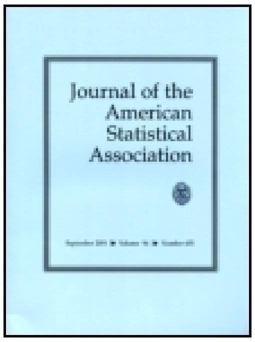


Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://amstat.tandfonline.com/loi/uasa20

Exact Optimal Confidence Intervals for Hypergeometric Parameters

Weizhen Wang

To cite this article: Weizhen Wang (2015) Exact Optimal Confidence Intervals for Hypergeometric Parameters, Journal of the American Statistical Association, 110:512, 1491-1499, DOI: 10.1080/01621459.2014.966191

To link to this article: http://dx.doi.org/10.1080/01621459.2014.966191

Exact Optimal Confidence Intervals for Hypergeometric Parameters

Weizhen WANG

For a hypergeometric distribution, denoted by Hyper(M, N, n), where N is the population size, M is the number of population units with some attribute, and n is the given sample size, there are two parametric cases: (i) N is unknown and M is given; (ii) M is unknown and N is given. For each case, we first show that the minimum coverage probability of commonly used approximate intervals is much smaller than the nominal level for any n, then we provide exact smallest lower and upper one-sided confidence intervals and an exact admissible two-sided confidence interval, a complete set of solutions, for each parameter. Supplementary materials for this article are available online.

KEY WORDS: Capture-recapture sampling; Coverage probability; One-sided interval; Set inclusion; Two-sided interval.

1. INTRODUCTION

In survey sampling on a finite population, a simple random sample is typically selected without replacement, in which case a hypergeometric distribution models the observation. More precisely, we select a sample of size n without replacement from a finite population of N units, among which M units have some attribute, and we observe X, the number of units having the attribute in the sample. Then X follows Hyper(M, N, n) with probability mass function

$$p_h(x; M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \quad \text{for any} \quad \max\{0, n+M-N\} \le x \le \min\{M, n\}. \quad (1)$$

The sample size n is a given positive integer. However, either N or M can be the unknown parameter of interest as discussed in the following two paragraphs.

In a capture and recapture sampling plan without replacement, we randomly capture M (known) units from a population of N (unknown) units, then tag and release them. Later, a simple random sample of size n is recaptured from the population, and $X \sim$ Hyper(M, N, n), the number of tagged units in this sample, is observed. Here N is the unknown parameter, with range $[\max\{n, M\}, +\infty)$. One goal of this article is to estimate N using exact $1-\alpha$ confidence intervals $C(X)$ of the following forms: $[L(X), +\infty)$ and $[\max\{n, M\}, U(X)]$ (one-sided) and $[L(X), U(X)]$ (two-sided). The word "exact" means that the coverage probability of $C(X)$ is at least $1-\alpha$, that is,

$$\text{Cover}_C(N) \overset{\text{def}}{=} P(N \in C(X)) \ge 1-\alpha, \quad \text{for any} \quad N = \max\{n, M\}, \ldots, +\infty. \quad (2)$$

There have been many attempts to derive point estimators and the corresponding approximate intervals for N, see, for example, Petersen (1896), Lincoln (1930), Sekar and Deming (1949), Chapman (1951), Bailey (1951), and Seber (1982, 1986, 1992). Here are two widely used approximate intervals for N described in statistical textbooks (e.g., Thompson 2002, p. 235):

$$C_{N1}(X) = \left[\frac{nM}{X} \pm z_{\alpha/2}\left(\frac{Mn(M-X)(n-X)}{X^3}\right)^{1/2}\right], \quad (3)$$

where z_α is the upper 100αth percentile of a standard normal distribution, and

$$C_{N2}(X) = [\tilde{N} \pm z_{\alpha/2}\widehat{\text{se}}(\tilde{N})], \quad (4)$$

where

$$\tilde{N} = \frac{(n+1)(M+1)}{X+1} - 1,$$
$$\widehat{\text{se}}(\tilde{N}) = \left(\frac{(M+1)(n+1)(M-X)(n-X)}{(X+1)^2(X+2)}\right)^{1/2}.$$

The first interval is based on the Lincoln–Petersen estimator, and an obvious drawback is that there is no definition for $X = 0$. If we let $C_{N1}(0) = [\max\{n, M\}, +\infty)$, then it is too wide and also counter-intuitive since a large estimate of N would be expected for the smallest possible observation $X = 0$. The second interval, which is based on the Chapman estimator, overcomes this problem and always has a finite length, including when $X = 0$. However, as shown later, this interval has a zero minimum coverage probability. Both intervals have poor coverage probabilities, see Figure 1. We provide in Theorem 1 an exact admissible two-sided interval for N, that is, an interval whose proper subinterval is of level strictly less than $1-\alpha$.

On the other hand, it is also common in practice that the population size N is known and one wants to estimate M, the unknown number of population units with an attribute, based on X, the observed number of units with the attribute in a simple random sample of size n. This is equivalent to estimating the population proportion $p = M/N$. Again, X follows Hyper(M, N, n). This case arises, for example, when one wants to estimate the proportion of voters in a given population favoring a proposition. The following Wald-type interval is typically recommended in

Weizhen Wang is Visiting Professor, College of Applied Sciences, Beijing University of Technology, Beijing 100124, P. R. China. He is also Professor, Department of Mathematics and Statistics, Wright State University, Dayton, OH 45435 (E-mail: weizhen.wang@wright.edu). The author thanks a referee, the associate editor, the co-editor, and Professor Daniel T. Voss for their helpful comments and suggestions that substantially improved the article and also thanks Yunyang E. Wang for editing the article.

Journal of the American Statistical Association
December 2015, Vol. 110, No. 512, Theory and Methods
DOI: 10.1080/01621459.2014.966191

