to describe the two-stage methodology included later in Section 3. A part of such preliminaries is also found in Haner and Zacks (2013) and De (2014).

In order to obtain a sequence of uncorrelated multinormal vectors $\mathbf{Y_1}, \mathbf{Y_2}, \ldots$, from the original observations $\mathbf{X_1}, \mathbf{X_2}, \ldots$, let us first use the transformation $\mathbf{Y}_i = H\mathbf{X}_i$, for $i = 1, 2, \ldots$, where H is the $m \times m$ Helmert orthogonal matrix (Mukhopadhyay 2000, pp. 197-201), namely

$$H = \begin{bmatrix} \frac{1}{\sqrt{m}} & \frac{1}{\sqrt{m}} & \cdots & & \frac{1}{\sqrt{m}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & 0 \quad \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\sqrt{m(m-1)}} & & \cdots & \frac{1}{\sqrt{m(m-1)}} & -\frac{(m-1)}{\sqrt{m(m-1)}} \end{bmatrix}.$$

The transformed random vectors $\mathbf{Y_1}, \mathbf{Y_2}, \ldots$ are distributed as $N_m(\mathbf{0},\, \sigma^2\Lambda)$, where

$$\Lambda = HRH^T = \text{diag}(\lambda_1, \ldots, \lambda_m)$$

is a diagonal matrix, and $\lambda_1, \ldots, \lambda_m$ are the eigenvalues of R. In an equi-correlated model, the correlation matrix $R = (1-\rho)I_m + \rho J$, where I_m is the $m \times m$ identity matrix and $J = \mathbf{1}\mathbf{1}^T$ is a $m \times m$ matrix of 1's. The eigenvalues have the following form (Rao 1973, p. 67)

$$\lambda_1 = 1 + (m-1)\rho \quad \text{and} \quad \lambda_2 = \cdots = \lambda_m = 1 - \rho.$$

Note that the condition $-\frac{1}{m-1} < \rho < 1$ is needed to ensure $\lambda_1 > 0$. Given $\mathbf{Y_1}, \ldots, \mathbf{Y_n}$, the likelihood

function of (σ^2, ρ) is

$$L(\sigma^2, \rho; V_{1n}, V_{2n}) \propto (\sigma^2)^{-\frac{nm}{2}} (1-\rho)^{-\frac{n(m-1)}{2}} (1+(m-1)\rho)^{-\frac{n}{2}}$$
$$\times \exp\left\{-\frac{1}{2\sigma^2}\left(\frac{V_{1n}}{1+(m-1)\rho} + \frac{V_{2n}}{1-\rho}\right)\right\},$$

where

$$V_{1n} = \sum_{i=1}^{n} Y_{i1}^2 \text{ and } V_{2n} = \sum_{i=1}^{n}\sum_{j=2}^{m} Y_{ij}^2.$$

Observe that (V_{1n}, V_{2n}) is a complete sufficient statistic (Lehmann 1997) of (σ^2, ρ), and thus, it is natural to consider estimators of unknown parameters based on (V_{1n}, V_{2n}). The properties of (V_{1n}, V_{2n}) that are crucial in the sequel for the derivation of exact distributions are as follows:

(i) $V_{1n} \sim \sigma^2 (1+(m-1)\rho)\, \chi^2_n$,

(ii) $V_{2n} \sim \sigma^2 (1-\rho)\, \chi^2_{n(m-1)}$, and

(iii) V_{1n} and V_{2n} are independent.

Zacks and Ramig (1987) show that the maximum likelihood estimator of σ^2 and ρ are respectively

$$\widehat{\sigma}_n^2 = \frac{V_{1n} + V_{2n}}{nm} \qquad \text{and} \tag{2.1}$$

$$\widehat{\rho}_n = \frac{V_{1n} - V_{2n}(m-1)^{-1}}{V_{1n} + V_{2n}}. \tag{2.2}$$

Both these estimators are strongly consistent and asymptotically normal. The asymptotic distribution of $\sqrt{n}\left(\widehat{\sigma}_n^2 - \sigma^2\right)$ is

$$N\left(0,\, \frac{2\sigma^4}{m}\left\{1+(m-1)\rho^2\right\}\right) \text{ as } n \to \infty.$$

Thus, using the delta method, $\sqrt{n}(\ln \widehat{\sigma}_n^2 - \ln \sigma^2)$ is asymptotically distributed as

$$N\left(0,\, \frac{2}{m}\left\{1+(m-1)\rho^2\right\}\right) \text{ as } n \to \infty. \tag{2.3}$$

For some prefixed $\delta > 1$ and $\alpha \in (0,1)$, we need to find an expression of the required optimal fixed-sample-size n such that

$$P\left(\delta^{-1}\widehat{\sigma}_n^2 < \sigma^2 < \delta\widehat{\sigma}_n^2\right) \geq 1-\alpha \text{ approximately.} \tag{2.4}$$

Using (2.3) to solve (2.4), we obtain the sample size n as the smallest integer:

$$n \geq \beta(\delta)\left\{1+(m-1)\rho^2\right\},$$

where $\beta(\delta) = 2z^2_{\alpha/2} / \left(m \ln(\delta)^2\right)$ and $z_{\alpha/2}$ is the upper $\alpha/2$-quantile of a standard normal distribution.

Let the notation $\lfloor x \rfloor$ represent the largest integer less than x. Then, the required *optimal fixed-sample-size* needed to achieve (2.4) is approximately given by

$$n_0 \equiv n_0(\rho, \delta) = \left\lfloor \beta(\delta) \left\{1 + (m-1)\rho^2\right\} \right\rfloor + 1, \tag{2.5}$$

had the actual value of ρ been known.

But since ρ is indeed an unknown parameter, one must collect observations in two stages or more to obtain the fixed-accuracy interval in (1.1). In the next section, we develop a two-stage estimation methodology in the original spirit of Stein (1945, 1949) to achieve (2.4).

3. TWO-STAGE SAMPLING

In this section, we present a two-stage sampling strategy to construct a fixed-accuracy confidence interval of σ^2 using a fixed pilot sample size k. Later, we will discuss some possible choices of a pilot sample size and their advantages or disadvantages.

Stage I: Fix the pilot sample size k and draw an initial random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ from m-variate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\sigma^2 R$. Transform these observations to $\mathbf{Y}_i = H\boldsymbol{X}_i$, $i = 1, \ldots, k$, and obtain (V_{1k}, V_{2k}) and the MLE $\widehat{\rho}_k$ as in (2.2) based on $\mathbf{Y}_1, \ldots, \mathbf{Y}_k$. Then, obtain the estimated sample size

$$K^* \equiv K^*(\delta) = \lfloor \beta(\delta) \left\{1 + (m-1)\widehat{\rho}_k^2\right\} \rfloor + 1. \tag{3.1}$$

Suppose N is the stopping variable, that is, the final sample size for this two-stage methodology. If $K^* \leq k$, we stop sampling and set $N = k$ and $\widehat{\sigma}_N^2 = \widehat{\sigma}_k^2$ using (2.1). Otherwise, we proceed to Stage II.

Stage II: If $K^* > k$, we draw remaining $(K^* - k)$ random samples from $N_m\left(\mathbf{0},\, \sigma^2 R\right)$ independent of the initial k samples and label the observations as $\boldsymbol{X}_{k+1}^*, \ldots, \boldsymbol{X}_{K^*}^*$. Transform these observations to $\mathbf{Y}_i^* = H\boldsymbol{X}_i^*$, $i = k+1, \ldots, K^*$. We define

$$V_{1(K^*-k)}^* = \sum_{i=k+1}^{K^*} Y_{i1}^{*2} \text{ and } V_{2(K^*-k)}^* = \sum_{i=k+1}^{K^*} \sum_{j=2}^{m} Y_{ij}^{*2},$$

and set $N = K^*$. The final estimator of σ^2 is

$$\widehat{\sigma}_N^2 = \frac{V_{1k} + V_{1(N-k)}^* + V_{2k} + V_{2(N-k)}^*}{mN}.$$

The fixed-accuracy confidence interval for σ^2 is

$$\left(\delta^{-1}\widehat{\sigma}_N^2,\, \delta\widehat{\sigma}_N^2\right)$$

where the final sample size is given by

$$N \equiv N(\delta) = \max\left\{k,\, K^*(\delta)\right\}. \tag{3.2}$$

4. EXACT DISTRIBUTION OF N AND $\widehat{\sigma}^2_N$

In order to obtain the exact distribution of N, let us first derive the exact distribution of $\widehat{\rho}_n$. Suppose $F_{n,\,n(m-1)}$ denotes a random variable having F-distribution with numerator and denominator degrees of freedom n and $n(m-1)$ respectively. For any fixed positive integer k, let $G(.)$ be the distribution function of $F_{k,\,k(m-1)}$ and $F_{\widehat{\rho}_k}$ be the distribution function of $\widehat{\rho}_k$.

