Two-Stage and Sequential Estimation of Parameter N of Binomial Distribution When p Is Known

Shyamal K. De[1] **and Shelemyahu Zacks** [2]

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

[2]Department of Mathematical Sciences, Binghamton University, Binghamton, NY 13902

Abstract: In this article, we propose two-stage and purely sequential procedures to construct *bounded width* and *prescribed proportional closeness* confidence intervals for the unknown parameter N of $B(N, p)$ distribution where the parameter p is assumed to be known. The exact distributions of the stopping variables and the estimators of N at stopping are derived for all cases. The coverage probabilities of the proposed interval estimator are computed exactly and are shown to be nearly the same as the prescribed level.

Keywords: Exact computations; Bounded width; Proportional closeness interval; Sequential sampling; Stopping variable; Two-stage sampling.

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

1. INTRODUCTION

The problem of estimating the parameters of a binomial distribution is a classical problem that has been studied in many research articles. Some of the articles are those of Haldane (1941), Feldman and Fox (1968), McCabe (1973), Bluementhal and Dehiya (1981), Olkin, Petkau, and Zidek (1981), Raftery (1988), Hall (1994), and DasGupta and Rubin (2005). More specifically, suppose $X \sim B(N, p)$ is a Binomially distributed random variable (number of successes among N trials), when the parameters N and p are unknown. Estimation of N and p is an important inference problem that arises in ecology, software reliability, and other problems of estimating the size of a population. Most of the papers mentioned above deal with the properties of estimators when both parameters (N, p) are unknown. This is a difficult problem since, when both parameters are unknown, there is no joint maximum likelihood estimator of both N and p. Various modified estimators of N and p were suggested in the literature in order to overcome this problem. DasGupta and Rubin (2005) study the common estimators and propose a new one which is more efficient. All estimators of N and p suggested in the above references are based on a random sample $\{X_1, \ldots, X_m\}$ from

Address correspondence to Shyamal K. De, School of Mathematical Sciences, National Institute of Science Education and Research, P.O. Bhimpur-Padanpur, Khordha, Odisha 752050, India; Tel: +91 (674) 249-4083; Fax: +91 (674) 249-4083; E-mail: sde@niser.ac.in

the same distribution. The problem that we study in the present paper is how large should m be in order to achieve a prescribed precision of the estimators. To the best of our knowledge, this problem has not been studied before. The likelihood function of (N, p), given a sample of m observations, is

$$L(N, p \mid X_1, \ldots, X_m) = p^{S_m}(1-p)^{mN-S_m} \prod_{i=1}^{m} \binom{N}{X_i} \tag{1.1}$$

where $S_m := \sum_{i=1}^{m} X_i$. Accordingly, the minimal sufficient statistics, when N is unknown, is the order statistics $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(m)}$. Thus, for $m \geq 2$, there is no uniformly most efficient estimator of N. In order to shed more light on the precision problem, we simplify the model and assume that p is known.

If $p = 1$, then $P(X = N) = 1$, and it is sufficient to have $m = 1$. Therefore, we restrict our attention to the case when $p < 1$. For our study, it is analytically convenient to use the moment estimator of N, given by $\overline{N} := S_m/mp$. Clearly,

$$E_{N,p}[\overline{N}] = N \quad \text{and} \quad V_{N,p}[\overline{N}] = \frac{N(1-p)}{mp}.$$

If N is bounded by a known N^*, one can determine m to make $V_{N,p}[\overline{N}]$ as small as required. The present paper deals with the problem, when such an upper bound is not available.

The sample maximum, $X_{(m)}$, is a strongly consistent estimator of N, that is, $\lim_{m\to\infty} X_{(m)} = N$ with probability 1. However, as will be shown, the sample maximum is a very inefficient estimator of N. We know that $X_{(m)} \leq N$ with probability 1. It is interesting to study under what conditions on N, m and p, our moment estimator $\overline{N} < X_{(m)}$. In Table 1, we present estimates of $P(X_{(m)} > \overline{N})$ based on 10,000 simulation runs in each case. We observe that the estimates of $P(X_{(m)} > \overline{N})$ are very small even when N is relatively small and p is relatively close to 1. We also investigate the efficiency of $X_{(m)}$ as an estimator of N and present the expected values and the mean squared errors (MSE) of $\overline{N}$ and $X_{(m)}$. The expected value and MSE of $X_{(m)}$ are

$$E_{N,p}[X_{(m)}] = N - \sum_{j=0}^{N-1} \{B(j; N, p)\}^m$$

and

$$MSE_{N,p}[X_{(m)}] = 2N \sum_{j=0}^{N-1} \{B(j; N, p)\}^m - \sum_{j=0}^{N-1} j\{B(j; N, p)\}^m,$$

where $B(j; N, p)$ is the cumulative distribution function of $B(N, p)$. We present in Table 2 the exact values of expectation and MSE of $\overline{N}$ and $X_{(m)}$ and observe that $X_{(m)}$ is very inefficient for small values of p.

McCabe (1973) studied a similar problem, in somewhat more general terms. He considered a doubly indexed set of i.i.d. random variables, $\{X_{ij} : i = 1, \ldots, n;\ j = 1, 2, \ldots, k\}$, satisfying that $E[X_{ij}] = \mu \neq 0$, and $V\{X_{ij}\} = \sigma^2$ where $0 < \sigma^2 < \infty$. The observed random variables are the sums $Y_j = \sum_{i=1}^{n} X_{ij}$, where the number of terms n in each sum is the same but it is unknown. In order to estimate n, the number of replicas, k, has to be determined to obtain certain properties of the estimators of n. If $X_{ij} \stackrel{d}{=} B(1, p)$ then $Y_j \stackrel{d}{=} B(n, p)$. This special case of estimating n is similar to the problem studied in this paper. However,

there are methodological differences between the work of McCabe and our work. McCabe (1973) studied the question of how large k should be, by employing robust inequalities, which guarantee that estimation errors can decrease fast as n is large. There is no numerical study of the effectiveness of these inequalities for small samples. Our study employs a different approach. We derive the exact distributions of the random sample size, K, required in a two-stage or sequential sampling in order to attain fixed-width confidence intervals for the unknown parameter N. Our results are valid for all N not just for large ones.