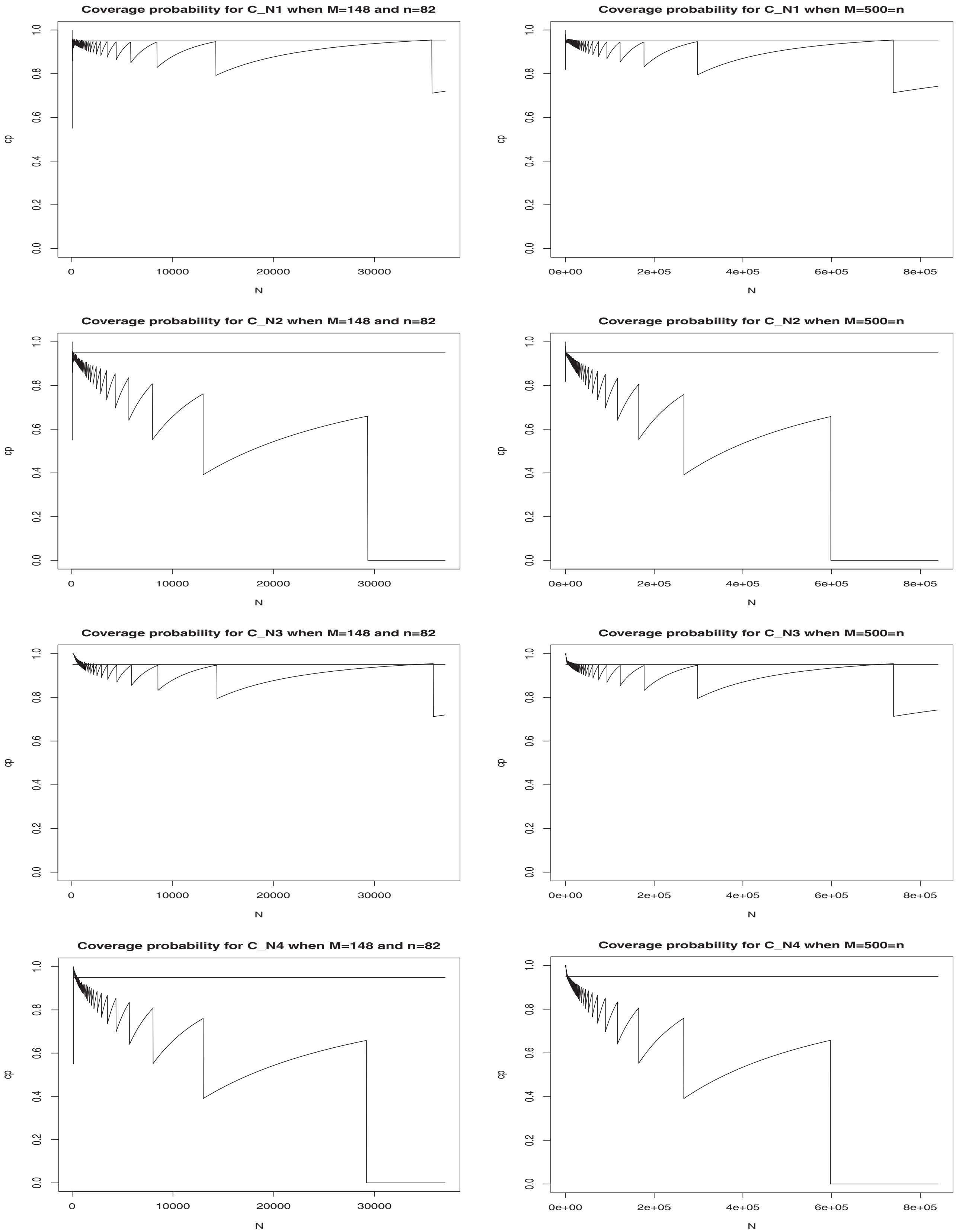


Figure 1. Coverage probabilities (cp) of four approximate 95% intervals, $C_{N1}(X)$, $C_{N2}(X)$, $C_{N3}(X)$, and $C_{N4}(X)$, plotted versus N, with a reference line at 0.95.

Table 1. An upper bound b for the minimum coverage probability for $C_{N1}(X)$

Level	$1-\alpha=0.9$			$1-\alpha=0.95$			$1-\alpha=0.99$		
$n(=M)$	100	500	1000	100	500	1000	100	500	1000
b	0.683	0.685	0.685	0.711	0.713	0.713	0.754	0.756	0.756

practice,

$$C^M(X) = \left[\frac{NX}{n} \pm t_{\alpha/2,n-1}\left(\frac{(N-n)NX(n-X)}{n^2(n-1)}\right)^{1/2}\right], \tag{5}$$

where $t_{\alpha,n-1}$ is the upper 100αth percentile of a t-distribution with $n-1$ degrees of freedom (Thompson 2002, p. 40). We will show in Lemma 7 that $C^M(X)$ has a coverage probability as low as n/N for any α. A similar phenomenon is observed for the Wald interval of a binomial proportion, see Agresti and Coull (1998) and Brown, Cai, and DasGupta (2001). There are three exact $1-\alpha$ intervals for M by Cochran (1977, p. 57), Konijn (1973, p. 79), and Buonaccorsi (1987). As pointed out by Buonaccorsi (1987), the last two intervals are identical and are a proper subset of the first. We will further improve the Konijn (1973) interval. The second goal of this article is to estimate M with exact $1-\alpha$ confidence intervals $C(X)$ of the following forms: $[L(X), N]$ and $[0, U(X)]$ (one-sided) and $[L(X), U(X)]$ (two-sided). Similar to (2), an exact $1-\alpha$ interval $C(X)$ for M satisfies

$$\text{Cover}_C(M) \overset{\text{def}}{=} P(M \in C(X)) \geq 1-\alpha, \quad \text{for any } M = 0, 1, \ldots, N. \tag{6}$$

Theorem 2 describes an admissible two-sided interval for M.

Section 2 contains a complete interval construction for N. Three exact optimal intervals based on X, including two smallest one-sided intervals and one admissible two-sided interval, are proposed. In Section 3, we establish similar results for M. Section 4 contains discussion. All proofs are given in the Appendix.

2. ESTIMATING THE POPULATION SIZE N

We first show poor coverage for four approximate intervals that justifies the usage of exact intervals, then provide two smallest one-sided intervals, and lastly describe how to derive an admissible two-sided interval from the two smallest one-sided intervals.

2.1 Coverage of Four Approximate Intervals for N

We now investigate the coverage probabilities of four approximate intervals, including $C_{N1}(X)$ and $C_{N2}(X)$ for N introduced in (3) and (4).

Lemma 1. For any $\alpha \in (0,1)$, write $C_{N1}(X) = [L_{N1}(X), U_{N1}(X)]$ and let u_1 be the smallest integer larger than $U_{N1}(1)$. Then the minimum coverage probability of the interval $C_{N1}(X)$ given in (3) is less than or equal to $\min\{a, b\}$, where $a = \frac{\min\{n,M\}}{\max\{n,M\}+1}$ and $b = p_h(0; M, u_1)$ with p_h given in (1).

Is $\min\{a, b\}$ a sharp upper bound for the minimum coverage probability of $C_{N1}(X)$? Our numerical study provides a positive answer. If the ratio of n/M is far away from 1, then $a < b$; otherwise, $b < a$. For example, if the second sample size n is half or double the first sample size M, then the minimum coverage probability is about 0.5. To have a better (closer to $1-\alpha$) coverage probability, one would think of choosing $n \approx M$ to make a close to 1. This, however, is not successful because the minimum coverage bounded by b is still much smaller than the nominal level as shown in Table 1 and Figure 1. For example, when $M = 500 = n$ and $\alpha = 0.05$, $a = 500/501$ is close to 1, but $b = 0.713$. So the minimum coverage of $C_{N1}(X)$, which is no larger than b, is much less than the nominal level 0.95.

Many researchers recommend $C_{N2}(X)$ over $C_{N1}(X)$ since the Chapman estimator $\tilde{N}$ is less biased (Wittes 1972). However, $C_{N2}(X)$ has a worse coverage than $C_{N1}(X)$.

Lemma 2. For any $\alpha \in (0,1)$ and any M and n, the minimum coverage probability of the interval $C_{N2}(X)$ given in (4) is equal to zero.