Lemma 4.1. *The distribution function of $\widehat{\rho}_k$ is given by*

$$F_{\widehat{\rho}_k}(x) = \begin{cases} 0 & \text{if } x \le -\frac{1}{m-1} \\ G\left[\dfrac{(1-\rho)\{m-(m-1)(1-x)\}}{(1-x)\{1+(m-1)\rho\}}\right] & \text{if } -\frac{1}{m-1} < x < 1 \\ 1 & \text{if } x \ge 1. \end{cases}$$

Proof. Since $V_{1n} \sim \sigma^2(1+(m-1)\rho)\chi^2_n$ and $V_{2n} \sim \sigma^2(1-\rho)\chi^2_{n(m-1)}$ independently, by definition of F-distribution, $F_{k,\,k(m-1)} = \frac{(1-\rho)(m-1)}{1+(m-1)\rho}\frac{V_{1k}}{V_{2k}}$. Recall the expression of $\widehat{\rho}_k$ from (2.2) and write

$$\begin{aligned}\widehat{\rho}_k &= 1 - \frac{m}{m-1}\frac{V_{2k}}{V_{2k}+V_{1k}} = 1 - \frac{m}{m-1}\left(1+\frac{1+(m-1)\rho}{(1-\rho)(m-1)}F_{k,\,k(m-1)}\right)^{-1} \\ &= 1 - \frac{m(1-\rho)}{(1-\rho)(m-1)+\{1+(m-1)\rho\}F_{k,\,k(m-1)}}.\end{aligned}$$

Therefore, using the above expression of $\widehat{\rho}_k$, we have

$$\begin{aligned}F_{\widehat{\rho}_k}(x) &= P_\theta\left(\widehat{\rho}_k \le x\right) = P_\theta\left((1-\rho)(m-1)+\{1+(m-1)\rho\}F_{k,\,k(m-1)} \le \frac{m(1-\rho)}{1-x}\right) \\ &= G\left[\frac{(1-\rho)\{m-(m-1)(1-x)\}}{(1-x)\{1+(m-1)\rho\}}\right] \quad \text{whenever } \frac{-1}{m-1} < x < 1.\end{aligned}$$

Clearly, if $x \le \frac{-1}{m-1}$, then $m-(m-1)(1-x) \le 0$ and hence $F_{\widehat{\rho}_k}(x) = 0$. □

Note that the distribution function of $\widehat{\rho}_n$ does not depend on σ^2. Next, we derive the exact distribution of N and show that it also does not depend on σ^2. Note that

$$P_\theta(N=k) = P_\theta(K^* \le k) = P_\theta\left(\widehat{\rho}^2_k \le \left(\frac{k}{\beta}-1\right)\frac{1}{m-1}\right).$$

The last equality is obtained from (3.1). Denote $\xi_k = \left(\frac{k}{\beta}-1\right)\frac{1}{m-1}$. It is easy to see that

$$P_\theta(N=k) = \begin{cases} 0 & \text{if } \xi_k < 0 \\ F_{\widehat{\rho}_k}(\sqrt{\xi_k}) - F_{\widehat{\rho}_k}(-\sqrt{\xi_k}) & \text{if } \xi_k > 0 \text{ and } \sqrt{\xi_k} < \frac{1}{m-1} \\ F_{\widehat{\rho}_k}(\sqrt{\xi_k}) & \text{if } \xi_k > 0 \text{ and } \frac{1}{m-1} < \sqrt{\xi_k} < 1 \\ 1 & \text{if } \xi_k > 0 \text{ and } \sqrt{\xi_k} > 1. \end{cases}$$

Therefore, we have

$$P_\theta(N=k) = \begin{cases} 0 & \text{if } k \le \beta \\ F_{\widehat{\rho}_k}(\sqrt{\xi_k}) - F_{\widehat{\rho}_k}(-\sqrt{\xi_k}) & \text{if } \beta < k \le \frac{m\beta}{m-1} \\ F_{\widehat{\rho}_k}(\sqrt{\xi_k}) & \text{if } \frac{m\beta}{m-1} < k \le m\beta \\ 1 & \text{if } k > m\beta. \end{cases} \tag{4.1}$$

For $n > k$, the cumulative distribution function of N, denoted by F_N, is given by

$$F_N(n) = P_\theta(N \le n) = P_\theta(k \le n,\, K^* \le n) = P_\theta(K^* \le n).$$

Note that

$$P_\theta(K^* \le n) = P_\theta\left(\widehat{\rho}_k^2 \le \left(\frac{n}{\beta} - 1\right)\frac{1}{m-1}\right).$$

Following similar arguments as in (4.1), we have, for $n > k$,

$$F_N(n) = \begin{cases} 0 & \text{if } n < \beta \\ F_{\widehat{\rho}_k}(\sqrt{\xi_n}) - F_{\widehat{\rho}_k}(-\sqrt{\xi_n}) & \text{if } \beta \le n < \frac{m\beta}{m-1} \\ F_{\widehat{\rho}_k}(\sqrt{\xi_n}) & \text{if } \frac{m\beta}{m-1} \le n < m\beta \\ 1 & \text{if } n \ge m\beta. \end{cases} \tag{4.2}$$

Now, let us derive the exact distribution of $\widehat{\sigma}_N^2$. Define $W_{in} = V_{in}/\sigma^2$ and $W_{in}^* = V_{in}^*/\sigma^2$ for $i = 1, 2$. Let us denote the cumulative distribution functions of W_{1n} and W_{2n} by $F_{W_{1n}}$ and $F_{W_{2n}}$ respectively. Since $W_{1n} \sim (1 + (m-1)\rho)\,\chi_n^2$ and $W_{2n} \sim (1-\rho)\,\chi_{n(m-1)}^2$ independently,

$$F_{W_{1n}}(x) = F_{\chi_n^2}\left[\frac{x}{1+(m-1)\rho}\right] \quad \text{and} \quad F_{W_{2n}}(x) = F_{\chi_{n(m-1)}^2}\left[\frac{x}{1-\rho}\right],$$

where $F_{\chi_n^2}$ is the CDF of χ_n^2 distribution. The cumulative distribution function of $\widehat{\sigma}_N^2$, denoted as $F_{\widehat{\sigma}_N^2}$, is given by

$$\begin{aligned} F_{\widehat{\sigma}_N^2}(x) &= P_\theta(\widehat{\sigma}_N^2 \le x) = P_\theta(V_{1k} + V_{2k} + V_{1(N-k)}^* + V_{2(N-k)}^* \le mNx) \\ &= P_\theta\left(W_{1k} + W_{2k} + W_{1(N-k)}^* + W_{2(N-k)}^* \le \frac{mNx}{\sigma^2}\right) \\ &= \int_0^\infty \int_0^\infty P_\theta\left(W_{1j}^* + W_{2j}^* \le \frac{m(k+j)x}{\sigma^2} - w_1 - w_2\right) dF_{W_{2k}}(w_2)\; dF_{W_{1k}}(w_1). \end{aligned}$$

The last equality is obtained by conditioning on $(W_{1k}, W_{2k}) = (w_1, w_2)$ where j is the value of $(N-k)$ when $(W_{1k}, W_{2k}) = (w_1, w_2)$, that is,

$$j = j(w_1, w_2) = \max\left\{k, \left\lfloor \beta(\delta)\left(1 + (m-1)\left(\frac{w_1 - w_2/(m-1)}{w_1 + w_2}\right)^2\right)\right\rfloor + 1\right\} - k.$$

We also use independence of (W_{1k}, W_{2k}) and (W^*_{1j}, W^*_{2j}) to express the last equality. Next, we condition on $W^*_{2j} = w^*_{2j}$ and use the independence of W^*_{1j} and W^*_{2j} to derive the CDF of $\widehat{\sigma}^2_N$ as follows:

$$F_{\widehat{\sigma}^2_N}(x) = \int_0^\infty \int_0^\infty \int_0^{\frac{m(k+j)x}{\sigma^2} - w_1 - w_2} F_{W^*_{1j}}\left(\frac{m(k+j)x}{\sigma^2} - w_1 - w_2 - w^*_{2j}\right) \times dF_{W^*_{2j}}(w^*_{2j}) dF_{W_{2k}}(w_2)\ dF_{W_{1k}}(w_1). \tag{4.3}$$

Finally, the exact coverage probability yielded by $\left(\delta^{-1}\widehat{\sigma}^2_N,\ \delta\widehat{\sigma}^2_N\right)$ is

$$P_\theta\left(\frac{\widehat{\sigma}^2_N}{\delta} < \sigma^2 < \delta\widehat{\sigma}^2_N\right) = P_\theta\left(\frac{\sigma^2}{\delta} < \widehat{\sigma}^2_N < \delta\sigma^2\right) = F_{\widehat{\sigma}^2_N}(\delta\sigma^2) - F_{\widehat{\sigma}^2_N}\left(\sigma^2/\delta\right). \tag{4.4}$$

From (4.3) and (4.4), we remark that the coverage probability does not depend on the true value of σ^2. In Section 6, we will show that the exact coverage probability derived in (4.4) nearly equals the target $(1-\alpha)$.

5. CHOICE OF A PILOT SAMPLE SIZE AND ASYMPTOTIC RESULTS

Performance of any two-stage sampling methodology depends heavily upon proper selection of a pilot sample size. We will now discuss two possible choices, namely k_1 and k_2, of a pilot sample size. We summarize advantages or disadvantages, if any, of using either choice of a pilot sample size. We follow up with some desirable asymptotic optimality properties of our proposed two-stage methodology.

5.1. First Choice of a Pilot Sample Size

From the definition of the optimal fixed-sample-size n_0 in (2.5), we note that $n_0 \geq \beta(\delta)$. Therefore, one must collect at least $(\lfloor\beta(\delta)\rfloor + 1)$ samples to satisfy (2.4). This gives us a very natural choice of the pilot sample size

$$k_1 \equiv k_1(\delta) = \lfloor\beta(\delta)\rfloor + 1.$$

Next, we propose the first two-stage sampling approach with a pilot sample size k_1.