In the present paper, we restrict our attention to the case of known p and study the problem of obtaining a set estimator of N with a prescribed precision. In other words, we study the required number of repetitions, M, which guarantee a bounded width confidence interval for N or a prescribed proportional closeness interval of N. To obtain a bounded width confidence interval for N one has to resort to two-stage or sequential procedures. In Section 2, we develop formulas for the exact computation of the distribution of M and the related estimators of N at stopping, in two-stage procedures. In Section 3, we study the distribution of M in sequential procedures. If the true value of N is 25, for example, one would like to require that the width of the interval estimator will not exceed $2\delta = 2$. The expected value of M depends on p. In the case of $p = 0.1$, the required number of repetitions should be about 609 for the coverage probability of the interval estimator to be about 0.9. Moreover, if $N = 100$ and $\delta = 1$, the required number of repetitions should be about 2435 under similar requirements. For many applications these demands are too prohibiting. In order to alleviate the situation, we introduce the proportional closeness requirement in Section 5. We are satisfied with $\delta = 0.1N$. The number of required repetitions to satisfy proportional closeness criteria is reduced significantly when N is large.

Note that the estimator $\overline{N}$ need not belong to the parameter space of N, that is, $\{1, 2, 3, \ldots\}$. Thus, we use a slightly modified moment estimator. Based on m i.i.d. observations, $X_1, \ldots, X_m$ from $B(N, p)$ distribution, a natural estimator of N, assuming p is known, is the moment type estimator given by

$$\widehat{N}_m := \max\left\{1, \left\lfloor \frac{1}{mp}\sum_{i=1}^{m} X_i \right\rfloor + 1\right\} = \max\left\{1, \left\lfloor \frac{S_m}{mp} \right\rfloor + 1\right\}, \tag{1.2}$$

where the notation $\lfloor x \rfloor$ represents the largest integer less than x. In this way, yields the value 1 even if $S_m = 0$. If m is sufficiently large, $P(S_m = 0)$ is negligible even if $N = 1$ and p is not too small. Note that $\widehat{N}_m$ defined in this way yields a slightly biased estimator and also that $\widehat{N}_m \in \{1, 2, \ldots\}$ with probability one. We would like to construct a confidence interval $[L, U]$ for N such that $1 \leq L \leq U$ with probability 1. Since N can only be positive integers $P(N \in [L, U]) = P(N \in A)$ where $A = \{L, L + 1, \ldots, U\}$. The cardinality of A, denoted as $|A|$, is given by $U - L + 1$. We are interested to construct a bounded width confidence interval $[L, U]$, or equivalently, a bounded cardinality confidence set A.

In this work, we consider

$$L = \max\{1, \widehat{N}_m - \delta\} \quad \text{and} \quad U = \widehat{N}_m + \delta - 1$$

for a prefixed $\delta \in \{1, 2, \ldots\}$.

Note that L and U are functions of sample size m and we must choose m such that

$$P_N\left(L \leq N \leq U\right) \geq 1-\alpha \quad \text{for all } N = 1, 2, \ldots. \tag{1.3}$$

The confidence interval $[L, U]$ is actually a *bounded width* confidence interval with an upper bound of 2δ. Since N is unbounded one cannot construct a bounded width confidence interval for N on the basis of a single sample. One needs at least two-stage sampling procedure. Indeed, a confidence interval at level $(1-\alpha)$ must have a coverage probability of at least $(1-\alpha)$ for all N. Note that

$$\sqrt{m}\left(\frac{S_m}{mp} - N\right) \quad \xrightarrow{\mathcal{L}} \quad N\left(0, \ \frac{Nq}{p}\right) \quad \text{as } m \to \infty,$$

where $q = 1-p$. For large m and prefixed δ, we can write

$$\begin{aligned} P\left(\left\lfloor\frac{S_m}{mp}\right\rfloor + 1 - \delta \leq N \leq \left\lfloor\frac{S_m}{mp}\right\rfloor + \delta\right) &= P\left(\left\lfloor\frac{S_m}{mp}\right\rfloor \geq N - \delta\right) - P\left(\left\lfloor\frac{S_m}{mp}\right\rfloor \geq N + \delta\right) \\ &= P\left(\frac{S_m}{mp} \leq N + \delta\right) - P\left(\frac{S_m}{mp} \leq N - \delta\right) \\ &\approx \Phi\left(\frac{\delta\sqrt{m}}{\sqrt{Nq/p}}\right) - \Phi\left(\frac{-\delta\sqrt{m}}{\sqrt{Nq/p}}\right) = 2\Phi\left(\frac{\delta\sqrt{m}}{\sqrt{Nq/p}}\right) - 1. \end{aligned}$$

In order to obtain $(1-\alpha)$ coverage probability, the sample size must satisfy

$$m \geq \frac{\chi^2_{1-\alpha} Nq}{\delta^2 p},$$

where $\chi^2_{1-\alpha}$ is the $(1-\alpha)$-quantile of a chi-square distribution with 1 degree of freedom. Therefore, the minimum or optimal sample size to satisfy (1.3) is given by

$$n_0 \equiv n_0(\delta) := \left\lfloor \frac{\chi^2_{1-\alpha} Nq}{\delta^2 p} \right\rfloor + 1. \tag{1.4}$$

Since the true value of N is not known in (1.4), one must estimate N and n_0 based on a pilot sample of size k (prefixed) in the first stage. If $\widehat{n}_0$ is the estimate of n_0, then in the second stage (if necessary) additional $(\widehat{n}_0 - k)$ samples are drawn and N is estimated based on $\widehat{n}_0$ number of samples. Below, we describe the two-stage sampling strategy for bounded width confidence interval estimation of N.