The interval $C_{N2}(X)$ has poor and unstable coverage, especially for large N (see Figure 1). Another attempt to improve $C_{N1}(X)$, pointed out by a referee, is

$$C_{N3}(X) = \left[\frac{nM}{X} \pm z_{\alpha/2}\left(\frac{n^2M(M-X)}{X^3}\right)^{1/2}\right], \tag{7}$$

which was proposed by Bailey (1951). This interval is very similar to $C_{N1}(X)$ except that the standard error ignores the finite population correction. Bonett (1988) concluded from a simulation study that $C_{N3}(X)$ was more efficient than $C_{N1}(X)$. Our study indicates the improvement in coverage probability is not good enough. Lemma 3 provides a sharp bound for the minimum coverage probability for $C_{N3}(X)$ (see Table 2 for the bound). The proof is similar to Lemma 1 and is omitted.

Lemma 3. For any $\alpha \in (0,1)$, write $C_{N3}(X) = [L_{N3}(X), U_{N3}(X)]$ and let u_1' be the smallest integer larger than $U_{N3}(1)$. Let $a' = \frac{M}{n+1}$ and $b' = p_h(0; M, u_1')$ with p_h given in (1). Then the minimum coverage probability of the interval $C_{N3}(X)$ given in (7) is less than or equal to $\min\{a', b'\}$ if $M < n$, and is less than or equal to b' if $M \geq n$.

Bailey (1951) described another approximate interval for N,

$$C_{N4}(X) = \left[\frac{(n+1)M}{X+1} \pm z_{\alpha/2}\left(\frac{M^2(n+1)(n-X)}{(X+1)^2(X+2)}\right)^{1/2}\right], \tag{8}$$

which always has a finite length. Consequently, it has a zero minimum coverage probability as stated in Lemma 4. The proof is similar to Lemma 2 and is omitted.

Lemma 4. For any $\alpha \in (0,1)$ and any M and n, the minimum coverage probability of the interval $C_{N4}(X)$ given in (8) is equal to zero.

Table 2. An upper bound, b' or $\min\{a', b'\}$, for the minimum coverage probability for $C_{N3}(X)$

Level	$1-\alpha=0.9$			$1-\alpha=0.95$			$1-\alpha=0.99$		
$n(=M)$	100	500	1000	100	500	1000	100	500	1000
b'	0.683	0.685	0.685	0.712	0.713	0.713	0.755	0.756	0.756
$n(=2M)$	100	500	1000	100	500	1000	100	500	1000
$\min\{a', b'\}$	0.495	0.499	0.500	0.495	0.499	0.500	0.495	0.499	0.500

Figure 1 gives the coverage probabilities for $C_{N1}(X)$ through $C_{N4}(X)$ in two cases: $M=148$ and $n=82$, and $M=500=n$. It is clear that $C_{N2}(X)$ and $C_{N4}(X)$ are worse than the other two, and $C_{N1}(X)$ and $C_{N3}(X)$ have similar coverage when N is large. However, none of them is recommended for use in practice due to poor coverage even for large sample sizes. This strongly motivates the use of exact intervals.

2.2 Two Exact Smallest One-Sided Intervals for N

Intuitively, a large value of N tends to yield a small value of X. Then the confidence limits of $C(X)$, $L(X)$, and $U(X)$ are nonincreasing in X. Here are two results to derive the exact smallest lower and upper one-sided $1-\alpha$ confidence intervals for N.

Lemma 5. For any $\alpha \in (0,1)$, let

$$L_S(x)=\begin{cases}\max\{n, M\}, & \text{if } x=\min\{n,M\};\\ \max\left\{N : \sum_{i=x+1}^{\min\{n,M\}} p_h(i; M, N-1) \geq 1-\alpha\right\}, & \text{otherwise.}\end{cases} \tag{9}$$

Then, among all $1-\alpha$ confidence intervals for N of the form $[L(X), \infty)$ with nonincreasing $L(X)$, $[L_S(X), \infty)$ is the smallest, that is, $L(X) \leq L_S(X)$.

Lemma 6. For any $\alpha \in (0,1)$, let

$$U_S(x)=\begin{cases}+\infty, & \text{if } x=0;\\ \min\left\{N : \sum_{i=0}^{x-1} p_h(i; M, N+1) \geq 1-\alpha\right\}, & \text{otherwise.}\end{cases} \tag{10}$$

Then, among all $1-\alpha$ confidence intervals for N of the form $[0, U(X)]$ with nonincreasing $U(X)$, $[0, U_S(X)]$ is the smallest, that is, $U_S(X) \leq U(X)$.

2.3 An Exact Admissible Two-Sided Interval $C_I(X)$ for N

Here is the first major result of this article. We are only interested in confidence intervals: (i) of level $1-\alpha$; (ii) with both lower and upper confidence limits nonincreasing in X. Let $[L_O(X), +\infty)$ and $[\max\{n, M\}, U_O(X)]$ be the smallest one-sided $1-\alpha/2$ intervals for N derived in Lemmas 5 and 6, respectively. Then $C_O(X)=[L_O(X), U_O(X)]$ is an exact interval of level $1-\alpha$ satisfying the two properties. However, can $C_O(X)$ be improved? More precisely, is there a proper subinterval of $C_O(X)$ still of level $1-\alpha$ and with nonincreasing limits? We propose an iterative algorithm (Algorithm I) that generates an exact interval, $C_I(X)=[L_I(X), U_I(X)]$, which is a subset of $C_O(X)$ and is admissible under the set inclusion criterion (Wang 2006). Therefore, the interval $C_I(X)$ is recommended for estimating N in practice due to reliability and optimality. The author has developed an R code to compute $C_I(X)$.

Table 3 contains the 95% interval $C_O(X)=[L_O(X), U_O(X)]$ and the proposed interval $C_I(X)=[L_I(X), U_I(X)]$ when $M=10$ and $n=8$. The improvement is substantial. The minimum coverage probabilities for the two intervals $C_O(X)$ and $C_I(X)$ are 0.9712426 and 0.950026, respectively. Note that the length of $C_O(X)$, $U_O(X)-L_O(X)$, is decreasing in X. We construct $C_I(x)$ by squeezing each interval $C_O(x)$ (i.e., we choose the shortest interval by lifting up $L_O(x)$ and pressing down $U_O(x)$) at $x=0$ first, at $x=1$ sec, and so on, until x reaches the last value $\min\{n, M\}$, with the minimum coverage probability always maintained at least at $1-\alpha$ during the whole process. The procedure ends in $(\min\{n,M\}+1)$ steps. In each step the shortest interval always exists, and we will also show uniqueness. Since in each step we pick the shortest interval, the resulting interval $C_I(X)$ is admissible.

For illustration purposes, consider Table 3 with given $L_O(X)$ and $U_O(X)$. In Step 1, we improve $C_O(x)$ at $x=0$ by lifting $L_O(0)=31$ up as large as possible while keeping the rest of the $C_O(x)$'s unchanged. The process stops at $L_I(0)=36$ since this value makes the interval of level 0.95 and $L_I(0)=37$ would yield a confidence level less than 0.95. In this step, we only solve for one unknown $L_I(0)$ because $U_I(0)$ must be infinity to assure the 0.95 level. In Step 2, we take the interval obtained in Step 1 and improve it at $x=1$. That is, lift $L_O(1)=22$ up to an integer c, press $U_O(1)=3168$ down to an integer d, keep the interval unchanged for any $x \neq 1$, and maintain the level at at least 0.95 during this process. There are only a finite number of choices for the pair (c,d) since $22 \leq c \leq d \leq 3168$. So, pick the pair $(c^*, d^*)=(24, 1568)$ that has the smallest $d-c$ among all such pairs (c,d), and define $[L_I(1), U_I(1)]=[24, 1568]$. In this step, two unknowns $L_I(1)$ and $U_I(1)$ are determined. The remaining steps are similar to Step 2 and the construction ends in Step 9.