Stage I: Draw an initial random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k_1}$ from m-variate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\sigma^2 R$. Based on this pilot sample, obtain the MLE $\widehat{\rho}_{k_1}$ of ρ and evaluate

$$K^* \equiv K^*(\delta) = \lfloor\beta(\delta)\{1 + (m-1)\widehat{\rho}^2_{k_1}\}\rfloor + 1.$$

Since $K^* \geq k_1$ almost surely, the final sample size for this choice of pilot sample size becomes

$$N_1 = N_1(\delta) = \max\{k_1(\delta), K^*(\delta)\} = \lfloor\beta(\delta)\{1 + (m-1)\widehat{\rho}^2_{k_1}\}\rfloor + 1. \tag{5.1}$$

At this moment, one must proceed to the second stage and collect more samples.

Stage II: Draw $(N_1 - k_1)$ independent samples $\boldsymbol{X}_{k_1+1}, \ldots, \boldsymbol{X}_{N_1}$ from $N_m\left(\mathbf{0},\ \sigma^2 R\right)$ which are also independent of the initial k_1 samples. The point estimator of σ^2 is $\widehat{\sigma}^2_{N_1}$ and the fixed-accuracy confidence

interval for σ^2 is $\left(\delta^{-1}\widehat{\sigma}^2_{N_1},\, \delta\widehat{\sigma}^2_{N_1}\right)$.

Now, we present the asymptotic optimality properties enjoyed by the two-stage methodology for the first choice of the pilot sample size. First, we state and prove a lemma that would be required to establish the optimality results.

Lemma 5.1. *For the maximum likelihood estimator $\widehat{\rho}_n$ in* (2.2)*, there exists constants L and U such that*

$$-\frac{L}{n} + O(n^{-2}) \leq E_\theta\left[\widehat{\rho}^2_n - \rho^2\right] \leq \frac{U}{n} + O(n^{-2}).$$

Proof. Recall from the proof of Lemma 4.1 that

$$F_{n,\, n(m-1)} = \left[\frac{(1-\rho)(m-1)}{1+(m-1)\rho}\right]\frac{V_{1n}}{V_{2n}}.$$

Therefore, $\widehat{\rho}_n$ from (2.2) can be written as

$$\widehat{\rho}_n = \frac{\{1+(m-1)\rho\}F_{n,\, n(m-1)} - (1-\rho)}{\{1+(m-1)\rho\}F_{n,\, n(m-1)} + (m-1)(1-\rho)}.$$

Let us define $a := 1+(m-1)\rho$, $b := 1-\rho$, $c := (m-1)(1-\rho)$, and the function $h(x) := \left(\frac{ax-b}{ax+c}\right)^2$ for $x \geq 0$. Note that $h\left(F_{n,\, n(m-1)}\right) = \widehat{\rho}^2_n$ and $h(1) = \left(\frac{a-b}{a+c}\right)^2 = \rho^2$. The first two derivatives of $h(x)$ are

$$h'(x) = \frac{2a(b+c)(ax-b)}{(ax+c)^3} \quad \text{and} \quad h''(x) = \frac{2a^2(b+c)(3b+c-2ax)}{(ax+c)^4}. \tag{5.2}$$

By Taylor series expansion of $h\left(F_{n,\, n(m-1)}\right)$ and using (5.2), we have

$$\widehat{\rho}^2_n = \rho^2 - \frac{2a(b+c)(a-b)}{(a+c)^3}\left(F_{n,\, n(m-1)} - 1\right) + \frac{a^2(b+c)(3b+c-2a\zeta_n)}{(a\zeta_n+c)^4}\left(F_{n,\, n(m-1)} - 1\right)^2, \tag{5.3}$$

where ζ_n is a random variable between 1 and $F_{n,\, n(m-1)}$. Now, note that the first two raw moments of $F_{n,\, n(m-1)}$ distribution are

$$E_\theta\left[F_{n,\, n(m-1)}\right] = \frac{n(m-1)}{n(m-1)-2} = \left(1 - \frac{2}{n(m-1)}\right)^{-1} = 1 + \frac{2}{n(m-1)} + O(n^{-2}), \tag{5.4}$$

and

$$\begin{aligned} E_\theta\left[F^2_{n,\,n(m-1)}\right] &= E_\theta\left[\left(\chi^2_n/n\right)^2\right] E_\theta\left[\left(\chi^2_{n(m-1)}/n(m-1)\right)^{-2}\right] \\ &= \left[\frac{n^2+2n}{n^2}\right]\left[\frac{n^2(m-1)^2}{(n(m-1)-2)(n(m-1)-4)}\right] \\ &= \left[1+\frac{2}{n}\right]\left[1-\frac{2}{n(m-1)}\right]^{-1}\left[1-\frac{4}{n(m-1)}\right]^{-1} \\ &= 1+\frac{2}{n}\left(\frac{m+2}{m-1}\right)+O(n^{-2}). \end{aligned}$$

Therefore, we have

$$E_\theta\left[\left(F^2_{n,\,n(m-1)}-1\right)^2\right] = \frac{2m}{n(m-1)}+O(n^{-2}). \tag{5.5}$$

Taking expectation on both sides of (5.3) and using (5.4), we have

$$E_\theta\left[\widehat{\rho}^2_n\right] = \rho^2+\frac{4a(b+c)(a-b)}{n(m-1)(a+c)^3}+a^2(b+c)E_\theta\left[\frac{(3b+c-2a\zeta_n)}{(a\zeta_n+c)^4}\left(F_{n,\,n(m-1)}-1\right)^2\right]+O(n^{-2}). \tag{5.6}$$

Let $f(x) := (3b+c-2ax)/(ax+c)^4$. In order to find an upper and lower bound for $E_\theta\left[\widehat{\rho}^2_n\right]$, we will bound $E_\theta\left[f(\zeta_n)\left(F_{n,\,n(m-1)}-1\right)^2\right]$ from above and below. From the first derivative test, we conclude that $f(x)$ has a global minimum at $x=\frac{2b+c}{a}$ and the minimum value of $f(x)$ is $f\left(\frac{2b+c}{a}\right)=\frac{-1}{16(b+c)^3}$. Thus, using (5.5), we have a lower bound

$$E_\theta\left[f(\zeta_n)\left(F^2_{n,\,n(m-1)}-1\right)^2\right] \geq \frac{-1}{16(b+c)^3}\left[\frac{2m}{n(m-1)}\right]+O(n^{-2}). \tag{5.7}$$

Note that the global maximum of $f(x)$ is attained at $x=0$ the maximum value of $f(x)$ is $(3b+c)/c^4$. Therefore, using (5.5), we have an upper bound

$$E_\theta\left[f(\zeta_n)\left(F^2_{n,\,n(m-1)}-1\right)^2\right] \leq \frac{3b+c}{c^4}\left[\frac{2m}{n(m-1)}\right]+O(n^{-2}). \tag{5.8}$$

Finally, using (5.6), (5.7), and (5.8), we obtain

$$-\frac{L}{n}+O(n^{-2}) \leq E_\theta\left(\widehat{\rho}^2_n-\rho^2\right) \leq \frac{U}{n}+O(n^{-2}),$$

where

$$U=\frac{2ma^2(b+c)(3b+c)}{c^4(m-1)}+\frac{4a(b+c)(a-b)}{(a+c)^3(m-1)} \quad \text{and} \quad L=\frac{ma^2}{8(b+c)^2(m-1)}+\frac{4a(b+c)(b-a)}{(a+c)^2(m-1)}.$$

□

Next, we state and prove the asymptotic optimality properties of our two-stage sampling strategy based on k_1 pilot samples.

Theorem 5.1. *For the choice of pilot sample size as k_1, we have*

1. $\lim_{\delta\downarrow 1} \frac{N_1(\delta)}{n_0(\delta)} = 1$ *almost surely;*
2. $E_\theta\left[N_1(\delta) - n_0(\delta)\right] = O(1)$ *(Asymptotic second-order efficiency);*
3. $\lim_{\delta\downarrow 1} P_\theta\left(\frac{\widehat{\sigma}^2_{N_1(\delta)}}{\delta} < \sigma^2 < \delta\widehat{\sigma}^2_{N_1(\delta)}\right) = 1-\alpha$ *(Asymptotic consistency).*

Proof. 1. Recall that $N_1(\delta) = \lfloor \beta(\delta)\{1+(m-1)\widehat{\rho}^2_{k_1(\delta)}\}\rfloor + 1$ and $n_0(\delta) = \lfloor \beta(\delta)\{1+(m-1)\rho^2\}\rfloor + 1$. The proof of part 1 is immediate since $k_1(\delta)\to\infty$ as $\delta\downarrow 1$ and $\widehat{\rho}_n$ is a strongly consistent estimator of ρ.

2. Note that $\beta(\delta)\left\{1+(m-1)E_\theta\left[\widehat{\rho}^2_{k_1(\delta)}\right]\right\} \le E_\theta\left[N_1(\delta)\right] \le \beta(\delta)\left\{1+(m-1)E_\theta\left[\widehat{\rho}^2_{k_1(\delta)}\right]\right\} + 1$. Using Lemma 5.1, we can write

$$E_\theta\left[N_1(\delta)\right] \ge \beta(\delta)\left\{1+(m-1)\left[\rho^2 - \frac{L}{k_1(\delta)} + O\left(\frac{1}{k_1(\delta)^2}\right)\right]\right\} \ge n_0(\delta) - 1 - (m-1)L + O\left(\frac{1}{k_1(\delta)}\right).$$

Similarly, using Lemma 5.1, we have

$$E_\theta\left[N_1(\delta)\right] \le n_0(\delta) + (m-1)U + 1 + O\left(\frac{1}{k_1(\delta)}\right).$$

The above inequalities yield

$$-1 + (m-1)L + o(1) \le E_\theta\left[N_1(\delta) - n_0(\delta)\right] \le 1 + (m-1)U + o(1).$$

This completes the proof of part 2.