There is extensive literature on two-stage and sequential sampling for attaining bounded length confidence intervals. The two-stage sampling method presented in Section 2, is called a "stein-like procedure". A recent paper reviewing Stein-like procedures is that of Zacks (2015). In particular, we mention the papers of De (2014), De and Mukhopadhyay (2015), Mukhopadhyay (2005), Mukhopadhyay et al. (2010), Stein (1945), Zacks and Mukhopadhyay (2007a, 2007b), and the book of Zacks (2009). Sequential procedures for attaining proportional closeness in reliability estimation were derived also in Zacks (2001).

2. TWO-STAGE SAMPLING

In this section, we present a two-stage sampling strategy to construct a bounded width confidence interval of N using a fixed pilot sample size k.

Stage I: Fix the pilot sample size k and draw an initial random sample $X_1, \ldots, X_k$ from $B(N, p)$ distribution. Estimate N by S_k/kp and obtain the estimated sample size

$$K^* \equiv K^*(\delta) = \left\lfloor \frac{\chi^2_{1-\alpha} S_k q}{k\delta^2 p^2} \right\rfloor + 1. \tag{2.1}$$

Suppose M is the stopping variable, that is, the final sample size for this two-stage procedure. If $K^* \leq k$, we stop sampling and set $M = k$ and $\widehat{N}_M = \widehat{N}_k$. Otherwise, proceed to Stage II.

Stage II: If $K^* > k$, we draw remaining $(K^* - k)$ random samples from $B(N, p)$ independent of the initial k samples and label the observations as $X^*_{k+1}, \ldots, X^*_{K^*}$. Set $M = K^*$. Define $S_k := \sum_{i=1}^{k} X_i$ and $S^*_{K^*-k} := I(K^* > k) \sum_{i=k+1}^{K^*} X^*_i$. The final estimator of N is

$$\widehat{N}_M := \left\lfloor \frac{1}{Mp} \left(S_k + S^*_{M-k} \right) \right\rfloor + 1,$$

where the final sample size is $M \equiv M(\delta) = \max\{k, K^*(\delta)\}$. The bounded width confidence interval for N is

$$\left[L_M \, , \widehat{N}_M + \delta \right) \quad \text{where } \; L_M := \max\left\{1, \widehat{N}_M - \delta\right\}.$$

3. EXACT DISTRIBUTION OF M AND $\widehat{N}_M$

Let us denote the cumulative distribution function (c.d.f.) of a $B(N, p)$ random variable by $B(j; N, p)$ and the probability mass function (p.m.f.) by $b(j; N, p)$. The CDF of M, denoted as F_M, is given by

$$\begin{aligned} F_M(m) &= P(M \leq m) = P\left(K^* \leq m \, , k \leq m\right) \\ &= P\left(\frac{\chi^2_{1-\alpha} S_k q}{k\delta^2 p^2} \leq m, \; k \leq m \right) = I(m \geq k) B\left(\frac{m}{\xi}; kN, p \right), \end{aligned} \tag{3.1}$$

where $\xi := (q\chi^2_{1-\alpha}/kp^2\delta^2)$. Clearly, $F_M(m) = 0$ for $m < k$. Therefore, the probability mass function of M is $P(M = m) = F_M(m) - F_M(m-1)$ for $m = k, k+1, \ldots$. Using this probability mass function, we also compute $E(M)$ and $SD(M)$ which are presented in Table 3.

Next, we derive the exact distribution of $\widehat{N}_M$. Suppose for given $S_k = s$, the stopping variable $M =$

$m(s) = \max\{k, \lfloor \xi s \rfloor + 1\}$. The cdf of $\widehat{N}_M$, denoted by $F_{\widehat{N}_M}$, is given as

$$\begin{aligned} F_{\widehat{N}_M}(n) &= P\left(\widehat{N}_M \le n\right) = P\left(S_k + S^*_{M-k} \le npM\right) \\ &= \sum_{s=0}^{Nk} P\left(s + S^*_{m(s)-k} \le npm(s) | S_k = s\right) P(S_k = s) \\ &= \sum_{s=0}^{Nk} B\left(npm(s) - s; (m(s)-k)N, p\right) b(s; Nk, p), \end{aligned} \tag{3.2}$$

where $n \in \{1, 2, \ldots\}$. Based on (3.2), we can compute the coverage probability CP_{TS} as follows

$$CP_{TS} = P\left(\max\{1, \widehat{N}_M - \delta\} \le N < \widehat{N}_M + \delta\right) = F_{\widehat{N}_M}(N + \delta) - F_{\widehat{N}_M}(N - \delta). \tag{3.3}$$

From the above cumulative distribution function, we compute the probability mass function of $\widehat{N}_M$ as

$$p_{\widehat{N}_M}(n) = P(\widehat{N}_M = n) = F_{\widehat{N}_M}(n) - F_{\widehat{N}_M}(n-1) \quad \text{for } n = 2, 3, \ldots,$$

and $p_{\widehat{N}_M}(1) = P(\widehat{N}_M = 1) = F_{\widehat{N}_M}(1)$. The expectation and standard deviation of $\widehat{N}_M$ is also computed using $p_{\widehat{N}_M}$. Table 3 illustrates the performance of the two-stage stopping variable M.