A formal description of Algorithm I that generates the interval $C_I(X)$ for N is given below. In each step of $k=1,\ldots,\min\{n,M\}+1$, we compute an interval $C_k(X)=[L_k(X), U_k(X)]$, which is the same as $C_{k-1}(x)$ for $x \leq k-2$ and $k \geq 2$. The last interval obtained is defined to be $C_I(X)$. Clearly, $C_I(x)=C_k(x)$ for $x \leq k-1$.

Table 3. Two 95% confidence intervals, $C_O(X)$ and $C_I(X)$, when $M=10$ and $n=8$

x	0	1	2	3	4	5	6	7	8
$U_O(x)$	∞	3168	301	108	57	35	24	17	13
$U_I(x)$	∞	1568	208	84	47	35	23	16	12
$L_I(x)$	36	24	18	16	14	13	12	11	10
$L_O(x)$	31	22	18	16	14	13	12	11	10

Step 1. We determine $C_I(0) = [L_I(0), U_I(0))$ by squeezing $C_O(0) = [L_O(0), U_O(0))$. The upper limit has to be $+\infty$ (otherwise the minimum coverage probability is zero). So only lift the lower limit up as large as possible and keep the $C_O(x)$'s for $x > 0$ unchanged. In particular, for any given integer $a \geq L_O(0)$, let

$$C_{a,+\infty}(x) = \begin{cases} (a, +\infty), & \text{if } x = 0; \\ C_O(x), & \text{if } x > 0 \end{cases} \tag{11}$$

if the minimum coverage of $C_{a,+\infty}(X)$ is at least $1 - \alpha$. Such an interval exists since $C_{a,+\infty}(X) = C_O(X)$ is one such interval. Then $C_1(X)$ is the $C_{a,+\infty}(X)$ with the largest a, denoted by a^*, which is obviously unique. In this step, we obtain $C_1(X) = C_{a^*,+\infty}(X)$ and $C_I(0) = C_1(0) = [a^*, +\infty)$.

By induction, the interval is derived up to Step $k - 1$ for $k > 1$, that is, $C_{k-1}(X)$ is constructed and $C_I(x)$ is constructed for $x \leq k - 2$.

Step k. We need to determine $C_I(x) = [L_I(x), U_I(x)]$ at $x = k - 1$. For each pair of integers (c, d) with $L_{k-1}(k - 1) \leq c \leq d \leq U_{k-1}(k - 1)$, let

$$C_{c,d}(x) = \begin{cases} C_I(x), & \text{if } x \leq k - 2; \\ [c, d], & \text{if } x = k - 1; \\ [L_{k-1}(x), \min\{U_{k-1}(x), d\}], & \text{otherwise} \end{cases} \tag{12}$$

if the minimum coverage of $C_{c,d}(X)$ is at least $1 - \alpha$. Several facts are worth mentioning: (i) $C_{c,d}(X)$ exists because $C_{k-1}(X)$ is one such interval; (ii) the minimum function in the third line of (12) is to make sure the interval has nonincreasing limits; (iii) since there are only finitely many such pairs (c, d), there exists a pair (c^*, d^*) so that $d^* - c^*$ is the smallest $d - c$ among all such pairs (c, d); and (iv) the smallest difference $D = d^* - c^*$ is obviously unique. However, there may exist several pairs of (c^*, d^*) that have the same difference D. To obtain a unique interval, we pick the pair, denoted by (c_k^*, d_k^*), so that d_k^* is the smallest d^* and the difference $d_k^* - c_k^* = D$. Then the pair (c_k^*, d_k^*) is unique. In this step, we define

$$C_k(X) = C_{c_k^*,d_k^*}(X), \text{ and } C_I(k - 1) = C_k(k - 1) = [c_k^*, d_k^*]. \tag{13}$$

The interval construction is complete by induction.

Fact (iii) implies that we pick a best (shortest) interval in each step, and this guarantees $C_I(X)$ to be admissible. Fact (iv) assures a unique output from Algorithm I. Based on our extensive numerical study, for the levels of 0.9, 0.95, and 0.99, at each step $k(> 1)$ there is only one pair (c^*, d^*) with the difference D; multiple pairs with the difference D may occur only when $1 - \alpha$ is small. In any case, $C_I(X)$ is well defined and Algorithm I generates a unique output. To summarize the results we have

Theorem 1. Following Algorithm I above, a unique exact $1 - \alpha$ confidence interval $C_I(X) = [L_I(X), U_I(X)]$ for N is generated. Both confidence limits are nonincreasing in X, and any proper subinterval of $C_I(X)$ is of level less than $1 - \alpha$.

Example 1. Pollock et al. (1990) gave data to estimate the size of a population of bobwhite quail in a Florida field research station. Capture–recapture sampling was employed: $M = 148$ quail were trapped, banded over a 20-day period, and released, then $n = 82$ birds were captured, of which $X = 39$ had bands, so were recaptured. We obtain two 95% confidence intervals: $C_O(39) = [260, 394]$ and the improvement $C_I(39) = [263, 394]$. It takes 12 min for an HP 2760 laptop (Intel (R) Core(TM) i5-2520M CPU@2.50GHz RAM=8GB) to compute $C_I(x)$ for $x = 0, \ldots, 82$.

Unlike the case of one-sided intervals in which the smallest interval always exists, the smallest two-sided interval typically does not exist. Wang (2006) provided a necessary and sufficient condition for the existence of the smallest interval for a binomial proportion. That is, it exists only when the number of trials n and/or the level $1 - \alpha$ is very small. A similar result is expected for the hypergeometric case. So in practice using an admissible two-sided interval for N is an optimal choice. Here, $C_I(X)$ is constructed by squeezing each $C_O(x)$ in the order $x = 0, \ldots, \min\{n, M\}$. One may use any order on the x-values to improve any existing level $1 - \alpha$ interval. For example, one may squeeze the trivial constant interval $C_T(X) \equiv [\max\{n, M\}, +\infty)$ in the order $x = \min\{n, M\}, \ldots, 0$. This, however, yields an admissible interval that is short for large x but might be long for small x. Our initial choice of $C_O(X)$ with the order $x = 0, \ldots, \min\{n, M\}$ makes the resulting interval $C_I(X)$ perform well in general.