3. Let us define the random variables $U_i := \frac{1}{m}\sum_{j=1}^m Y_{ij}^2 - \sigma^2$ for $i = 1,\ldots,n$. It is not difficult to see that $\widehat{\sigma}^2_n = (V_{1n}+V_{2n})/mn$ is the sample mean of n independent random variables $U_1,\ldots,U_n$. That is,

$$\widehat{\sigma}^2_n - \sigma^2 = \frac{1}{n}\sum_{i=1}^n U_i, \text{ where } E_\theta(U_i) = 0 \text{ and } V_\theta(U_i) = \frac{2\sigma^4}{m}\left[1+(m-1)\rho^2\right], i = 1,\ldots,n.$$

Recall that $\ln(\delta)^2\beta(\delta) = 2z^2_{\alpha/2}/m$ and note

$$(\ln(\delta))^2\left[\beta(\delta)\{1+(m-1)\rho^2\}\right] \le (\ln(\delta))^2 n_0(\delta) \le (\ln(\delta))^2\left[\beta(\delta)\{1+(m-1)\rho^2\}+1\right].$$

This yields $\lim_{\delta\downarrow 1}(\ln(\delta))^2 n_0(\delta) = \frac{2z^2_{\alpha/2}}{m}\{1+(m-1)\rho^2\}$. Using part 1 of Theorem 5.1, we have

$$\lim_{\delta\downarrow 1}(\ln(\delta))^2 N_1(\delta) = \frac{2z^2_{\alpha/2}}{m}\{1+(m-1)\rho^2\} \in (0,\infty) \quad \text{almost surely.}$$

Therefore, by Anscombe's CLT (Anscombe 1952) for stopped random walks (see Gut 2009, p. 17), we can

write

$$\sqrt{N_1(\delta)}\left\{\widehat{\sigma}^2_{N_1(\delta)} - \sigma^2\right\} \quad \xrightarrow{\mathcal{L}} \quad N\left(0,\ \frac{2\sigma^4}{m}\left\{1+(m-1)\rho^2\right\}\right) \quad \text{as } \delta \downarrow 1.$$

See also Mukhopadhyay and Chattopadhyay (2012). Applying the delta method, we obtain

$$\sqrt{N_1(\delta)}\left\{\ln\left(\widehat{\sigma}^2_{N_1(\delta)}\right) - \ln\left(\sigma^2\right)\right\} \quad \xrightarrow{\mathcal{L}} \quad N\left(0,\ \frac{2}{m}\left\{1+(m-1)\rho^2\right\}\right) \quad \text{as } \delta \downarrow 1.$$

Note that (5.1) yields $\beta(\delta)\{1+(m-1)\rho^2\}/N_1(\delta) \to 1$ almost surely as $\delta \downarrow 1$. Therefore, by Slutsky's theorem

$$\sqrt{\beta(\delta)\{1+(m-1)\rho^2\}}\left\{\ln\left(\widehat{\sigma}^2_{N_1(\delta)}\right) - \ln\left(\sigma^2\right)\right\} \quad \xrightarrow{\mathcal{L}} \quad N\left(0,\ \frac{2}{m}\left\{1+(m-1)\rho^2\right\}\right) \quad \text{as } \delta \downarrow 1. \tag{5.9}$$

Using (5.9), we have

$$\begin{aligned}\lim_{\delta\downarrow 1} P_\theta\left(\frac{\widehat{\sigma}^2_{N_1(\delta)}}{\delta} < \sigma^2 < \delta\widehat{\sigma}^2_{N_1(\delta)}\right) &= 2\Phi\left(\frac{\sqrt{\beta(\delta)\{1+(m-1)\rho^2\}}\ln\delta}{\sqrt{\frac{2}{m}(1+(m-1)\rho^2)}}\right) - 1 \\ &= 2\Phi\left(z_{\alpha/2}\right) - 1 = 1-\alpha.\end{aligned}$$

Hence, the proof of Theorem 5.1 is complete. □

Remark 5.1. Even though the first choice of pilot sample size yields several desirable asymptotic properties, it may suffer from serious drawbacks in practice. The concepts of asymptotic consistency and asymptotic first- or second-order efficiency properties respectively come from Chow and Robbins (1965) and Ghosh and Mukhopadhyay (1981). Note that $k_1(\delta) = \lfloor 2z^2_{\alpha/2}/\left(m\ln(\delta)^2\right)\rfloor + 1$. Hence, for fixed α and δ, as the dimension m increases, the pilot sample size decreases. The numerical study illustrates that for large m, the value of k_1 may be too small to produce an accurate estimator of ρ. As a consequence, N_1 may be an inaccurate estimate of the optimal fixed-sample-size n_0, and hence, the performance of the two-stage methodology will be poor. The coverage probability produced by our two-stage methodology may be lower than the desired level $(1-\alpha)$ for any fixed α.

To avoid such poor performance (when m is large) due to a small pilot sample size, we propose another choice of a pilot sample size, k_2, in Subsection 5.2.

5.2. Second Choice of a Pilot Sample Size

In order to overcome the drawback of the first choice of a pilot sample size and suggest an alternative choice, we follow the idea of Mukhopadhyay (1980). Conceptually, consider an auxiliary sequential stopping variable

$$T = \inf\left\{n \geq 1:\ n \geq \beta(\delta)\left[1+(m-1)(\widehat{\rho}^2_n + n^{-r})\right]\right\},$$

where $r > 0$. Note that, in the definition of T, we have estimated ρ^2 by $\widehat{\rho}_n^2 + n^{-r}$. From the above definition of a stopping variable, we have

$$T \geq \frac{(m-1)\beta(\delta)}{T^r} \text{ a.s.}, \quad \text{that is, } T \geq \{(m-1)\beta(\delta)\}^{\frac{1}{1+r}} \text{ a.s.}$$

Thus, we will choose the pilot sample size to be

$$k_2 \equiv k_2(\delta) = \lfloor \{(m-1)\beta(\delta)\}^{\frac{1}{1+r}} \rfloor + 1.$$

With this choice of a pilot sample size, the corresponding two-stage final sample size is defined as

$$N_2 \equiv N_2(\delta) = \max\{k_2(\delta), K^*\}, \quad \text{where } K^* = \beta(\delta)\left\{1 + (m-1)\widehat{\rho}_{k_2(\delta)}^2\right\}. \tag{5.10}$$

Now, we establish some asymptotic optimality properties of the two-stage methodology involving k_2 as $\delta \downarrow 1$. First, we note that $k_2(\delta)$ is such that $k_2(\delta) \to \infty$ and $k_2(\delta)/n_0(\delta) \to 0$ as $\delta \downarrow 1$. Next, we again state and prove a theorem that describes three desirable asymptotic properties of our two-stage methodology.

Theorem 5.2. *The two-stage stopping variable $N_2(\delta)$ satisfies the following*

1. $\lim_{\delta\downarrow 1} \frac{N_2(\delta)}{n_0(\delta)} = 1$ *almost surely;*
2. $\lim_{\delta\downarrow 1} \frac{E_\theta\{N_2(\delta)\}}{n_0(\delta)} = 1$ *(Asymptotic first-order efficiency) ;*
3. $\lim_{\delta\downarrow 1} P_\theta\left(\frac{\widehat{\sigma}^2_{N_2(\delta)}}{\delta} < \sigma^2 < \delta\widehat{\sigma}^2_{N_2(\delta)}\right) = 1 - \alpha$ *(Asymptotic consistency) .*

Proof. 1. Note that $N_2(\delta) \to \infty$ a.s. as $\delta \downarrow 1$. Since $\widehat{\rho}_n$ is a strongly consistent estimator of ρ, the estimator $\widehat{\rho}_{k_2(\delta)} \to \rho$ almost surely as $\delta \downarrow 1$. Using the definition in (5.10), we have

$$\left\lfloor \beta(\delta)\left\{1 + (m-1)\widehat{\rho}_{k_2(\delta)}^2\right\} \right\rfloor + 1 \leq N_2(\delta) \leq k(\delta) + \beta(\delta)\left\{1 + (m-1)\widehat{\rho}_{k_2(\delta)}^2\right\} + 1. \tag{5.11}$$

Dividing all three parts of (5.11) by $n_0(\delta) = \lfloor \beta(\delta)\left\{1 + (m-1)\rho^2\right\} \rfloor + 1$ and taking limit as $\delta \downarrow 1$, we prove the first part of Theorem 5.2.

2. Since $N_2(\delta) \leq k(\delta) + \lfloor m\beta(\delta) \rfloor + 1 < \infty$ almost surely for all $\delta > 1$, we write

$$N_2(\delta)/n_0(\delta) \leq \frac{(m-1)\beta(\delta) + 1 + m\beta(\delta) + 1}{n_0(\delta)\beta(\delta)\{1 + (m-1)\rho^2\}} \leq \frac{2m\beta(\delta)}{\beta(\delta)} + \frac{2}{n_0(\delta)} \leq 2m + 2.$$

Therefore, by dominated convergence theorem and using first part of 5.2, we obtain

$$\lim_{\delta\downarrow 1} \frac{E_\theta\left\{N_2(\delta)\right\}}{n_0(\delta)} = E_\theta\left\{\lim_{\delta\downarrow 1} \frac{N_2(\delta)}{n_0(\delta)}\right\} = 1.$$

3. The proof of the third part of Theorem 5.2 is very similar to that of the part 3 of Theorem 5.1. Hence, we omit the proof. □

6. NUMERICAL STUDIES

In this numerical study, we investigate the following items:

(1) How to evaluate the *exact* distribution of stopping variable N defined in (3.2), the exact distribution of $\widehat{\sigma}^2_N$, and the exact coverage probability numerically.