4. SEQUENTIAL ESTIMATION PROCEDURE

Take an initial sample of size m_0 and collect afterwards samples one by one until the sample size m satisfies the inequality

$$m \ge \frac{\chi^2_{1-\alpha} S_m q}{m p^2 \delta^2}. \tag{4.1}$$

Let us define $C := (\delta^2 p^2)/q\chi^2_{1-\alpha}$. The stopping variable or stopping time for this sequential procedure is

$$M_S \equiv M_S(\delta) = \inf\left\{m \ge m_0 : S_m \le Cm^2\right\} = m_0 + \inf\left\{n \ge 0 : S_{m_0+n} \le \beta_n\right\}, \tag{4.2}$$

where $\beta_n := \lfloor C(m_0 + n)^2 \rfloor + 1$ for $n = 0, 1, 2 \ldots$. The estimator of N at stopping is

$$\widehat{N}_{M_S} := \left\lfloor \frac{S_{M_S}}{pM_S} \right\rfloor + 1 \tag{4.3}$$

and the final bounded width interval is

$$\left[L_{M_S}\ , \widehat{N}_{M_S} + \delta\right) \quad \text{where } L_{M_S} := \max\{1, \widehat{N}_{M_S} - \delta\}.$$

4.1. Exact Distribution of Sequential Stopping Time M_S

In order to derive the exact distribution of M_S, we will first define a defective probability mass function or pmf as

$$g(j,n) := P(S_{m_0+n} = j,\ M_S > m_0 + n), \quad \text{where } j, n \text{ are some nonnegative integers.}$$

One can compute this defective pmf via some recursive relation which we derive below. Note that

$$\{M_S > m_0 + n\} = \{S_{m_0} > \beta_0, S_{m_0+1} > \beta_1, \cdots, S_{m_0+n} > \beta_n\}.$$

Therefore,

$$g(j,0) = P(S_{m_0} = j, S_{m_0} > \beta_0) = I(j \geq \beta_0 + 1) b(j; m_0 N, p). \tag{4.4}$$

Similarly, one can find the recursive relation between these defective pmfs as

$$\begin{aligned}
g(j,n) &= P\left(S_{m_0+n} = j, S_{m_0} > \beta_0, S_{m_0+1} > \beta_1, \cdots, S_{m_0+n} > \beta_n\right) \\
&= I(j \geq \beta_n + 1) \sum_{l=\beta_{n-1}+1}^{j} P\left(S_{m_0+n} = j, S_{m_0} > \beta_0, \cdots, S_{m_0+n} > \beta_n, S_{m_0+n-1} = l\right) \\
&= I(j \geq \beta_n + 1) \sum_{l=\beta_{n-1}+1}^{j} P\left(X_{m_0+n} = j - l, S_{m_0} > \beta_0, \cdots, S_{m_0+n-1} > \beta_{n-1}, S_{m_0+n-1} = l\right) \\
&= I(j \geq \beta_n + 1) \sum_{l=\beta_{n-1}+1}^{j} P\left(X_{m_0+n} = j - l\right) P\left(M_S > m_0 + n - 1, S_{m_0+n-1} = l\right) \\
&= I(j \geq \beta_n + 1) \sum_{l=\beta_{n-1}+1}^{j} b(j - l; N, p) g(l, n-1).
\end{aligned} \tag{4.5}$$

Notice that $g(j,n) = 0$ if $j > (m_0 + n)N$. Next, we derive the exact probability mass function of M_S using the known defective pmfs.

$$\begin{aligned}
\psi(n) := P\left(M_S = m_0 + n\right) &= P\left(M_S > m_0 + n - 1\right) - P\left(M_S > m_0 + n\right) \\
&= \sum_{j=\beta_{n-1}+1}^{\infty} g(j, n-1) - \sum_{j=\beta_n+1}^{\infty} g(j,n) \\
&= \sum_{j=\beta_{n-1}+1}^{\infty} g(j, n-1) - \sum_{j=\beta_n+1}^{\infty} \sum_{l=\beta_{n-1}+1}^{j} b(j - l; N, p) g(l, n-1)
\end{aligned} \tag{4.6}$$

Interchanging the order of the double sum in equation 4.6, we have

$$\begin{aligned}\psi(n) &:= \sum_{j=\beta_{n-1}+1}^{\infty} g(j,n-1) - \sum_{l=\beta_{n-1}+1}^{\infty} g(l,n-1) \sum_{j=\beta_n+1}^{\infty} b(j-l;N,p) \\ &= \sum_{j=\beta_{n-1}+1}^{\infty} g(j,n-1) - \sum_{l=\beta_{n-1}+1}^{\infty} g(l,n-1)\left(1-B(\beta_n-l;N,p)\right) \\ &= \sum_{l=\beta_{n-1}+1}^{\beta_n} g(l,n-1)B(\beta_n-l;N,p). \end{aligned} \tag{4.7}$$

4.2. Exact Distribution of $\widehat{N}_{M_S}$

The sequential procedure stops at $M_S = m_0 + m$ if

$$S_{m_0} > \beta_0, \ldots, S_{m_0+m-1} > \beta_{m-1},\ \beta_{m-1}+1 \le S_{m_0+m} \le \beta_m.$$

The corresponding values of $\widehat{N}_{M_S}$ are

$$\left\lfloor \frac{\beta_{m-1}+1}{p(m_0+m)} \right\rfloor + 1, \ldots, \left\lfloor \frac{\beta_m}{p(m_0+m)} \right\rfloor + 1.$$

Accordingly,

$$\begin{aligned} P(\widehat{N}_{M_S}=n) = &\sum_{j=1}^{\beta_0} I\left\{ \left\lfloor \frac{j}{pm_0} \right\rfloor + 1 = n \right\} b(j;Nm_0,p) \\ &+ \sum_{m=1}^{\infty} \sum_{\beta_{m-1}+1}^{\beta_m} I\left\{ \left\lfloor \frac{j}{p(m_0+m)} \right\rfloor + 1 = n \right\} \sum_{l=\beta_{m-1}+1}^{j} g(l,m-1)b(j-l;N,p). \end{aligned} \tag{4.8}$$