Figure 2 compares the coverage probabilities of two 95% intervals $C_I(X)$ and $C_O(X)$ in two cases: $M = 148$ and $n = 82$ as in Example 1, and $M = 200 = n$ with larger sample sizes. It is clear that (i) both intervals have coverage at least 0.95, and (ii) $C_I(X)$ has a much closer coverage to 0.95 than $C_O(X)$. To compare interval lengths when $M = 200 = n$, we find that $C_I(x)$ is a proper subset of $C_O(x)$ at 156 out of 201 x-values and the substantial improvement in length occurs when x is small. The ratio of two total interval lengths is

$$\frac{\sum_{x=1}^{200}(U_I(x) - L_I(x))}{\sum_{x=1}^{200}(U_O(x) - L_O(x))} = 0.5455$$

(the case of $x = 0$ is excluded since $C_I(0)$ and $C_O(0)$ both have an infinite length), and the average interval length ratio is, if we define $\frac{+\infty}{+\infty} = 1$ and $\frac{0}{0} = 1$,

$$\frac{1}{201}\sum_{x=0}^{200}\frac{(U_I(x) - L_I(x))}{(U_O(x) - L_O(x))} = 0.9532.$$

If we select the second sample of size n with replacement, see Bailey (1951) and Bonett (1988), then X_R, the number of tagged units in this sample, follows a binomial distribution with n trials and a probability of success $p = M/N$. One can mimic the process described in this section to construct two smallest one-sided intervals and one admissible two-sided interval for N using X_R. The details are skipped.

3. ESTIMATING THE NUMBER OF UNITS WITH ATTRIBUTE M

Now return to the case of $X \sim \text{Hyper}(M, N, n)$ with M unknown and N known.

3.1 Coverage of the Interval $C^M(X)$ for M

Figure 3 indicates that the minimum coverage probability of the approximate interval $C^M(X)$ is equal to n/N and is achieved

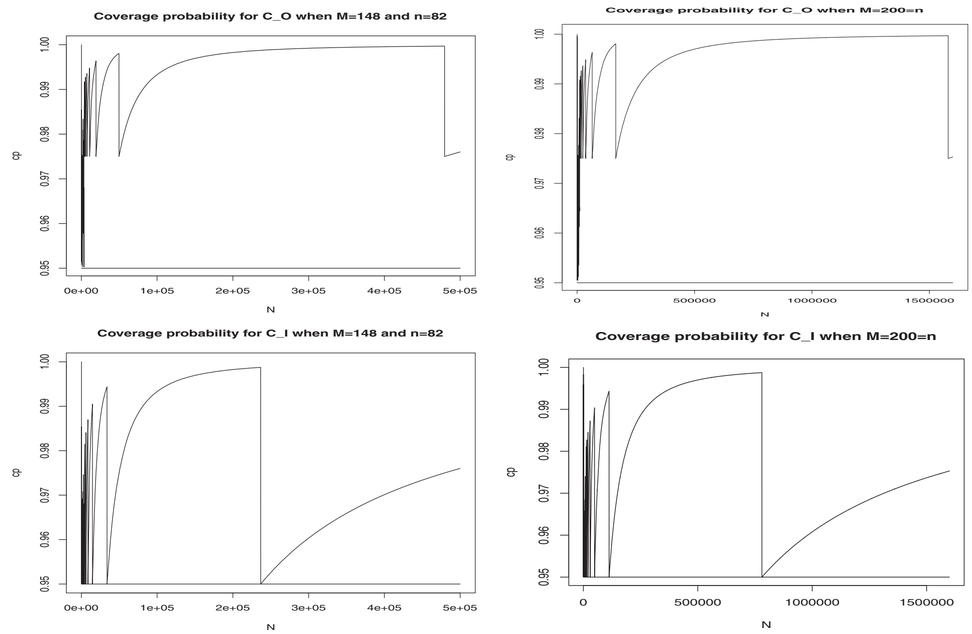


Figure 2. Coverage probabilities (cp) of two exact 95% intervals $C_O(X)$ and $C_I(X)$ (the proposed interval), plotted versus N, with a reference line at 0.95.

at $M = 1$ and $M = N - 1$. Therefore, there is no guarantee for $C^M(X)$ to capture M with the desired confidence level if the sample is much smaller than the population.

Lemma 7. For any $\alpha \in (0, 1)$, N and n, the $1 - \alpha$ Wald-type interval $C^M(X)$ for M given in (5) has a minimum coverage probability no larger than n/N.

3.2 Two Exact Smallest One-Sided Intervals for *M*

For an interval $C(X) = [L(X), U(X)]$ for M, there are two natural restrictions:

$$L(X) \text{ is a nondecreasing function in } X; \tag{14}$$

$$U(X) = N - L(n - X), \tag{15}$$

since a large M tends to yield a large value of X and $n - X \sim \text{Hyper}(N - M, N, n)$. Therefore, $U(X)$ is also nondecreasing. Under (14), the smallest lower one-sided $1 - \alpha$ confidence interval $[L_S^M(X), N]$ is derived in Lemma 8, and the smallest upper one-sided interval $[0, U_S^M(X)]$ is derived from $[L_S^M(X), N]$ following (15).

Lemma 8. For any $\alpha \in (0, 1)$, let

$$L_S^M(x) = \begin{cases} 0, & \text{if } x = 0; \\ \max\left\{M : \sum_{i=0}^{x-1} p_h(i; M-1, N) \geq 1 - \alpha\right\}, & \text{if } x > 0. \end{cases} \tag{16}$$

Then, among all $1 - \alpha$ confidence intervals for M of the form $[L(X), N]$ with nondecreasing $L(X)$, $[L_S^M(X), \infty)$ is the smallest, that is, $L(X) \leq L_S^M(X)$.

The proof is similar to Lemma 5 and is omitted.

Lemma 9. For any $\alpha \in (0, 1)$, let

$$U_S^M(X) = N - L_S^M(n - X). \tag{17}$$

Then, among all $1 - \alpha$ confidence intervals for M of the form $[0, U(X)]$ with nondecreasing $U(X)$, $[0, U_S^M(X)]$ is the smallest, that is, $U_S^M(X) \leq U(X)$.

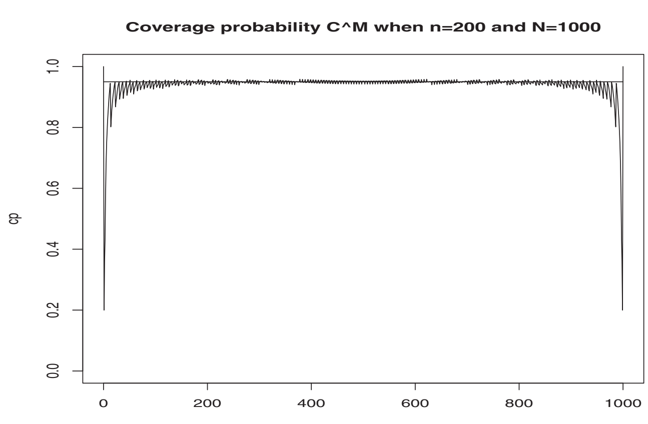


Figure 3. Coverage probability (cp) of the 95% interval $C^M(X)$, plotted versus M, with a reference line at 0.95.