(2) Address the exact computation of functionals of N and the CDF of $\widehat{\sigma}^2_N$ with appropriate numerical analysis and validate these with simulation studies.

(3) We may validate the results stated in Theorems 5.1 and 5.2 numerically. That is, we will show that the required expected sample size to achieve (2.4) is close to the optimal fixed-sample-size and the coverage probability associated fixed-accuracy confidence interval $\left(\delta^{-1}\widehat{\sigma}^2_N,\, \delta\widehat{\sigma}^2_N\right)$ is close to preset level $(1-\alpha)$.

We begin this study with numerical evaluation of the distribution functions of N and $\widehat{\sigma}^2_N$ as given in (4.2) and (4.3) respectively. Both distribution functions involve integrals which are computed using the "*integrate*" function from R which uses an adaptive Gauss–Kronrod quadrature method. We also simulate the distributions of N and $\widehat{\sigma}^2_N$ based on 10^5 replications. Note that the stopping variable or the final sample size N of the two-stage methodology is denoted by N_1 (or N_2) if k_1 (or k_2) is used as the pilot sample size.

Table 1 shows the pilot sample size k_1, the expected value and standard deviation of two-stage stopping variable N_1 for different choices of correlation ρ, dimension of data vector m, and δ. Here, $E(N_1)$ and $SD(N_1)$ represent *exact* values whereas $\widehat{E}(N_1)$ and $\widehat{SD}(N_1)$ represent the *simulated* values of expectation and standard deviation of the stopping variable. Moreover, Table 1 provides the exact and simulated coverage probabilities, denoted as CP and $\widehat{CP}$ respectively, associated with the two-stage methodology.

We observe that the exact values of expectation, standard deviation, and coverage probabilities are very similar to their simulated estimates which validates our numerical approximations of the exact formulas. We observe that, in most cases, the expected sample sizes are close to the optimal fixed-sample-size n_0 and the coverage probabilities are very close to the target 0.90.

Table 1 also provides numerical evidence in favor of the asymptotic results of Theorem 5.1. For fixed ρ and m, as δ decreases, both expected sample size and standard deviation of the stopping variable increase. For $\rho = 0.5$, as the dimension m increases from 5 to 20, the coverage probabilities seem to decrease and deviate from the target 0.90.

Recall Remark 5.1. Therefore, we feel the need of investigating this issue in more detail. We suspect that this may be due to the fact that the pilot sample size k_1 becomes smaller as m increases. In Table 2, we compare the performance of the two-stage methodology for two proposed choices of pilot sample sizes, k_1 and k_2. Based on results from our simulation study presented in Table 2, we observe that if the choice k_1 is used, the attained coverage probability gradually deceases as m increases from 25 to 150. When m is larger than 50, the coverage probabilities are much lower than the target 0.90. On the other hand, if k_2 is used with fixed $r = 0.5$, the attained coverage probabilities do not change much as m increases. With $r = 0.5$, the attained coverage probabilities remain close to the target 0.90. Table 2 illustrates that for large m, the difference between the expected sample size and the optimal fixed-sample-size is smaller if k_2 is

used. Moreover, the standard deviations of the stopping variable are also smaller when k_2 is used instead of k_1.

In Table 3, we evaluate the performance of the two-stage stopping variable N_2 for fixed $\delta = 1.1$, $\alpha = 0.10$ and a number of choices of correlation ρ, m, and the pilot sample size k_2 (or equivalently r). The exact values of expectation, standard deviation, and coverage probabilities are close to that of the simulated values which implies that the exact formulas developed in Section 4 are correct and the numerical approximations of the exact values are accurate. Moreover, the expected sample sizes required by the two-stage methodology are close to the optimal fixed-sample-size n_0 and the coverage probabilities are very close to the target 0.90. Therefore, the numerical results in Table 3 support the asymptotic results obtained in Theorem 5.2. We notice that for fixed m, if $|\rho|$ increases, the required sample size also increases which makes sense in the spirit of (2.5). For fixed ρ, as m increases, the required sample size decreases which can also be validated from (2.5).

Table 4 compares the expected values and standard deviations of N_2 and the attained coverage probabilities for different values of δ, α, and pilot sample size k_2 (or equivalently r). The expected sample sizes are close to the optimal fixed-sample-size and the coverage probabilities are very close to the desired level $(1 - \alpha)$ which again lend a hand to support the findings from Theorem 5.2. Table 4 illustrates that as δ becomes closer to 1 or α decreases, the expectation and the standard deviation of final sample size increases. From the cases we studied here, we can recommend that a value of $r = 0.3$ may be appropriate to use if $m = 3$ and $\rho = 0.1$.

The Figures 1, 2, 3, and 4 compare the empirical PMF of N_2 (based on simulated values of N_2) and the exact PMF of N_2. These pictures suggest that the empirical and the exact PMF agree with each other which validates the derived exact formulas. An interesting observation from Figures 1, 2, and 3 is that for $m = 2$, $\delta = 1.1$, and $\alpha = 0.10$, the distribution of N_2 is positively skewed, symmetric, but negatively skewed when $\rho = 0.3$, $\rho = 0.5$, and $\rho = 0.95$ respectively. This phenomenon has been observed for other combinations of m, δ, and α as well.

The Figures 5, 6, 7, and 8 compare the distribution function of the stopping variable N_2 for different combinations of ρ, m, δ, and α. Figure 5 illustrates that the CDF of N_2 for $\rho = 0$ stays higher than that of $\rho = \pm 0.4$ which implies that the stopping variable for $\rho = 0$ is stochastically smaller than the stopping variable for $\rho = \pm 0.4$.

This may not be very surprising because higher magnitude of common correlation between each components of data vector implies that each data vector becomes successively less informative. Hence, more data vectors are essential to estimate the optimal fixed-sample-size. Figure 6 shows that for a fixed choice of ρ, δ, and α, the stopping variable becomes stochastically larger as the dimension of data vector decreases from $m = 8$ to $m = 3$. This also makes sense to us because a higher value of m implies each data vector becomes successively more informative, and hence less data vectors are required to attain certain precision in estimating n_0. From Figures 7 and 8, we observe that the stopping variables become stochastically larger as δ gets closer to 1 or as α gets closer to 0.

Table 1. Comparison of the exact values and the simulated values of expectation, standard deviation, and coverage probability of the two-stage methodology using k_1 as the pilot sample size for $\alpha = 0.1$

ρ	m	δ	k_1	n^0	$E(N_1)$	$\widehat{E}(N_1)$	$SD(N_1)$	$\widehat{SD}(N_1)$	CP	$\widehat{CP}$
0.2	5	1.12	85	98	98.88	98.87	6.72	6.72	0.9010	0.9019
		1.10	120	139	139.33	139.37	7.98	8.00	0.9008	0.9010
		1.08	183	212	213.08	213.12	9.88	9.89	0.9005	0.9017
	20	1.12	22	38	38.39	38.41	9.62	9.66	0.8958	0.8962
		1.10	30	53	53.77	53.86	11.62	11.64	0.8969	0.8967
		1.08	46	81	81.83	81.76	14.35	14.40	0.8977	0.8973
0.5	5	1.12	85	169	168.74	168.82	17.12	17.12	0.8965	0.8978
		1.10	120	239	238.48	238.51	20.45	20.46	0.8973	0.8972
		1.08	183	366	365.64	365.70	25.48	25.53	0.8979	0.8965
	20	1.12	22	122	119.65	119.62	32.22	32.28	0.8786	0.8788
		1.10	30	172	169.66	169.74	38.20	38.25	0.8774	0.8863
		1.08	46	263	261.02	261.12	47.80	47.92	0.8793	0.8904

7. CONCLUDING REMARKS

This article develops a two-stage sampling methodology to construct a $100(1-\alpha)\%$ fixed-accuracy confidence interval in the sense of (1.1) for the common variance σ^2 in an equi-correlated multivariate normal distribution. The exact formulas for the distribution function of the final sample size N and $\widehat{\sigma}^2_N$ are derived, and it is shown that they can be computed accurately using some appropriate numerical integration techniques such as the "integrate" function in R.

The performance of the two-stage methodology depends heavily on a proper choice of a pilot sample size. We have proposed two different choices for a pilot sample size, namely k_1 and k_2. The first choice k_1 seems to perform well as long as the dimension m of the data vector is not too large. This choice is desirable in practice since Theorem 5.1 establishes second-order asymptotic optimality of the stopping variable.

Therefore, we recommend the use of k_1 as long as m is not sufficiently large (e.g., $m < 25$). On the other hand, if m is indeed large enough to make the pilot sample size k_1 too small, then we recommend the use of k_2 as the alternative choice of a pilot sample size. Clearly, this would require an appropriate selection of r. We recommend that the practitioner select r such that the initial sample size k_2 is not too small. Using the exact formulas from section 4, one may perform some numerical study assuming different values of ρ (for fixed α and δ) to select r appropriately.