The moments of $\widehat{N}_{M_S}$ can easily be computed from the probability mass function in 4.8. Moreover, the coverage probability is obtained as

$$CP_{M_S} := P\left(\max\{1, \widehat{N}_{M_S}-\delta\} \le N < \widehat{N}_{M_S}+\delta\right) = \sum_{n=N-\delta+1}^{N+\delta} P(\widehat{N}_{M_S}=n). \tag{4.9}$$

5. PRESCRIBED PROPORTIONAL CLOSENESS ESTIMATOR OF N

Apart from the bounded width interval estimation approach, another popular estimation approach is to construct estimators satisfying the prescribed proportional closeness criteria. Based on sample of size m, an estimator $\tilde{N}_m$ of N is called *prescribed proportional closeness estimator* with prescribed confidence level

$(1-\alpha) \in (0,1)$ and proportional closeness $\gamma \in (0,1)$ if

$$P\left(\frac{|\tilde{N}_m - N|}{N} \leq \gamma\right) \geq 1-\alpha. \tag{5.1}$$

From (5.1), the corresponding interval estimator $\left[\tilde{N}_m(1+\gamma)^{-1}, \tilde{N}_m(1-\gamma)^{-1}\right]$ is known as the prescribed proportional closeness interval estimator of N. For this approach, we consider the estimator $\tilde{N}_m = \frac{S_m}{mp}$ and find out the minimum sample size for which

$$P\left(\left|\frac{S_m}{mp} - N\right| \leq \gamma N\right) \geq 1-\alpha.$$

Using the asymptotic normality of $\frac{S_m}{mp}$, we can write

$$P\left(\left|\frac{S_m}{mp} - N\right| \leq \gamma N\right) \approx 2\Phi\left(\frac{\sqrt{m}\gamma N}{\sqrt{Nq/p}}\right) - 1.$$

In order to satisfy (5.1), the sample size m should be

$$m \geq \frac{\chi^2_{1-\alpha} q}{Np\gamma^2},$$

where $\chi^2_{1-\alpha}$ is the $(1-\alpha)$-quantile of a chi-square distribution with one degree of freedom. Therefore, the minimum or optimal sample size to satisfy (5.1) is given by

$$n_{opt} \equiv n_{opt}(\gamma) := \left\lfloor \frac{\chi^2_{1-\alpha} q}{Np\gamma^2} \right\rfloor + 1. \tag{5.2}$$

Since $N \geq 1$, an upper bound for n_{opt} is obtained by substituting $N = 1$ in (5.2). Thus, a sample size of $\left\lfloor \frac{\chi^2_{1-\alpha} q}{p\gamma^2} \right\rfloor + 1$ will produce prescribed proportional closeness interval for any N. The problem is that this sample size may be too large. For example, if $\gamma = 0.05$ and $p = 0.1$, this upper bound is 13,824. In order to achieve (5.1) with reasonable sample size, one can consider a two-stage or sequential sampling procedure where the unknown n_{opt} is estimated by estimating N first.

Based on m samples one would tend to propose the estimator of n_{opt} as $\left\lfloor \frac{m\chi^2_{1-\alpha} q}{S_m \gamma^2} \right\rfloor + 1$. Note that $P(S_m = 0) > 0$ and, therefore, the proposed estimator of n_{opt} can be undefined with positive probability. This is certainly not desirable. To overcome this issue, we propose to estimate n_{opt} by a slightly modified estimator $\left\lfloor \frac{m\chi^2_{1-\alpha} q}{\max\{S_m, \epsilon\}\gamma^2} \right\rfloor + 1$ for some fixed $\epsilon \in (0,1)$. For our numerical study, we consider $\epsilon = 0.01$. Below, we propose a two-stage procedure to construct a prescribed proportional closeness estimator of N.

5.1. Two-Stage Sampling

Stage I: Fix the pilot sample size r and draw an initial random sample $X_1, \ldots, X_r$ from $B(N, p)$ distribution. Estimate N by $\widehat{N}_r$ and obtain the estimated sample size

$$K_1^* \equiv K_1^*(\gamma) = \left\lfloor \frac{r\chi^2_{1-\alpha} q}{\max\{S_r, \epsilon\}\gamma^2} \right\rfloor + 1 = \left\lfloor \frac{\eta}{\max\{S_r, \epsilon\}} \right\rfloor + 1, \tag{5.3}$$

where $\eta := \frac{r\chi^2_{1-\alpha} q}{\gamma^2}$. Suppose T is the final sample size for this two-stage procedure. If $K_1^* \leq r$, we stop sampling and set $T = r$ and $\widehat{N}_T = \widehat{N}_r$. Otherwise, proceed to Stage II.

Stage II: If $K_1^* > r$, we draw remaining $(K_1^* - r)$ random samples from $B(N, p)$ independent of the initial r samples and label the observations as $X^*_{r+1}, \ldots, X^*_{K_1^*}$. Set $T = K_1^*$. Define $S_r := \sum_{i=1}^r X_i$ and $S^*_{K_1^*-r} := \sum_{i=r+1}^{K_1^*} X_i^*$ if $K_1^* > r$ and $S^*_{K_1^*-r} = 0$ otherwise. The final estimator of N is

$$\widehat{N}_T = \left\lfloor \frac{1}{Tp}\left(S_r + S^*_{T-r}\right) \right\rfloor + 1,$$

where the final sample size $T \equiv T(\gamma) = \max\{r, K_1^*(\gamma)\}$. The bounded width confidence interval for N is

$$\left[\left\lfloor \frac{S_r + S^*_{T-r}}{Tp(1+\gamma)} \right\rfloor + 1, \left\lfloor \frac{S_r + S^*_{T-r}}{Tp(1-\gamma)} \right\rfloor + 1 \right]. \tag{5.4}$$