3.3 An Exact Admissible Two-Sided Interval $C_I^M(X)$ for *M*

Here, we present the second major result of this article. Similar to Section 2.3, we start from two smallest one-sided $1 - \alpha/2$ intervals, $[L_O^M(X), N]$ and $[0, U_O^M(X)]$ for M constructed in Lemmas 8 and 9, then $C_O^M(X) = [L_O^M(X), U_O^M(X)]$ is a $1 - \alpha$ interval. In fact, $C_O^M(X)$ is equal to the Konijn (1973) interval, denoted by $[L_K^M(X), U_K^M(X)]$, since $L_K^M(x) = \max\{M : \sum_{i=x}^{n} p_h(i; M', N) \leq \alpha/2, \forall\ M' < M\}$ for $x > 0$ and $L_K^M(0) = 0$ (see Buonaccorsi 1987) are equivalent to (16) with α replaced by $\alpha/2$. We apply an iterative algorithm (Algorithm II) to $C_O^M(X)$ to obtain an admissible interval $C_I^M(X) = [L_I^M(X), U_I^M(X)]$. The process is similar to the construction of $C_I(X)$ but with two differences. (a) The length of $C_O^M(x)$ reaches its maximum when $x \approx [\frac{n}{2}]$, so the interval construction starts at $x = \frac{n}{2}$ if the sample size n is even and at $x = [\frac{n}{2}]$ if n is odd, then x goes

Table 4. Two 95% intervals, $C_O^M(X)$ and $C_I^M(X)$, when $n = 20$ and $N = 200$

x	0	1	2	3	4	5	6	7	8	9	10
$U_O^M(x)$	32	47	61	73	85	95	106	116	125	134	143
$U_I^M(x)$	29	45	61	70	85	95	104	114	123	130	138
$L_I^M(x)$	0	1	4	10	16	22	30	34	45	49	62
$L_O^M(x)$	0	1	3	8	13	19	26	33	40	48	57
x	20	19	18	17	16	15	14	13	12	11	
$U_O^M(x)$	200	199	197	192	187	181	174	167	160	152	
$U_I^M(x)$	200	199	196	190	184	178	170	166	155	151	
$L_I^M(x)$	171	155	139	130	115	105	96	86	77	70	
$L_O^M(x)$	168	153	139	127	115	105	94	84	75	66	

down by 1 in each step until x reaches 0. (b) In each step, we use (15) to determine $C_I^M(n-x)$ from $C_I^M(x)$ as follows:

$$L_I^M(n-x) = N - U_I^M(x), \text{ and } U_I^M(n-x) = N - L_I^M(x). \quad (18)$$

We determine up to four unknowns (but only up to two independent unknowns) in each step. Algorithm II in the Appendix generates $C_I^M(X)$ for an even sample size n. Theorem 2 lists some properties for $C_I^M(X)$. The proof is similar to Theorem 1 and is omitted. When n is odd, $C_I^M(X)$ has the same result, so details are skipped.

Theorem 2. For an even sample size n, a unique exact $1-\alpha$ confidence interval $C_I^M(X) = [L_I^M(X), U_I^M(X)]$ for M is generated following Algorithm II. This interval satisfies (14) and (15), and is admissible under the set inclusion criterion.

Table 4 contains the 95% interval $C_O^M(X)$ and the improved interval $C_I^M(X)$ when $n = 20$ and $N = 200$. The minimum coverage probabilities for $C_O^M(X)$ and $C_I^M(X)$ are 0.9517089 and 0.9500568, respectively, and the coverage probabilities are given in Figure 4. The ratio of the total lengths for $C_I^M(X)$ and $C_O^M(X)$ is equal to 0.9408.

The construction of $C_I^M(X)$ starts at $x = 10$, and in each step we determine $C_I^M(x)$ and $C_I^M(n-x)$. In Step 1, we only determine $C_I^M(10)$ because $x = 10 = n - x$ here.

We lift $L_O^M(10) = 57$ up to an integer a and press $U_O^M(10) = 143$ down to $N - a$ simultaneously, and keep the other $C_O^M(x)$'s unchanged. Name this interval $C_{a,N-a}^M(X)$. For each integer $a \geq 57$, compute the minimum coverage of $C_{a,N-a}^M(X)$ and find that it is at least 0.95 when $a = 62$, but is less than 0.95 when $a = 63$. So $a = 62$ is the largest integer to make $C_{a,N-a}^M(X)$ of level 0.95. Then define $C_I^M(10) = [62, 138]$. In Step 2, we determine $C_I^M(X)$ at $x = 9$ and $x = 11$ by squeezing $C_{a,N-a}^M(X) = [L_{a,N-a}^M(X), U_{a,N-a}^M(X)]$ with $a = 62$ from Step 1. Lift $L_{a,N-a}^M(9) = 48$ up to an integer c (press $U_{a,N-a}^M(11) = 152$ down to $N - c$) and press $U_{a,N-a}^M(9) = 134$ down to an integer d (lift $L_{a,N-a}^M(11) = 66$ up to $N - d$), and keep the rest of $C_{a,N-a}^M(x)$'s unchanged. Name this interval $C_{c,d}^M(X)$. For all $48 \leq c \leq d \leq 134$, we compute the minimum coverage for $C_{c,d}^M(X)$, and find the pair $(c^*, d^*) = (49, 130)$ such that $d^* - c^*$ is the smallest $d - c$ among those $C_{c,d}^M(X)$'s with minimum coverage at least 0.95. So, define $C_I^M(9) = [49, 130]$, and $C_I^M(11) = [70, 151]$ following (18). Repeat Step 2 for $x = 8, \ldots, 0$, and the construction ends in Step 11. A formal description of this process, Algorithm II, is given in the Appendix.

In the case of $n = 200$, $N = 1000$, and $1 - \alpha = 0.95$, $C_I^M(x)$ is a proper subset of $C_O^M(x)$ at 185 out of 201 x-values. The ratio of two total interval lengths and the average interval length ratio are

$$\frac{\sum_{x=0}^{200}\left(U_I^M(x) - L_I^M(x)\right)}{\sum_{x=0}^{200}\left(U_O^M(x) - L_O^M(x)\right)} = 0.9744$$

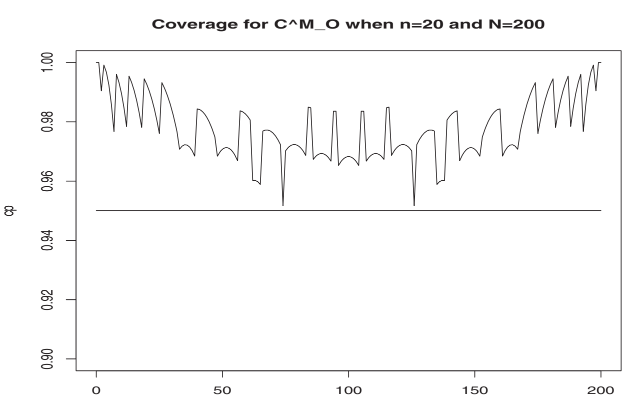
and

$$\frac{1}{201}\sum_{x=0}^{200}\frac{\left(U_I^M(x) - L_I^M(x)\right)}{\left(U_O^M(x) - L_O^M(x)\right)} = 0.9748,$$

respectively. On average, $C_I^M(X)$ reduces the length by 2.5%. To compare $C_I^M(X)$ with $C^M(X)$, the ratio of two total interval lengths is 1.0049. The interval $C_I^M(X)$ is just a little longer, but has a much larger minimum coverage 0.95 compared with 0.2 for $C^M(X)$.