Table 2. Comparing performances of the two-stage methodologies with pilot sample size k_1 against k_2 for $\alpha = 0.1$, and $\delta = 1.1$, and $r = 0.5$

ρ	m	n_0	two-stage methodology	$\widehat{E}(N)$	$\widehat{SD}(N)$	$\widehat{CP}$
0.4	25	116	using k_1	115.30	33.22	0.8790
			using k_2	115.63	19.88	0.8875
	50	106	using k_1	104.40	44.58	0.8571
			using k_2	105.77	19.14	0.8841
	100	101	using k_1	97.85	59.18	0.8185
			using k_2	100.90	18.49	0.8847
	150	99	using k_1	94.98	68.92	0.7849
			using k_2	99.25	18.26	0.8848
0.6	25	230	using k_1	225.10	48.73	0.8858
			using k_2	228.42	29.14	0.8922
	50	223	using k_1	212.61	66.87	0.8710
			using k_2	220.70	28.90	0.8918
	100	219	using k_1	199.80	89.38	0.8395
			using k_2	216.97	28.70	0.8921
	150	217	using k_1	191.70	104.18	0.8096
			using k_2	215.65	28.1	0.8917

ACKNOWLEDGMENT

We thank an Associate Editor and the reviewers for their helpful feedbacks.

REFERENCES

Anscombe, F. J. (1952). Large-Sample Theory of Sequential Estimation, *Proceedings of Cambridge Philosophical Society* 48: 600–607.

Banerjee, S. and Mukhopadhyay, N. (2014). A General Sequential Fixed-Accuracy Confidence Interval Estimation Methodology for a Positive Parameter: Illustrations Using Health and Safety Data, *submitted*.

Chow, Y. S. and Robbins, H. (1965). On the Asymptotic Theory of Fixed-Width Confidence Intervals for the Mean, *Annals of Mathematical Statistics* 36: 457–462.

Table 3. Comparison of the exact values and the simulated values of expectation, standard deviation, and coverage probability of the two-stage methodology using k_2 as pilot sample size for $\alpha = 0.1$, and $\delta = 1.1$

ρ	m	n_0	k_2	r	$E(N_2)$	$\widehat{E}(N_2)$	$SD(N_2)$	$\widehat{SD}(N_2)$	CP	$\widehat{CP}$
-0.05	3	200	20	1.00	205.73	205.75	8.41	8.47	0.9046	0.9047
			30	0.76	203.86	203.86	6.09	6.09	0.9033	0.9027
			60	0.465	201.95	201.96	3.66	3.68	0.9015	0.9019
-0.05	10	61	05	2.95	62.14	62.14	1.75	1.75	0.9031	0.9029
			10	1.75	61.76	61.76	1.32	1.33	0.9021	0.9025
			20	1.10	61.58	61.58	1.01	1.01	0.9015	0.8994
-0.05	15	42	05	2.95	41.68	41.68	0.7321	0.7320	0.9020	0.9025
			10	1.75	41.64	41.64	0.5648	0.5658	0.9019	0.9026
			20	1.12	41.64	41.64	0.4882	0.4881	0.9019	0.9009
0.1	3	203	20	1.00	210.09	210.20	15.22	15.13	0.9052	0.9059
			30	0.76	207.83	207.87	11.44	11.45	0.9038	0.9036
			60	0.465	205.45	205.51	7.35	7.45	0.9025	0.9024
0.1	10	65	05	2.95	70.88	70.88	16.94	16.95	0.9079	0.9075
			10	1.75	68.32	68.34	10.66	10.70	0.9053	0.9057
			20	1.10	66.92	66.93	6.91	6.95	0.9039	0.9031
0.1	15	46	05	2.95	49.52	49.53	13.84	13.89	0.9065	0.9059
			10	1.75	47.76	47.81	8.82	8.83	0.9049	0.9053
			20	1.12	46.81	46.81	5.79	5.81	0.9037	0.9026
0.5	3	298	20	1.00	298.73	298.97	48.08	48.23	0.8952	0.8952
			30	0.76	298.52	298.55	40.08	40.21	0.8963	0.8962
			60	0.465	298.56	298.46	28.77	29.08	0.8975	0.8977
0.5	10	194	20	1.12	191.51	191.45	46.93	47.09	0.8847	0.8864
			30	0.86	192.31	192.33	38.90	39.06	0.8883	0.8909
			60	0.54	193.17	193.25	27.92	27.94	0.8921	0.8932
0.5	15	179	20	1.12	176.30	176.30	46.41	46.58	0.8802	0.8823
			30	0.86	177.21	177.32	38.45	38.55	0.8861	0.8876
			60	0.55	178.17	178.19	27.60	27.56	0.8903	0.8904

Table 4. Comparison of the exact values and the simulated values of expectation, standard deviation, and coverage probability of the two-stage methodology using k_2 as pilot sample size for $\rho = 0.1$, and $m = 3$

δ	$1-\alpha$	n_0	k_2	r	$E(N_2)$	$\widehat{E}(N_2)$	$SD(N_2)$	$\widehat{SD}(N_2)$	CP	$\widehat{CP}$
1.15	0.9	95	27	0.60	97.17	97.20	5.70	5.73	0.9041	0.9044
			37	0.45	96.51	96.50	4.63	4.59	0.9031	0.9030
			56	0.30	95.90	95.86	3.57	3.51	0.9022	0.9023
1.15	0.95	134	33	0.60	137.17	137.22	7.05	7.09	0.9522	0.9529
			47	0.45	136.32	136.34	5.61	5.65	0.9516	0.9517
			73	0.30	135.60	135.61	4.27	4.30	0.9510	0.9514
1.15	0.99	231	46	0.60	235.09	235.06	9.88	9.82	0.9903	0.9897
			68	0.45	233.93	233.93	7.76	7.77	0.9901	0.9898
			111	0.30	232.98	232.98	5.82	5.83	0.9900	0.9899
1.05	0.9	773	98	0.60	779.09	779.04	20.91	20.89	0.9011	0.9007
			157	0.45	776.95	776.96	16.00	16.12	0.9006	0.8990
			280	0.30	775.37	775.44	11.69	11.78	0.9004	0.8995
1.05	0.95	1098	122	0.60	1104.39	1104.35	26.17	26.24	0.9506	0.9494
			199	0.45	1101.87	1101.93	19.94	20.05	0.9503	0.9492
			367	0.30	1100.03	1100.07	14.38	14.43	0.9501	0.9491
1.05	0.99	1896	171	0.60	1903.90	1904.09	37.42	37.71	0.9906	0.9901
			290	0.45	1900.59	1900.70	28.13	28.23	0.9904	0.9905
			558	0.30	1898.30	1898.29	20.01	19.97	0.9899	0.9901

De, S. K. (2014). Modified Three-Stage Sampling for Fixed-Width Interval Estimation of the Common Variance of Equi-Correlated Normal Distributions, *Sequential Analysis* 33: 87–111.

Ghezzi, D. J. and Zacks, S. (2005). Inference on the Common Variance of Correlated Normal Random Variables, *Communications in Statistics – Theory and Methods* 34: 1517–1531.

Ghosh, B. K. and Sen, P. K. (1991). *Handbook of Sequential Analysis, edited volume*, New York: Dekker.

Ghosh, M. and Mukhopadhyay, N. (1981). Consistency and Asymptotic Efficiency of Two-Stage and Sequential Procedures, *Sankhya, Series A* 43: 220–227.

Ghosh, M., Mukhopadhyay, N., and Sen, P. K. (1997). *Sequential Estimation*, New York: Wiley.

Gut, A. (2009). *Stopped Random Walks: Limit Theorems and Applications*, 2^{nd} edition, New York: Springer.

Haner, D. M. (2012). On Two-Stage and Three-Stage Sampling Procedures for Fixed-Width Interval Estimation of the Common Variance of Correlated Normal Distributions, *Ph.D. Dissertation*, Binghamton University, New York, USA.

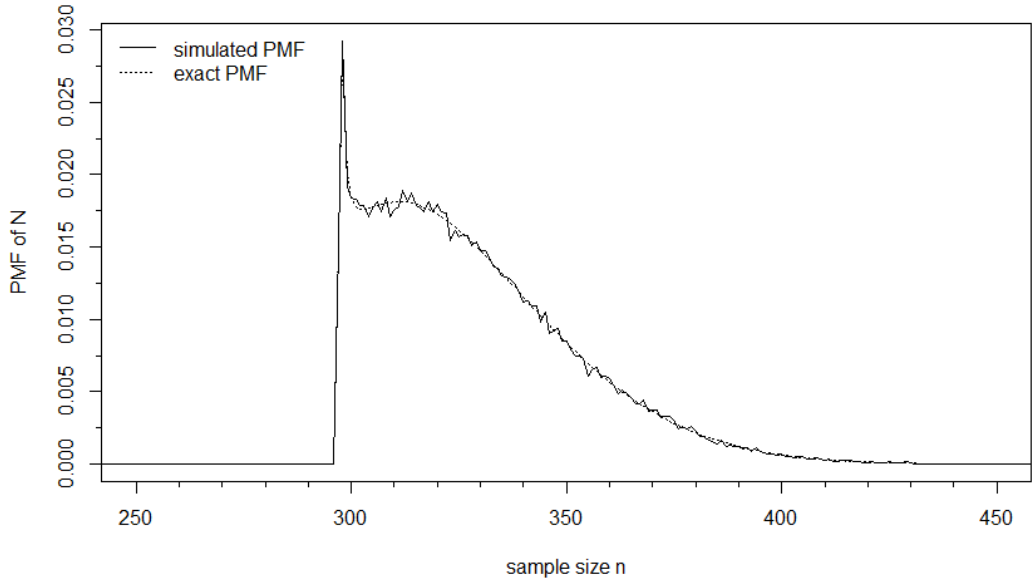

Figure 1. Comparison of empirical PMF and exact PMF of two-stage stopping variable N_2 for $\rho = 0.3$, $m = 2$, $r = 0.5$, $\alpha = 0.1$, and $\delta = 1.1$.