5.2. Exact Distribution of T and $\widehat{N}_T$

For $m \geq r$, the cumulative distribution function of T, denoted as F_T, is given by

$$\begin{aligned} F_T(m) = P(T \leq m) &= P\left(K_1^* \leq m\,, r \leq m\right) \\ &= P\left(\frac{\eta}{\max\{S_r, \epsilon\}} \leq m\right) \\ &= 1 - P\left(\max\{S_r, \epsilon\} < \frac{\eta}{m}\right) \\ &= 1 - I\left(m < \frac{\eta}{\epsilon}\right) B\left(\left\lfloor \frac{\eta}{m} \right\rfloor; rN, p\right). \end{aligned}$$

Clearly, $F_M(m) = 0$ for $m < r$. Therefore, the probability mass function of T is $P(T = m) = F_T(m) - F_T(m-1)$ for $m = r, r+1, \ldots$. Using this pmf, we also compute $E(T)$ and $SD(T)$ which is presented in Table 5.

Next, we derive the exact distribution of $\widehat{N}_T$. Suppose for given $S_r = s$, the stopping variable $T =$

$t(s) \equiv t = \max\{r, \lfloor \eta / \max\{s, \epsilon\} \rfloor + 1\}$. Then the cdf of $\widehat{N}_T$, denoted by $F_{\widehat{N}_T}$ is given as

$$\begin{aligned} F_{\widehat{N}_T}(n) &= P\left(\widehat{N}_T \leq n\right) = P\left(\frac{S_r + S^*_{T-r}}{Tp} \leq n\right) \\ &= \sum_{s=0}^{\infty} P\left(s + S^*_{t(s)-r} \leq npt(s) | S_r = s\right) P(S_r = s) \\ &= \sum_{s=0}^{\infty} B\left(npt(s) - s; (t(s)-r)N, p\right) b(s; rN, p), \end{aligned} \tag{5.5}$$

where $n \in \{1, 2, \ldots\}$. The expectation and standard deviation of $\widehat{N}_T$ can easily be computed using the cdf $F_{\widehat{N}_T}(n)$. Moreover, based on (5.4), we can compute the coverage probability CP_T as follows

$$\begin{aligned} CP_T &= P\left(\left\lfloor \frac{S_r + S^*_{T-r}}{Tp(1+\gamma)} \right\rfloor + 1 \leq N \leq \left\lfloor \frac{S_r + S^*_{T-r}}{Tp(1-\gamma)} \right\rfloor + 1\right) \\ &= \sum_{s=0}^{\infty} P\left(\left\lfloor \frac{s + S^*_{t(s)-r}}{t(s)p(1+\gamma)} \right\rfloor + 1 \leq N \leq \left\lfloor \frac{s + S^*_{t(s)-r}}{t(s)p(1-\gamma)} \right\rfloor + 1 \;\middle|\; S_r = s\right) P(S_r = s) \\ &= \sum_{s=0}^{\infty} \left\{ P\left(\left\lfloor \frac{s + S^*_{t(s)-r}}{t(s)p(1-\gamma)} \right\rfloor + 1 \geq N\right) - P\left(\left\lfloor \frac{s + S^*_{t(s)-r}}{t(s)p(1+\gamma)} \right\rfloor + 1 > N\right) \right\} P(S_r = s) \\ &= \sum_{s=0}^{\infty} \left\{ P\left(\frac{s + S^*_{t(s)-r}}{t(s)p(1-\gamma)} > N - 1\right) - P\left(\frac{s + S^*_{t(s)-r}}{t(s)p(1+\gamma)} > N\right) \right\} P(S_r = s) \\ &= \sum_{s=0}^{\infty} \left\{ P\left(S^*_{t(s)-r} \leq Npt(s)p(1+\gamma) - s\right) - P\left(S^*_{t(s)-r} \leq (N-1)pt(s)p(1-\gamma) - s\right) \right\} P(S_r = s) \\ &= \sum_{s=0}^{\infty} \{ B\left(Npt(s)(1+\gamma) - s; N(t(s)-r), p\right) \\ &\qquad - B\left((N-1)pt(s)(1-\gamma) - s; N(t(s)-r), p\right) \} \, b(s; rN, p) \end{aligned}$$

6. NUMERICAL STUDY

In this section, we present the results obtained from simulation study as well as from exact calculation of distributions of stopping time and estimators at stopping. Table 3 illustrates the performance of our proposed two-stage procedure to construct bounded width confidence interval of N for a given choice of pilot sample size $k = 30$ and confidence level $(1 - \alpha) = 0.9$. In this study, $E[M]$, $\widehat{SD[M]}$, $E\left[\widehat{N}_M\right]$, and CP_{TS} represent the exact values of the expectation of stopping variable M, standard deviation of M, expected value of $\widehat{N}_M$, and the coverage probability obtained from the two-stage procedure. Similarly, $\widehat{E[M]}$, $\widehat{SD[M]}$, $\widehat{E\left[\widehat{N}_M\right]}$, and $\widehat{CP}_{TS}$ represent the simulated values of the expectation and standard deviation of M, the expected value of $\widehat{N}_M$, and the coverage probability obtained from the two-stage procedure. All simulated values presented here are based on 50,000 repetitions. Table 3 shows that expected sample size is very close to the optimal sample size in all the scenarios, and the coverage probabilities are approximately the same as the desired level of 0.9. The expected value of the point estimator of N at stopping is very close to that of the

true value of N. Therefore, we have strong evidence that our proposed two-stage procedure is performing well. Moreover, all the simulated values match with the corresponding exact values which validates all the correctness of the derived exact formulas.