4. DISCUSSION

Extensive studies—see, for example, Newcombe (1998) and Pires and Amado (2008)—have been done to investigate the



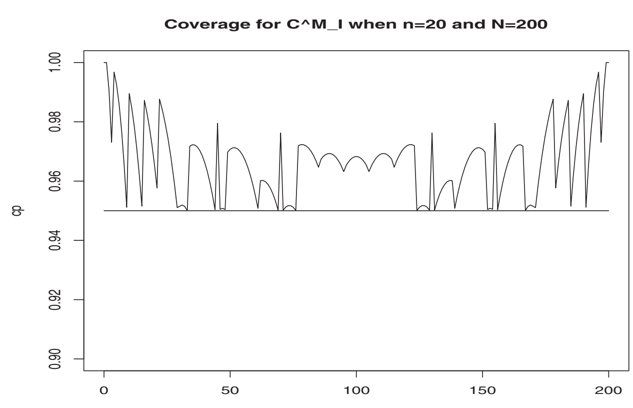


Figure 4. Coverage probabilities (cp) of two exact 95% intervals $C_O^M(X)$ and $C_I^M(X)$ (the proposed interval), plotted versus M, with a reference line at 0.95.

performance of existing approximate intervals for a binomial proportion that relates to M. The Agresti–Coull (1998) interval is considered a successful way to estimate the proportion since a recent study shows that its minimum coverage probability is close to the commonly used nominal levels, 0.9, 0.95, and 0.99, when the sample size is large. Other intervals, including the Wilson (1927) interval and the Jeffery-prior interval (Brown, Cai, and DasGupta 2001), have minimum coverage probabilities much smaller than the nominal levels even for huge sample sizes, pointed out by Huwang (1995). From a mathematical point of view, the Agresti–Coull interval has the same drawback. How to derive an approximate interval that is easy to compute and also has a good coverage probability is still open.

This challenge motivates the use of exact intervals even though intensive computation is typically involved. In this article, Algorithms I and II are introduced to generate exact admissible intervals to infer the unknown parameter, N or M, of hypergeometric distribution, which plays a central role in the random sampling. The results are clean, and are established by proofs, not simulations. The same method can be used to improve any existing $1-\alpha$ interval for the parameter of any single-parameter distribution family with a discrete sample space, including, for example, a Poisson mean and a negative binomial proportion.

Casella (1986) proposed an algorithm to refine any $1-\alpha$ interval for a binomial proportion. He lifted up the lower limit only while we lift up the lower limit and press down the upper limit simultaneously. Our method is more flexible since there exist nonsymmetric intervals, including $C_I(X)$ in Section 2. In each step of Algorithms I and II, the shortest interval is chosen to assure both admissibility and correct coverage. The R codes to compute $C_I(X)$ and $C_I^M(X)$ are available in the online supplementary materials.

In the computation, many minimum coverage probabilities must be evaluated during interval construction to make sure that they never go below $1-\alpha$. Wang (2007) proposed a technique that greatly simplifies the calculation of minimum coverage, and a variation of her method is applied in our R codes. For example, it takes 1.5 min to compute all 101 intervals $C_I^M(x)$, $x = 0, \ldots, 100$, when $n = 100$ and $N = 200$.

APPENDIX: PROOFS AND ALGORITHM II

Proof of Lemma 1. We show that the coverage probability at two points, $u_0 = \max\{n, M\} + 1$ and u_1, are equal to a and b, respectively, which establishes the lemma. First note that the interval $C_{N1}(\min\{n, M\})$ reduces to a point, $\max\{n, M\}$. Then

$$\mathrm{Cover}_{C_{N1}}(u_0) \le 1 - P(X = \min\{n, M\}) = 1 - \frac{\binom{M}{\min\{n,M\}}\binom{u_0-M}{n-\min\{n,M\}}}{\binom{u_0}{n}} = a.$$

Second, $U_{N1}(x)$ is nonincreasing in x and $u_1 > U_{N1}(1)$. Then

$$\mathrm{Cover}_{C_{N1}}(u_1) = P(X = 0) = b.$$

□

Proof of Lemma 2. Since $C_{N2}(X)$ is finite, pick an integer N_0 larger than all upper limits of $C_{N2}(X)$. Then N_0 is not in $C_{N2}(X)$, which implies $\mathrm{Cover}_{C_{N2}}(N_0) = 0$. □

Proof of Lemma 5. For any integer $x \in [0, \min\{n, M\}]$, let

$$G_x = \left\{ N : \sum_{x' > x} p_h(x'; M, N') \ge 1 - \alpha, \forall\ N' < N \right\}.$$

Following Theorem 4 in Wang (2010),

$$L_S(x) = \begin{cases} \max\{n, M\}, & \text{if } G_x = \emptyset; \\ \sup G_x, & \text{otherwise.} \end{cases}$$

It is easy to see that $G_x = \emptyset$ if $x = \min\{n, M\}$; and

$$G_x = \left\{ N : \sum_{i=x+1}^{\min\{n,M\}} p_h(i; M, N-1) \ge 1 - \alpha \right\}$$

for $x < \min\{n, M\}$ because X is stochastically decreasing in N. □

Proof of Theorem 1. Here, we outline the proof and skip the tedious details. First, the interval $C_I(X)$ exists, since $C_{a,+\infty}(x)$ in (11) and $C_{c,d}(x)$ in (12) exist, and $C_I(x)$ is equal to one of them for any x. Second, $C_I(X)$ is of level $1-\alpha$ since all $C_{a,+\infty}(X)$ and $C_{c,d}(X)$ are of level $1-\alpha$. Third, $C_I(X)$ is unique due to the uniqueness of D and d^* in Step k (fact iv). Finally, $C_I(X)$ is admissible because D is the smallest difference $d - c$ (fact iii in Step k). □

Proof of Lemma 7. Since $C^M(0) = [0, 0]$ does not contain 1, the lemma follows:

$$\begin{aligned} \mathrm{Cover}_{C^M}(1) &= P(1 \in C^M(X), X > 0 | M = 1) \\ &\le P(X > 0 | M = 1) = \frac{n}{N}. \end{aligned}$$

□

Algorithm II

Algorithm II is for the computation of $C_I^M(X)$ for M. For illustration purposes, consider the case of an even n. The case of an odd n is similar and is omitted.

Step 1-II: We determine $C_I^M(\frac{n}{2}) = [L_I^M(\frac{n}{2}), U_I^M(\frac{n}{2})]$ by squeezing $[L_O^M(\frac{n}{2}), U_O^M(\frac{n}{2})]$ with a restriction $U_I^M(\frac{n}{2}) = N - L_I^M(\frac{n}{2})$. For any given integer $a \ge L_O^M(\frac{n}{2})$, let

$$C_{a,N-a}^M(x) = \begin{cases} \left[L_O^M(x), \min\left\{N - a, U_O^M(x)\right\}\right], & \text{if } x < \frac{n}{2}; \\ [a, N - a], & \text{if } x = \frac{n}{2}; \\ \left[N - \min\left\{N - a, U_O^M(n - x)\right\},\right. & \\ \quad \left. N - L_O^M(n - x)\right], & \text{if } x > \frac{n}{2} \end{cases} \tag{A.1}$$

if its minimum coverage is at least $1-\alpha$. The minimum in the first line of (A.1) is to make interval $C_{a,N-a}^M(X)$ satisfy (14) and the third line is for (15). Such an interval $C_{a,N-a}^M(X)$ exists because $C_O^M(X)$ is one such interval. Let $C_1^M(X)$ be the $C_{a,N-a}^M(X)$ with the largest a, denoted by a^*, which is obviously unique. In this step, we obtain $C_1^M(X) = C_{a^*,+\infty}^M(X)$ and $C_I^M(\frac{n}{2}) = C_1^M(\frac{n}{2}) = [a^*, N - a^*]$.