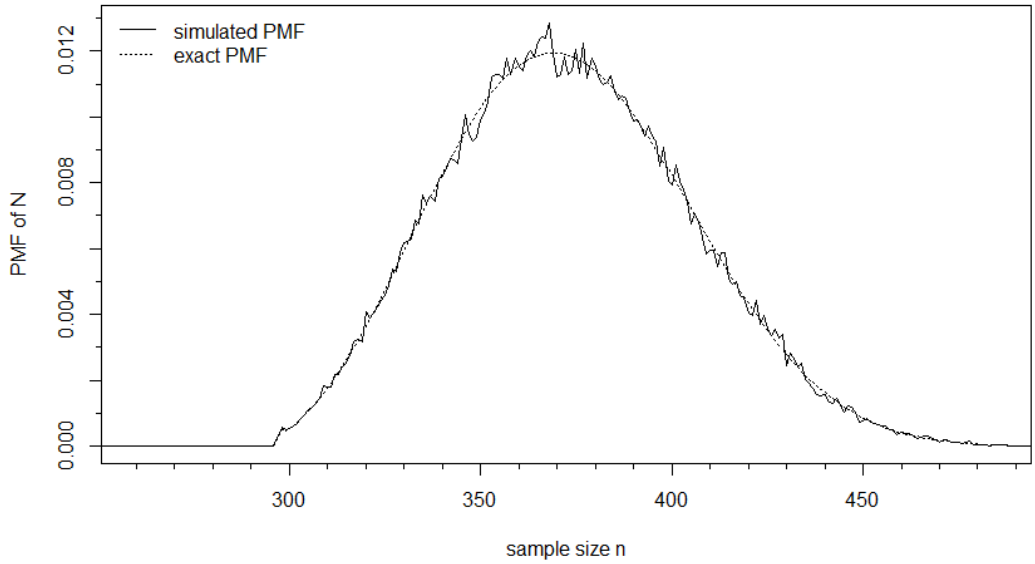

Figure 2. Comparison of empirical PMF and exact PMF of two-stage stopping variable N_2 for $\rho = 0.5$, $m = 2$, $r = 0.5$, $\alpha = 0.1$, and $\delta = 1.1$

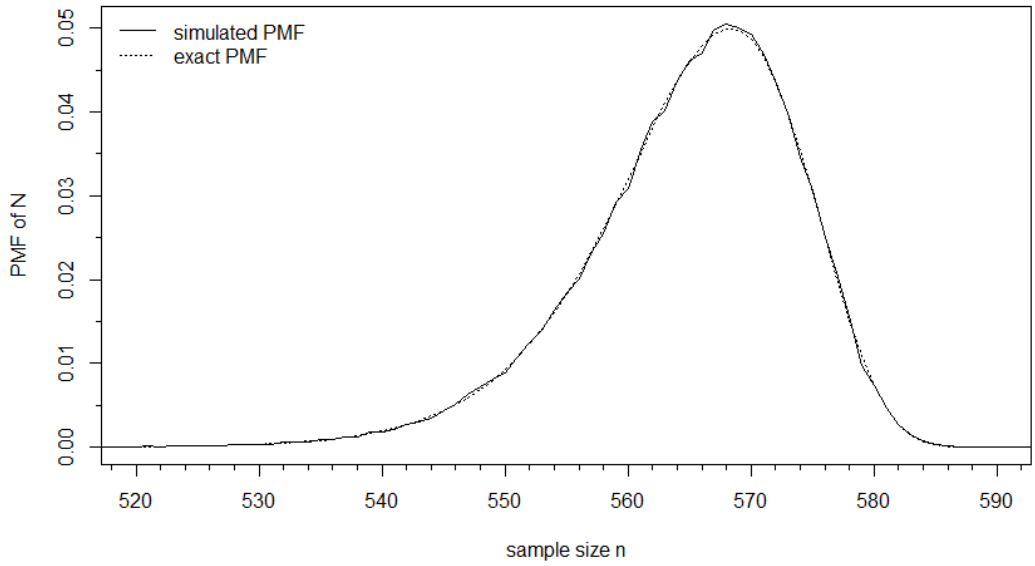

Figure 3. Comparison of empirical PMF and exact PMF of two-stage stopping variable N_2 for $\rho = 0.95$, $m = 2$, $r = 0.5$, $\alpha = 0.1$, and $\delta = 1.1$.

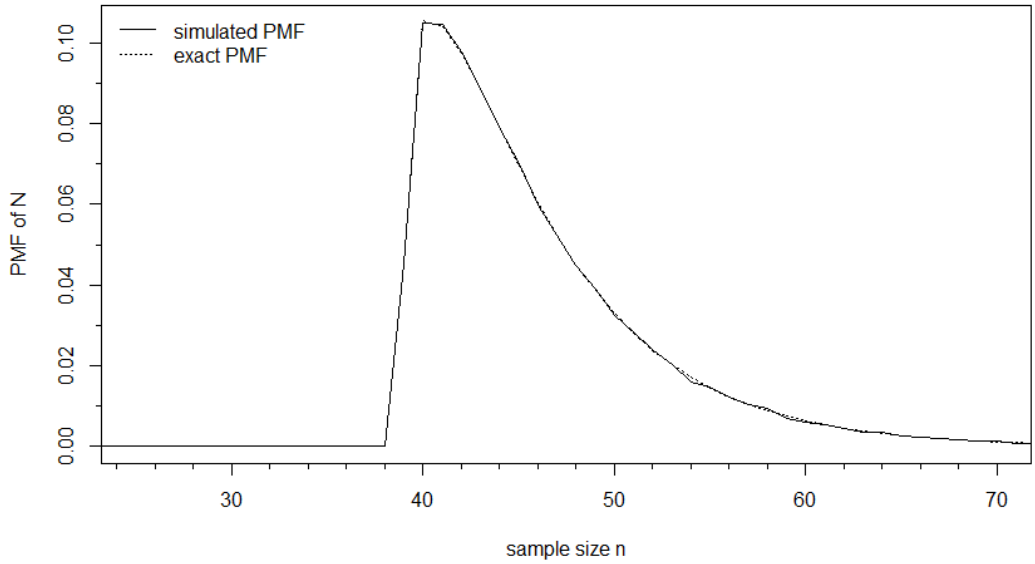

Figure 4. Comparison of empirical PMF and exact PMF of two-stage stopping variable N_2 for $\rho = 0.1$, $m = 15$, $r = 1.2$, $\alpha = 0.1$, and $\delta = 1.1$.

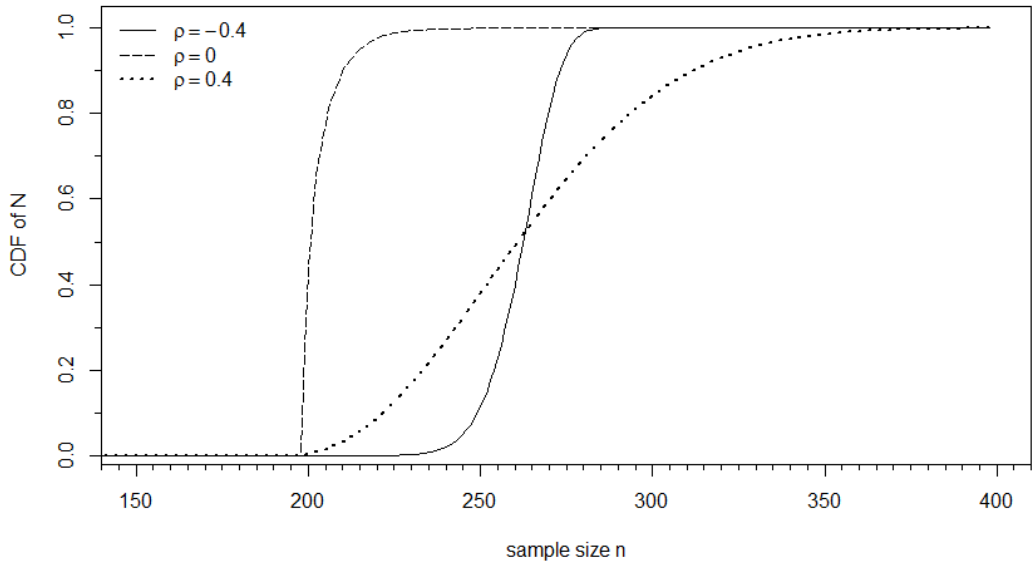

Figure 5. Comparison of the CDFs of two-stage stopping variable N_2 for different values of ρ using $k = 30$, $\alpha = 0.1$, $\delta = 1.1$, and $m = 3$.