Table 4 presents the performance of purely sequential procedure to construct bounded width confidence interval of N. The exact computation of expectation and standard deviation of M_S, expectation of $\widehat{N}_{M_S}$, and the coverage probability CP_{M_S} take quite some time. Therefore, in this article we present only two cases just as examples. For $m_0 = 30$, $\alpha = 0.1$, $\delta = 1$, $N = 25$, and $p = 0.6$, we obtain $E[M_S] = 45.581$, $SD[M_S] = 1.154$, $E\left[\widehat{N}_{M_S}\right] = 25.489$, and $CP_{M_S} = 0.9057$. For $m_0 = 30$, $\alpha = 0.1$, $\delta = 1$, $N = 25$, and $p = 0.3$, we obtain $E[M_S] = 158.27$, $SD[M_S] = 3.853$, $E\left[\widehat{N}_{M_S}\right] = 25.493$, and $CP_{M_S} = 0.9006$. In Table 4, we present the simulated values of the expected value and standard deviation of stopping time M_S, expected value of the estimator of N at stopping, and the coverage probability $\widehat{CP}_{M_S}$. The above two examples of exact computation result match with the simulated values which validate the correctness of the derived exact formulas. Note that the expected sample size and the coverage probability for the sequential procedure is similar to that of the two-stage procedure. However, the standard deviation of the stopping time is much lower for the sequential procedure than that of the two-stage procedure. Hence, the sequential stopping variable M_S may be preferred over the two-stage stopping variable M.

Table 5 illustrates the performance of our proposed two-stage procedure to construct prescribed proportional closeness interval of N for a given choice of pilot sample size $r = 30$ and confidence level $(1 - \alpha) = 0.9$. Here, $E[T]$, $\widehat{SD[T]}$, $E\left[\widehat{N}_T\right]$, and CP_T represent the exact values of the expectation of stopping variable T, standard deviation of T, expected value of $\widehat{N}_T$, and the coverage probability obtained from the two-stage procedure. Similarly, $\widehat{E[T]}$, $\widehat{SD[T]}$, $\widehat{E\left[\widehat{N}_T\right]}$, and $\widehat{CP}_T$ represent the simulated values of the expectation and standard deviation of T, the expected value of $\widehat{N}_T$, and the coverage probability obtained from the two-stage procedure. This table also reports $\overline{W}$, the average cardinality (or equivalently one can view this as the width of the confidence interval) of the obtained confidence sets from 10,000 simulations. Table 5 shows that expected sample size is close to the optimal sample size in almost all the scenarios, and the coverage probabilities are higher than the desired level of 0.9. The expected value of the point estimator of N at stopping is very close to that of the true value of N. Therefore, we have evidence that our proposed two-stage procedure is performing well. Moreover, all the simulated values match with the corresponding exact values which validates all the correctness of the derived exact formulas.

ACKNOWLEDGMENT

The authors gratefully acknowledge Professor Rasul A. Khan for suggesting this problem and for his assistance and encouragement. We also thank the editor, Professor Nitis Mukhopadhyay, for providing his valuable feedback that helped improve this paper.

Table 1. Estimated probabilities of $X_{(m)} \geq \overline{N}$ for different values of N, p, and m

N	p	m	$\widehat{P(X_{(m)} \geq \overline{N})}$
10	0.8	20	0.0022
10	0.6	20	0.0016
10	0.4	20	0
20	0.9	20	0.0013
20	0.8	20	0.0013
20	0.7	20	0.00004
50	0.95	20	0.0001
50	0.9	20	0.00005
50	0.8	20	0
100	0.95	20	0.00004
100	0.9	20	0

Table 2. Expected values and mean squared errors of $\overline{N}$ and $X_{(m)}$ for different values of N, p, and m

N	p	m	$E[\overline{N}]$	$V[\overline{N}]$	$E[X_{(m)}]$	$MSE[X_{(m)}]$
10	0.8	20	10	0.125	9.897	1.1354
10	0.5	20	10	0.5	7.899	25.001
10	0.1	20	10	4.5	3.027	97.8186
20	0.8	20	20	0.250	18.959	22.1268
20	0.5	20	20	1.000	14.116	138.5934
20	0.1	20	20	9.000	4.787	428.0409
50	0.8	20	50	0.635	44.947	268.7478
50	0.5	20	50	1.250	11.566	1102.5700
50	0.1	20	50	11.250	9.270	2887.2850

REFERENCES

Blumenthal, S. and Dahiya, R. C. (1981). Estimating the Binomial Parameter n, *Journal of American Statistical Association* 76: 903–909.

DasGupta, A. and Rubin, H. (2005). Estimation of the Binomial Parameter When Both n, p Are Unknown, *Journal of Statistical Planning and Inference* 130: 391–404.

De, S. K. (2014). Modified Three-Stage Sampling for Fixed-Width Interval Estimation of the Common

Table 3. Performances of the two-stage methodology for constructing bounded width confidence interval of N with pilot sample size $k = 30$ and $\alpha = 0.1$

δ	N	p	n_0	$E[M]$	$\widehat{E[M]}$	$SD[M]$	$\widehat{SD[M]}$	$E\left[\widehat{N}_M\right]$	$\widehat{E\left[\widehat{N}_M\right]}$	CP_{TS}	$\widehat{CP}_{TS}$
1	25	0.1	609	609.24	609.44	66.69	66.64	25.482	25.482	0.8982	0.9011
		0.3	158	158.34	158.40	8.80	8.81	25.481	25.479	0.8997	0.9009
		0.6	46	45.55	45.55	1.37	1.37	25.473	25.474	0.9006	0.9026
	100	0.1	2435	2435.49	2434.74	133.37	133.42	100.495	100.494	0.8995	0.9007
		0.3	632	631.79	631.80	17.61	17.60	100.495	100.495	0.9000	0.9008
		0.6	181	180.85	180.85	2.70	2.69	100.493	100.490	0.9004	0.8985
2	100	0.1	609	609.25	609.18	33.35	33.42	100.482	100.482	0.8996	0.8996
		0.3	158	158.32	158.34	4.41	4.42	100.481	100.478	0.9003	0.9003
		0.6	46	45.59	45.59	0.7304	0.7262	100.474	100.473	0.9010	0.9036
	1000	0.1	6088	6087.97	6087.30	105.44	105.00	1000.499	1000.493	0.9000	0.8988
		0.3	1579	1578.74	1578.77	13.92	13.89	1000.498	1000.506	0.9000	0.9004
		0.6	451	451.42	451.43	2.15	2.15	1000.498	1000.497	0.9002	0.9006

Variance of Equi-Correlated Normal Distributions, *Sequential Analysis* 33: 87–111.