By induction, suppose that an interval $C_{k-1}^M(X)$ is constructed for some $k - 1 \ge 1$ (note that $C_1^M(X)$ corresponds to $k = 2$), $C_I^M(x)$ is constructed for $x \in [\frac{n}{2} \pm (k-2)]$ ($C_I^M(x) = C_{k-1}^M(x)$ for $x \in [\frac{n}{2} \pm (k - 2)]$), and $C_{k-1}^M(X)$ satisfies (14) and (15).

Step k-II: In this step, we determine $C_k^M(X)$ and $C_I^M(x)$ at $x = \frac{n}{2} - (k-1)$ (then at $x = \frac{n}{2} + (k-1)$ by (15)) by modifying $C_{k-1}^M(X) = [L_{k-1}^M(X), U_{k-1}^M(X)]$. For any pair of integers (c, d) with $L_{k-1}^M(\frac{n}{2} - (k - 1)) \le c \le d \le U_{k-1}^M(\frac{n}{2} - (k-1))$, let

$$C_{c,d}^M(x) = \begin{cases} [L_{k-1}^M(x), \min\{d, U_{k-1}^M(x)\}], & \text{if } x < \frac{n}{2} - (k-1); \\ [c, d], & \text{if } x = \frac{n}{2} - (k-1); \\ C_{k-1}^M(x), & \text{if } \frac{n}{2} - (k-2) \le x \le \frac{n}{2} \end{cases} \tag{A.2}$$

(for $x > \frac{n}{2}$ define $C_{c,d}^M(x)$ following (15)) if its minimum coverage is at least $1-\alpha$. Interval $C_{c,d}^M(X)$ exists for some (c, d) because $C_{k-1}^M(X)$ is one such interval. Let D be the smallest difference of such $d - c$'s.

Furthermore, let (c^*, d^*) be the pair with the smallest d among pairs that have a difference D. Therefore, (c^*, d^*) exists and is unique. Then define $C_k^M(X) = C_{c^*,d^*}^M(X)$ and $C_I^M(x) = C_{c^*,d^*}^M(x) = [c^*, d^*]$ at $x = \frac{n}{2} - (k-1)$. The construction is complete in Step $(\frac{n}{2}+1)$ by induction.

Again, the minimum function in the first line of (A.2) is to make the interval $C_{c^*,d^*}^M(X)$ satisfy (14). Incidentally, in our extensive numerical study, the pair (c, d) that has the smallest difference D is typically unique for the commonly used confidence levels. Nevertheless, our selection of the smallest d in addition to the smallest D guarantees a unique output from Algorithm II. Since we in every step select the pair with the shortest difference D, the interval $C_I^M(X)$ is admissible.

SUPPLEMENTARY MATERIALS

Two R-codes, hyper-m.r and hyper-N.r, are provided to compute confidence intervals $C_I^M(X)$ for M and $C_I(X)$ for N, respectively.

[Received May 2013. Revised January 2014.]

REFERENCES

Agresti, A., and Coull, B. A. (1998), "Approximate is Better Than 'Exact' for Interval Estimation of Binomial Proportions," *The American Statistician*, 52, 119–126. [1493,1498]

Bailey, N. T. J. (1951), "On Estimating the Size of Mobile Populations From Recapture Data," *Biometrika*, 38, 293–306. [1491,1493,1495]

Bonett, D. G. (1988), "Small Sample Bias and Variance of Lincoln-Petersen and Bailey Estimators," *Biometrical Journal*, 30, 723–727. [1493,1495]

Buonaccorsi, J. P. (1987), "A Note on Confidence Intervals for Proportions in Finite Populations," *The American Statistician*, 41, 215–218. [1493,1496]

Brown, L. D., Cai, T. T., and DasGupta, A. (2001), "Interval Estimation for a Binomial Proportion," *Statistical Sciences*, 16, 101–133. [1493,1498]

Casella, G. (1986), "Refining Binomial Confidence Intervals," *The Canadian Journal of Statistics*, 14, 113–129. [1498]

Chapman, D. G. (1951), "Some Properties of the Hypergeometric Distribution with Application to Zoological Sample Censuses," *University of California Publications in Statistics*, 1, 131–160. [1491]

Cochran, W. G. (1977), *Sampling Techniques* (3rd ed.), New York: Wiley. [1493]

Huwang, L. (1995), "A Note on the Accuracy of an Approximate Interval for the Binomial Parameter," *Statistics & Probability Letters*, 24, 177–180. [1498]

Konijn, H. S. (1973), *Statistical Theory of Sample Survey Design and Analysis*, Amsterdam: North-Holland. [1493,1496]

Lincoln, F. C. (1930), "Calculating Waterfowl Abundance on the Basis of Banding Returns," Circular number 118, U.S. Department of Agriculture, Washington, DC. [1491]

Newcombe, R. G. (1998), "Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods," *Statistics in Medicine*, 17, 857–872. [1497]

Petersen, C. G. J. (1896), "The Yearly Immigration of Young Plaice into the Limfjord From the German Sea," *Report to the Danish Biological Station*, 6, 5–48. [1491]

Pires, A. M., and Amado, C. (2008), "Interval Estimators for a Binomial Proportion: Comparison of Twenty Methods," *Revstat-Statistical Journal*, 6, 165–197. [1497]

Pollock, K. H., Nichols, J. D., Brownie, C., and Hines, J. E. (1990), "Statistical Inference for Capture-Recapture Experiments," *Wildlife Society Monographs*, 107, 1–97. [1495]

Seber, G. A. F. (1982), *The Estimation of Animal Abundance* (2nd ed.), London: Griffin. [1491]

——— (1986), "A Review of Estimating Animal Abundance," *Biometrics*, 42, 267–292. [1491]

——— (1992), "A Review of Estimating Animal Abundance: II," *International Statistical Review*, 60, 129–166. [1491]

Sekar, C. C., and Deming, W. E. (1949), "On a Method of Estimating Birth and Death Rates and Extent of Registration," *Journal of the American Statistical Association*, 44, 101–115. [1491]

Thompson, S. K. (2002), *Sampling* (2nd ed.), New York: Wiley. [1491,1493]

Wang, H. (2007), "Exact Confidence Coefficients of Confidence Intervals for a Binomial Proportion," *Statistica Sinica*, 17, 361–368. [1498]

Wang, W. (2006), "Smallest Confidence Intervals for One Binomial Proportion," *Journal of Statistical Planning and Inference*, 136, 4293–4306. [1494,1495]

Wang, W. (2010), "On Construction of the Smallest One-sided Confidence Interval for the Difference of Two Proportions," *The Annals of Statistics*, 38, 1227–1243. [1498]

Wilson, E. B. (1927), "Probable Inference, the Law of Succession, and Statistical Inference," *Journal of the American Statistical Association*, 22, 209–212. [1498]

Wittes, J. T. (1972), "Note: On the Bias and Estimated Variance of Chapman's Two-Sample Capture-Recapture Population Estimate," *Biometrics*, 28, 592–597. [1493]