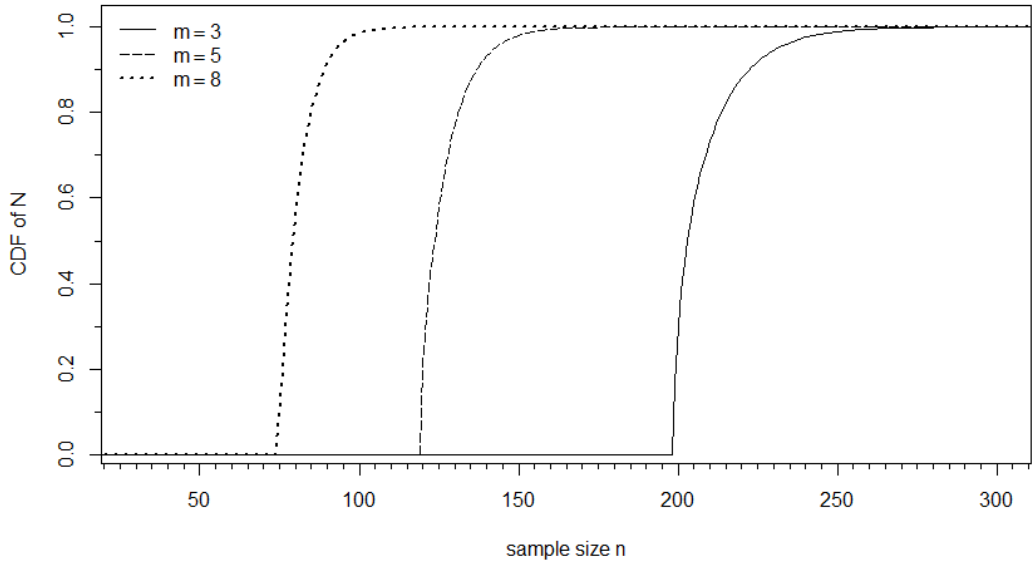

Figure 6. Comparison of the CDFs of two-stage stopping variable N_2 for different values of m using $k = 30$, $\alpha = 0.1$, $\delta = 1.1$, and $\rho = 0.1$.

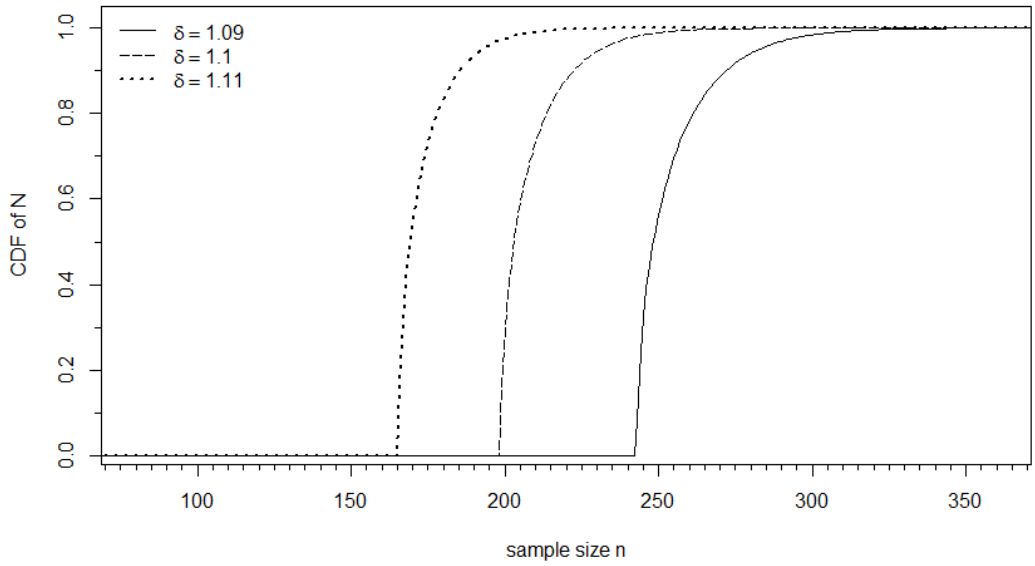

Figure 7. Comparison of the CDFs of two-stage stopping variable N_2 for different values of δ using $k = 30$, $\alpha = 0.1$, $m = 3$, and $\rho = 0.1$.

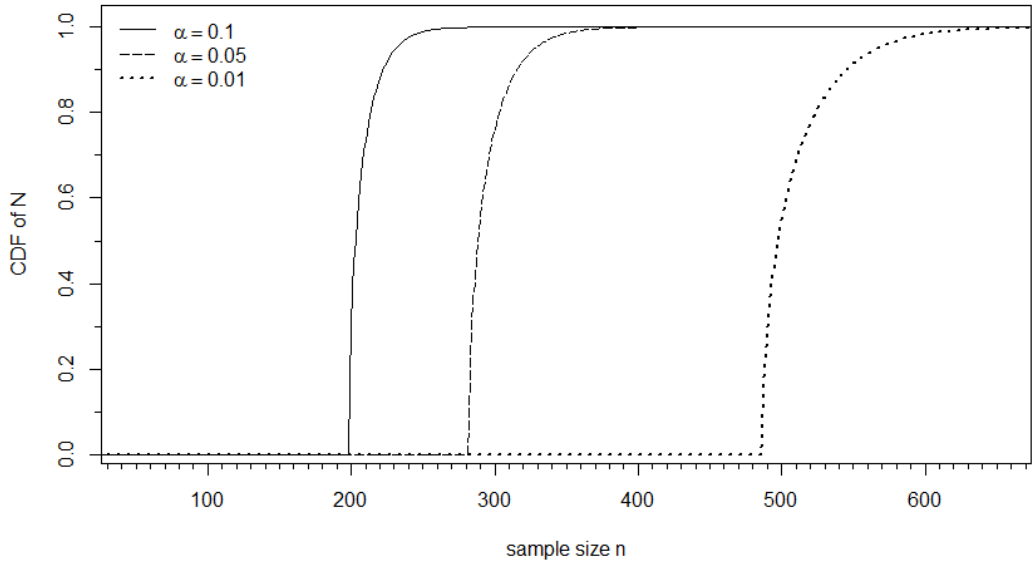

Figure 8. Comparison of the CDFs of two-stage stopping variable N_2 for different values of α using $k = 30$, $\delta = 1.1$, $m = 3$, and $\rho = 0.1$.

Haner, D. M. and Zacks, S. (2013). On Two-Stage Sampling for Fixed-Width Interval Estimation of the Common Variance of Equi-correlated Normal Distributions, *Sequential Analysis* 32: 1–13.

Khan, R. A. (1969). A General Method of Determining Fixed-Width Confidence Intervals, *Annals of Mathematical Statistics* 40: 704–709.

Lehmann, E. L. (1997). *Testing Statistical Hypotheses*, 2^{nd} edition, New York: Springer.

Mahalanobis, P. C. (1940). A Sample Survey of Acreage Under Jute in Bengal, with Discussion on Planning of Experiments, *Proceedings of Second Indian Statistical Conference*, Calcutta, India: Statistical Publishing Society.

Mukhopadhyay, N. (1980). A Consistent and Asymptotically Efficient Two-Stage Procedure to Construct Fixed Width Confidence Intervals for the Mean, *Metrika* 27: 281–284.

Mukhopadhyay, N. (1982). Stein's Two-Stage Procedure and Exact Consistency, *Scandinavian Actuarial Journal* 1982: 110–122.

Mukhopadhyay, N. (2000). *Probability and Statistical Inference*, New York: Dekker.

Mukhopadhyay, N. and Banerjee, S. (2014). Purely Sequential and Two-Stage Fixed-Accuracy Confidence Interval Estimation Methods for Count Data from Negative Binomial Distributions in Statistical Ecology: One-Sample and Two-Sample Problems, *Sequential Analysis* 33: 251–285.

Mukhopadhyay, N. and Chattopadhyay, B. (2012). A Tribute to Frank Anscombe and Random Central Limit Theorem from 1952, *Sequential Analysis* 31: 265–277.

Mukhopadhyay, N. and de Silva, B. M. (2009). *Sequential Methods and Their Applications*, Boca Raton: CRC.

Nadas, A. (1969). An Extension of a Theorem of Chow and Robbins on Sequential Confidence Intervals for the Mean, *Annals of Mathematical Statistics* 40: 667–671.

Rao, C. R. (1973). *Linear Statistical Inference*, 2^{nd} edition, New York: Wiley.

Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*, New York: Springer.

Stein, C. (1945). A Two Sample Test for a Linear Hypothesis Whose Power Is Independent of the Variance, *Annals of Mathematical Statistics* 16: 243–258.

Stein, C. (1949). Some Problems in Sequential Estimation, *Econometrica* 17: 77–78.

Willson, L. J. and Folks, J. L. (1983). Sequential Estimation of the Mean of the Negative Binomial Distribution, *Communications in Statistics-Sequential Analysis* 2: 55–70.

Zacks, S. (1966). Sequential Estimation of the Mean of a Log-Normal Distribution Having a Prescribed Proportional Closeness, *Annals of Mathematical Statistics* 37: 1439–1888.

Zacks, S. (2009a). *Stage-Wise Adaptive Designs*, New York: Wiley.

Zacks, S. (2009b). The Exact Distributions of the Stopping Times and Their Functionals in Two-Stage and Sequential Fixed-Width Confidence Intervals of the Exponential Parameter, *Sequential Analysis* 28: 69–81.

Zacks, S. and Khan, A. (2011). Two-Stage and Sequential Estimation of the Scale Parameter of a Gamma Distribution with Fixed-Width Intervals, *Sequential Analysis* 30: 297–307.

Zacks, S. and Mukhopadhyay, N. (2006a). Exact Risks of Sequential Point Estimators of the Exponential Parameter, *Sequential Analysis* 25: 203–220.

Zacks, S. and Mukhopadhyay, N. (2006b). Bounded Risk Estimation of the Exponential Parameter in a Two-Stage Sampling, *Sequential Analysis* 25: 437–452.

Zacks, S. and Mukhopadhyay, N. (2007). Distributions of Sequential and Two-Stage Stopping Times for Fixed-Width Confidence Intervals in Bernoulli Trials: Application in Reliability, *Sequential Analysis* 26: 425–441.

Zacks, S. and Ramig, P. F. (1987). Confidence Intervals for the Common Variance of Equicorrelated Normal Random Variables, in *Contributions to the Theory and Applications of Statistics*, volume in honor of Herbert Solomon, A. E. Gelfand, ed., pp. 511–544, New York: Academic Press.