De, S. K. and Mukhopadhyay, N. (2015). Fixed Accuracy Interval Estimation of the Common Variance in a Equi-Correlated Normal Distribution, *Sequential Analysis* 34: 1–23.

Feldman, D. and Fox, M. (1968). Estimation of the Parameter n in the Binomial Distribution, *Journal of American Statistical Association* 63: 150–158.

Haldane, J. B. S. (1941). The Fitting of the Binomial Distributions, *Annals of Eugenics* 11: 179–181.

Hall, P. (1994). On the Erratic Behavior of Estimators of N in the Binomial (N, p) Distribution, *Journal of American Statistical Association* 89: 344–352.

McCabe, G. P. (1973). Estimation of the Number of Terms in a Sum, *Journal of American Statistical Association* 68: 452–456.

Mukhopadhyay, N. (2005). A New Approach to Determine the Pilot Sample Size in Two-Stage Sampling, *Communications in Statistics – Theory & Methods* 34: 1275–1295.

Mukhopadhyay, N., DeSilva, B. M., and Waikar, V. B. (2010). On Two-Stage Confidence Interval Procedures and Their Comparisons for Estimating the Difference of Normal Means, *Sequential Analysis* 31: 1–20.

Olkin, I., Petkau, A. J., Zidek, J. V. (1981). A Comparison of the n Estimators for the Binomial Distributions, *Journal of American Statistical Association* 76: 637–642.

Raftery, A. (1988). Inference on the Binomail N Parameter: a Hierarchical Bayes Approach, *Biometrika* 75: 223–228.

Table 4. Performances of the sequential methodology for constructing bounded width confidence interval of N with pilot sample size $m_0 = 30$ and $\alpha = 0.1$

δ	N	p	n_0	$\widehat{E[M_S]}$	$\widehat{SD[M_S]}$	$\widehat{E\left[\widehat{N}_{M_S}\right]}$	$\widehat{CP}_{M_S}$
1	25	0.1	609	608.55	14.87	25.482	0.8958
		0.3	158	158.08	3.87	25.468	0.9007
		0.6	46	45.50	1.16	25.472	0.9035
	100	0.1	2435	2435.19	14.87	100.491	0.8998
		0.3	632	631.74	3.84	100.494	0.9017
		0.6	181	180.85	1.14	100.490	0.8994
2	100	0.1	609	609.09	7.39	100.483	0.9039
		0.3	158	158.25	1.94	100.464	0.9012
		0.6	46	45.57	0.6251	100.469	0.9057
	1000	0.1	6088	6088.00	7.32	1000.506	0.9049
		0.3	1579	1578.75	1.93	1000.514	0.8994
		0.6	451	451.42	0.6210	1000.503	0.8988

Table 5. Performances of the two-stage methodology for constructing prescribed proportional closeness interval of N with pilot sample size $r = 30$ and $\alpha = 0.1$

γ	N	p	n_{opt}	$E[T]$	$\widehat{E[T]}$	$SD[T]$	$\widehat{SD[T]}$	$E\left[\widehat{N}_T\right]$	$\widehat{E\left[\widehat{N}_T\right]}$	CP_T	$\widehat{CP}_T$	$\overline{W}$
0.05	25	0.1	390	394.85	395.44	44.36	44.65	25.52	25.51	0.9426	0.9424	3.507
		0.3	102	101.83	101.87	5.69	5.69	25.52	25.51	0.9434	0.9463	3.509
	50	0.1	195	196.47	196.51	15.40	15.55	50.54	50.53	0.9355	0.9361	6.013
		0.3	51	51.08	51.07	2.02	2.02	50.53	50.53	0.9364	0.9345	6.023
0.01	25	0.1	9740	9860.34	9830.73	1109.01	1106.45	25.50	25.50	0.9488	0.9526	1.904
		0.3	2526	2533.61	2534.28	142.36	141.24	25.50	25.50	0.9488	0.9506	1.903
	50	0.1	4870	4900.03	4898.22	384.46	381.76	50.50	50.50	0.9487	0.9500	2.000
		0.3	1263	1265.13	1264.14	50.10	49.54	50.50	50.50	0.9488	0.9494	2.001

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

Zacks, S. (2001). The Operating Characteristics of Sequential Procedures in Reliability, in *Handbook of Statistics*, vol. 20, N. Balakrishnan and C. R. Rao, eds., pp. 789–811, Amsterdam: Elsvier.

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

Zacks, S. (2015). Exact Evaluation of Two-Stage Stein-like Procedures – Review, *Sequential Analysis* 34: 461–482.

Zacks, S. and Mukhopadhyay, N. (2007a). Bounded Risk Estimation of Linear Combination of the Location and Scale Parameters of Exponential Distributions under Two-Stage Sampling, *Journal of Statistical Planning and Inference* 137: 3672 – 3686.

Zacks, S. and Mukhopadhyay, N. (2007b). Distributions of Sequential and Two-Stage Stopping Times for Fixed-Width Confidence Intervals in Bernoulli Trials: Applications in Reliability, *Sequential Analysis* 26: 425 – 442.